PROCEEDINGS OF SCIENCE

Distributed Data Management System at IHEP

Xuantong Zhang,[a,*] Xiaomei Zhang,[a] Haibo Li,[a] Yujiang Bi,[a] Hao Hu[a,b] and Haofan Wang[a]

[a] *Institute of High Energy Physics, Chinese Academy of Science,*
19B Yuquan Road, Shijingshan District, Beijing, China.

[b] *University of Science and Technology of China,*
No.96, JinZhai Road Baohe District, Hefei, Anhui, 230026, China.

E-mail: zhangxuantong@ihep.ac.cn

A distributed grid data management system has been established at the Institute of High Energy Physics (IHEP) since 2014, leveraging the DIRAC data management system. This system has been effectively utilized by the BESIII, JUNO, and CEPC experiments. To address the diverse data scales and intricate data management requirements associated with data production, IHEP developers have embarked on endeavors to create a more adaptable and experiment-oriented grid data management system based on Rucio, owing to its proven reliability, scalability, and automation capabilities. As a result, a production system for Monte Carlo (MC) data production has been implemented, relying on the DIRAC Data Management System (DMS) solution. Furthermore, work is underway to develop Rucio data transfer daemon plugins to facilitate non-grid data transfer for upcoming experiments at IHEP. Supporting infrastructure for distributed data management has also been established, encompassing StoRM and EOS storage elements, an IAM authentication service, as well as a third-party-copy active probing system.

International Symposium on Grids and Clouds (ISGC) 2023,
19 - 31 March 2023
Academia Sinica Taipei, Taiwan

[*] Speaker

1. Introduction

The Institute of High Energy Physics (IHEP), part of the Chinese Academy of Sciences, stands as the premier laboratory for particle physics research within China. The Computing Center of IHEP (IHEP-CC) plays a pivotal role in supporting experiments conducted both in China and internationally, driven by IHEP's scientific pursuits. The center infrastructure features approximately 50,000 CPU cores, 210 GPU cards for computational resources, and a storage capacity exceeding 75 PB on disk and 50 PB on tape.

Since 2009, in response to the rapidly increasing demands for raw data processing and Monte-Carlo (MC) data production from the BESIII experiment [1][2], and to strengthen collaboration with the experiment, IHEP embarked on a study and implementation of a distributed computing environment. The first distributed computing system, based on DIRAC [3] (Distributed Infrastructure with Remote Agent Control), was studied, developed, and deployed as the IHEPDIRAC project, becoming a production service for the BESIII experiment in 2012 [4]. Serving as a generic grid solution, DIRAC provides a Data Management System [5] (DMS) suitable for small sites and those lacking in grid computing maintainers. Currently, DIRAC-DMS primarily serves the JUNO experiment [6], a neutrino observatory located in southern China, which is planning to generate 2.4 PB of raw data and 0.6 PB of MC and calibration data every year [7].

In order to develop a more flexible grid data management system oriented towards the needs of future experiments, the Rucio system [8], originally developed for ATLAS experiment data management [9], is under consideration as an alternative distributed data management system at IHEP. Inspired by the BelleII experiment's DIRAC-Rucio integration data management solution [10], IHEP developers are also working on a customized Rucio-based distributed data management system for future experiments. Currently, the Rucio-based solution is in its initial stages of development for the HERD experiment [11], a high-energy cosmic detector aboard the China Space Station, which aims to generate 5.5 PB of data in 5 years and 15.5 PB in 10 years.

In recent years, WLCG [12] (Worldwide LHC Computing Grid) has been promoting the use of XrootD [13] and WebDav [14] protocols to replace the Gridftp protocol as the future third-party-copy (TPC) protocol for WLCG [15]. A new authentication and authorization model is also in development to replace the X.509 model for data access among sites [16]. At IHEP, StoRM [17] and EOS [18] have been chosen as the distributed data storage system for the JUNO and HERD experiments, with IAM serving as the token-based data authorization service.

1.1 DIRAC Data Management System

DIRAC-DMS currently serves as the primary data management system for the BESIII, JUNO, and CEPC experiments conducted at IHEP. In this system, a file or directory is defined using a Linux-like path structure, referred to as a Logic File Name (LFN). All LFNs collectively form the DIRAC File Catalog (DFC). Additionally, the Data Management System (DMS) offers dataset and metadata management capabilities for files and directories. This allows data to be queried based on datasets and metadata, facilitating further data processing or transfer activities. At IHEP, LFNs for production data are established using task names, software versions, and physical parameters. DIRAC-DMS provides users with command-line tools and a web-based user interface. Furthermore,

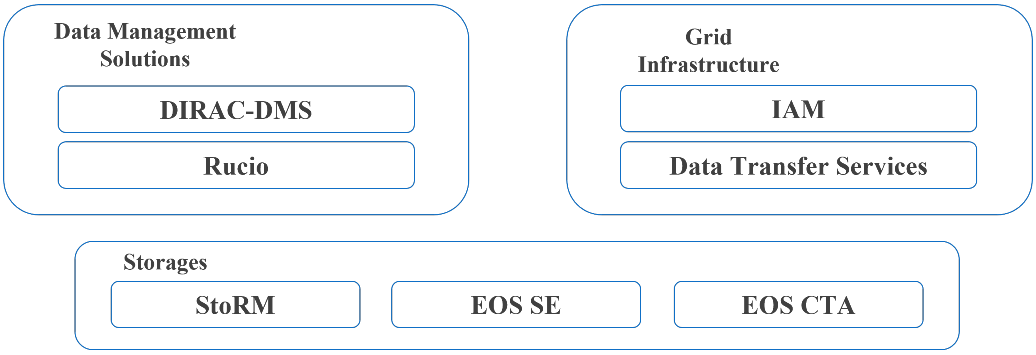

Figure 1: IHEP data management system components.

IHEP-DIRAC serves as a DIRAC extension, offering additional command-line tools built upon the DIRAC-DMS API for customized use cases, such as mass file registration and transfer.

DIRAC-DMS is highly involved in MC data production and raw data flow at IHEP. ProdSys is a massive MC data production task manager which is developed in DIRAC to automatically create and manage workflow, and multiple IHEP experiments software has already been merged in it. ProdSys uses DIRAC-DMS to manage raw data input, produced data registering and multi-sites data replication in its workflow. For raw data flow, depending on different experiments needs, DIRAC-DMS plays different roles in raw data transfer and archive. Following is a typical raw data flow,

- It initiates data processing triggers upon data arrival at the local storage facility at IHEP.
- It registers data from IHEP's local storage into the DIRAC-DFC.
- It replicates data from IHEP to cooperating data centers' disks and subsequently registers this data.
- It archives data onto tape storage and registers it within the DIRAC-DFC.
- It conducts data validation and monitors its status throughout the process.

DIRAC-DMS has been in operation at IHEP for nearly a decade. As of 2022, it continues to efficiently handle substantial data transfers with exceptional quality and speed. A total of 1.4 PB of data were successfully transferred, and 1 PB, comprising approximately 4 million files, were registered and managed within the DIRAC-DMS system in the year of 2022.

1.2 Rucio

The Rucio system exhibits remarkable scalability and places a strong emphasis on effective data management. Consequently, the IHEP Rucio solution is intricately designed for deep integration with experiment software, functioning as a backend service. Leveraging the extensive API capabilities of Rucio, it becomes straightforward to devise customized data logic catalogs tailored to the distinct data structures of various experiments. Furthermore, the development of user-oriented APIs within experiment software for streamlined data access is greatly facilitated. Additionally, an ongoing project involves the development of a plugin for non-gird data transfer tools at IHEP.

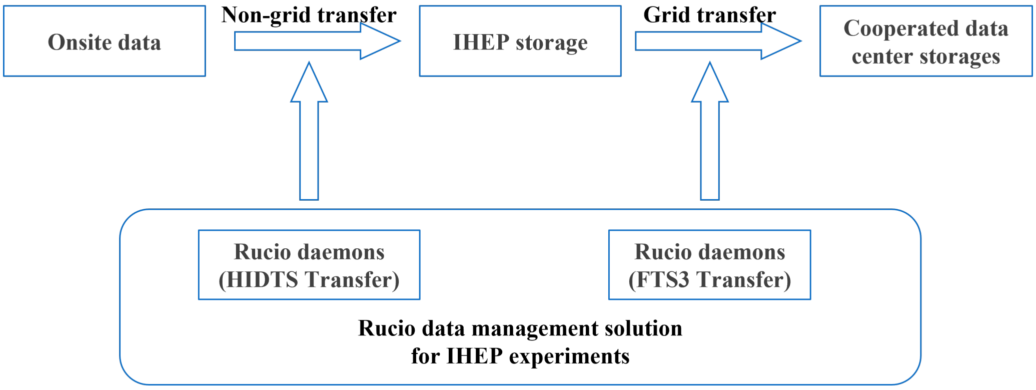
Onsite data
Non-grid transfer
IHEP storage
Grid transfer
Cooperated data center storages
Rucio daemons (HIDTS Transfer)
Rucio daemons (FTS3 Transfer)
Rucio data management solution for IHEP experiments

Figure 2: IHEP Rucio with HIDTS plugin for data management from onsite data to remote site distributed data.

Rucio utilizes Data Identifiers (DIDs) to construct a logical namespace for data. A standard Rucio DID is comprised of a "scope" and a "file name," linked together with a colon, resembling the format "scope:name." The file name within the DID lacks stringent naming conventions, affording developers the flexibility to customize it as needed. At IHEP, the scope is defined in terms of the data's status within the data flow, while the name follows a Linux-like file path structure. Differing from DIRAC-DMS, Rucio supports a two-level file collection design that involves datasets and containers. Files are exclusively collected within datasets, and multiple datasets can be housed within containers. Containers themselves can also be nested within other containers. Adhering to this organizational structure, a dataset DID at IHEP is defined as a directory path encompassing all files within that directory, excluding sub-directories. Conversely, a container DID is defined as a directory path that encompasses all sub-directories (which are also containers) and the dataset within the current directory. The distinction between datasets and containers hinges on the presence of a trailing '/'. Examples are shown in table 1.

DID type	Example	Definition
File	`temp:/herd/user/output.root`	File's logic name
Dataset	`temp:/herd/user/`	Files collection in a directory
Container	`temp:/herd/user`	Sub-directories containers collection

Table 1: DID Definition and Examples

IHEP's HIDTS stands as a non-grid data transfer service dedicated to the local storage site at IHEP. Its operational approach shares similarities with FTS3, although it utilizes proprietary transfer protocols rather than grid protocols and tokens. HIDTS primarily facilitates pre-transfers between experiment onsite locations and IHEP's grid storage elements. In order to manage HIDTS transfer jobs effectively, a new transfer plugin is currently in development within the Rucio framework. This HIDTS plugin will function as a transfer daemon within IHEP's Rucio system, enabling comprehensive data management, from onsite data to distributed grid data across remote sites, as the figure 2 shows.

Building upon these extensions, a user-centric API has been developed, empowering physicist users to seamlessly integrate data management into their data production and analysis software. To illustrate this, consider a Monte Carlo (MC) data flow scenario, where the DID scope can be defined to reflect data statuses such as "temp," "good," and "bad." All data management tasks within the following workflow can be conducted using the user-oriented API:

- Registering all raw MC data under the "temp" scope.
- Employing API-based data validation programs to assess data quality.
- In the case of data deemed "good," transitioning the scope to "good" status, followed by metadata registration.
- For data deemed less than satisfactory, relocating the scope to "bad," awaiting eventual deletion.

The Rucio solution is currently undergoing testing at IHEP. In the year 2022, the Rucio solution successfully executed data transfer missions from IHEP to European collaborator sites. More than 70 TB of data, comprising approximately 10 million files, were transferred seamlessly. Rucio demonstrated robust and stable performance following adjustments to data policies and configurations.

2. Grid Data Infrastructures

The grid infrastructure plays a pivotal role in supporting the data management service, encompassing key components such as the storage element (SE), token-based authentication system, and third-party-copy (TPC) protocols.

2.1 Storage Element

A storage element enables grid users to access storage resources using grid authentication. At IHEP, the SE is constructed using EOS and StoRM, both of which support TPC protocols, including XrootD and WebDav.

The EOS system, developed by CERN, is a robust exabyte-scale data storage solution. Within IHEP, EOS serves as the primary data storage system, housing raw and production data from experiments. To ensure data security, each experiment is assigned its independent EOS instance, wherein storage areas for grid users and local users are segregated. Additionally, an EOS-CTA, designated for tape data management, is slated for deployment in support of the JUNO experiment in April 2023, with its buffer also functioning as an EOS storage element.

Concurrently, the StoRM system at IHEP is designed to mount the Lustre file system as its backend storage area. Given that the Lustre file system forms the foundational infrastructure for local cluster users' data analysis at IHEP, StoRM serves as a vital bridge, facilitating access to local cluster data using grid protocols and tools. Unlike EOS, all experiments share a single StoRM service, with distinct storage areas employing different authentication methods to distinguish grid users from various experiments.

At IHEP, EOS serves as the primary storage solution for raw and production data, assuming the role of the source site for grid data distribution in experiments. Conversely, StoRM facilitates the accessibility of local data via grid technology, enabling external experiment collaborators, beyond the confines of IHEP, to access users' local data using grid protocols.

2.2 Third-Party-Copy Protocols

TPC protocols serve as the fundamental grid data transfer protocols, enabling grid users to perform direct data transfers between two SEs without the necessity of routing data through their local machines. Establishing a comprehensive authentication service is imperative to authorize grid user transfers. In alignment with WLCG's recommendations, XrootD and WebDav protocols are promoted, coupled with an Identity and Access Management (IAM) tokens service, forming the core of grid infrastructures.

At IHEP, an IAM service is currently under development to cater to the requirements of future experiments. In line with the WLCG tokens framework, this service supports OIDC, SAML, and X.509 certificate-based user authentication, while generating WLCG-Tokens and VOMS proxies for user authorization. These tokens play a pivotal role in authenticating grid users during data transfers between SEs.

Furthermore, an active TPC probing system has been designed for monitoring the functionality and speed of TPC operations, leveraging Gfal2 tools as its foundation. The probing data is systematically collected through Elasticsearch and presented on Kibana dashboards. Functionality tests for TPC are conducted at regular intervals of 30 minutes, while transfer speed tests are executed every 2 hours.

3. Conclusion

IHEP initiated its exploration of a distributed data management system with the inception of DIRAC-DMS. As WLCG techniques evolved, IHEP embraced additional systems such as Rucio, EOS, TPC, and a Token-based authentication service. Concurrently, dedicated efforts were channeled into the development of this system, aligning it with the specific requirements of ongoing experiments at IHEP. The experience garnered from designing this distributed data management system not only benefits the present experiments at IHEP but also holds valuable insights for future endeavors and high-energy physics centers embarking on similar experiments.

Acknowledgments

The authors would like to express their appreciation to the JUNO DCI group for their operation and use of DIRAC-DMS, Zuhao Li for discussing the HERD experiment's data flow, and colleagues at IHEP computing center for their support. This work is supported in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDA10010900, the National Natural Science Foundation of China (NSFC) under Grant No. 12188102, No. 12075268, and No. 12005247, and the Xie Jialin Foundation of the Institute of High Energy Physics, Chinese Academy of Sciences under Grant No. 2022000092.

References

[1] Besiii Collaboration, *The construction of the BESIII experiment*. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2009, 598(1): 7-11.

[2] Ablikim M, An Z H, Bai J Z, et al, *Design and construction of the BESIII detector*. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2010, 614(3): 345-399.

[3] Tsaregorodtsev A, Garonne V, Closier J, et al, *DIRAC–distributed infrastructure with remote agent control*. Proc. of CHEP2003. 2003.

[4] Deng Z Y, Li W D, Lin L, et al, *Experience of BESIII data production with local cluster and distributed computing model*. Journal of Physics: Conference Series. IOP Publishing, 2012, 396(3): 032031.

[5] Bargiotti M, Smith A C, *DIRAC Data Management: consistency, integrity and coherence of data*. Journal of Physics: Conference Series. IOP Publishing, 2008, 119(6): 062013.

[6] Adam T, An F, An G, et al, *JUNO conceptual design report*. arXiv preprint arXiv:1508.07166, 2015.

[7] Andronico G, Xiaomei Z, Weidong L, *Jiangmen Underground Neutrino Observatory computing requirements and infrastructure*. XXIX International Symposium on Lepton Photon Interactions at High Energies. August 5-10. 2019: 122.

[8] Barisits M, Beermann T, Berghaus F, et al, *Rucio: Scientific data management*. Computing and Software for Big Science, 2019, 3: 1-19.

[9] Garonne V, Vigne R, Stewart G, et al, *Rucio–The next generation of large scale distributed system for ATLAS Data Management*. Journal of Physics: Conference Series. IOP Publishing, 2014, 513(4): 042021.

[10] Serfon C, Mashinistov R, De Stefano J S, et al, *Integration of Rucio in Belle II*. EPJ Web of Conferences. EDP Sciences, 2021, 251: 02057.

[11] Zhang S N, Adriani O, Albergo S, et al, *The high energy cosmic-radiation detection (HERD) facility onboard China's Space Station*. Space Telescopes and Instrumentation 2014: Ultraviolet to Gamma Ray. SPIE, 2014, 9144: 293-301.

[12] Shiers J, *The worldwide LHC computing grid (worldwide LCG)*. Computer physics communications, 2007, 177(1-2): 219-223.

[13] Adye T, Bockelman B, Ellis K, et al, *XRootD Third Party Copy for the WLCG and HLLHC*. EPJ Web of Conferences. EDP Sciences, 2020, 245: 04034.

[14] Bockelman B, Ceccanti A, Furano F, et al, *Third-party transfers in WLCG using HTTP*. EPJ Web of Conferences. EDP Sciences, 2020, 245: 04031.

[15] Forti A, *Modernizing Third-Party-Copy Transfers in WLCG*. EPJ Web of Conferences (to be published). 2019.

[16] Bockelman B, Ceccanti A, Collier I, et al, *WLCG Authorisation from X. 509 to Tokens*. EPJ Web of Conferences. EDP Sciences, 2020, 245: 03001.

[17] Shoshani A, Sim A, Gu J. *Storage resource managers: Middleware components for grid storage*. In: NASA Conference Publication. NASA; 2002. p. 209-224.

[18] Peters A J, Janyst L. *Exabyte scale storage at CERN*. In: Journal of Physics: Conference Series. IOP Publishing; 2011. 331(5):052015.