ISSN 1547-4771, Physics of Particles and Nuclei Letters, 2022, Vol. 19, No. 5, pp. 547–549. © Pleiades Publishing, Ltd., 2022.
Russian Text © The Author(s), 2022, published in Pis'ma v Zhurnal Fizika Elementarnykh Chastits i Atomnogo Yadra, 2022.

COMPUTER TECHNOLOGIES IN PHYSICS

CERN-JINR-INP-KazNU Data Center: Current Status and Plans

N. Balashov[a], N. Burtebayev[b, c], V. Korenkov[a, b, d], N. Kutovskiy[a], Ye. Mazhitova[a, b, *], I. Satyshev[a, b], and R. Semenov[a, d]

[a] *Joint Institute for Nuclear Research, Dubna, 141980 Russia*
[b] *Institute of Nuclear Physics, Almaty, Kazakhstan*
[c] *Al-Farabi Kazakh National University, Almaty, Kazakhstan*
[d] *Plekhanov Russian University of Economics, Moscow, Russia*
**e-mail: emazhitova@jinr.ru*

Received January 26, 2022; revised April 20, 2022; accepted April 20, 2022

Abstract—Modern scientific projects generate a huge amount of data that needs to be stored, processed and analyzed. It is often impossible to solve such tasks within a single data center. Therefore, it is necessary to prepare a distributed infrastructure consisting of hardware, specialized software and communication channels. One of the important preliminary steps in building such an infrastructure is to study existing solutions and select a suitable model for distributed data storage, processing, and analysis. Building a distributed infrastructure requires software capable of solving tasks such as authentication and authorization, the creation of an information system, monitoring tools, the management of computing tasks, data storage and transfer. In this paper, a study of existing solutions was carried out and a particular model was selected for creating a CERN-JINR-INP-KazNU data center, which will be further integrated into a distributed infrastructure.

DOI: 10.1134/S1547477122050089

INTRODUCTION

To solve scientific problems and process data of Kazakhstani experiments, as well as to participate in international experiments, it was decided to create a CERN-JINR-INP-KazNU data center and integrate it into the distributed information and computing environment (DICE) built on the resources of the JINR Member States' organizations.

It is planned to use this center to work with such experiments as Monte-Carlo simulation for {MPD, SPD, BM@N}@NICA, Baikal-GVD, to analyze data obtained in the CLOUD/PS215 international project at CERN in the period 2009–2023 and data obtained in Nur-Sultan in the period 2017–2023, to study cosmic ray interactions at high-altitude stations located at an altitude of 3340 meters above sea level in the Tian Shan mountains—"Hadron-9", "Hadron-44" and "INKA-55"; variations of secondary cosmic rays due to changes in the characteristics of the surface electric field and changes in the meteorological parameters of the surface atmosphere are investigated. The material base of the experimental research in the ongoing project is a cosmophysical complex (Nur-Sultan), which includes a CARPET detector, a neutron detector and an electrostatic fluxmeter.

These resources can also be used for the comprehensive study of the fundamental properties of quark-gluon plasma within the ALICE project at the CERN Large Hadron Collider, for the investigation of the structure of azimuthal-rapidity distributions of events obtained at the Hadron-44 facility and comparison with the data of the CMS detector, as well as for experimental and theoretical studies of the characteristics of exotic stable and radioactive nuclei (differential cross sections for elastic scattering of ^{13}C ions by ^{9}Be nuclei at $E_{l.s.c} = 19.5$ and 16.25 MeV in the range of angles 8°–48° in the laboratory system were measured on the DC-60 cyclotron of the Institute of Nuclear Physics of the Republic of Kazakhstan).

INTERWARE FOR DICE

Over the past decades, distributed computing has come a long way from small research projects to significant scientific projects, for example, the Worldwide LHC Computing Grid (WLCG) project [1] for processing data from the Large Hadron Collider at CERN, which combined the computing resources of institutes from different countries and allowed processing about 100 PB of data. A promising area is the integration of various distributed computing technologies for data management. With the development of technologies such as supercomputers, cloud services or voluntary computing, it becomes necessary to

include them in the general management system so that their use does not create additional difficulties for users. The major issue here is the choice of software that enables to build such an infrastructure and organize computations with sufficient reliability and minimal costs.

Building a distributed infrastructure requires software capable of solving tasks such as authentication and authorization, the creation of an information system and monitoring tools, the management of computing tasks, data storage and transfer. This has already been addressed by the WLCG project, and the dedicated EMI software [2] has been developed to provide all the necessary components to create a distributed infrastructure. EMI plugins have been elaborated to optimize distributed computing for the structure, flows, and data processing procedures of a particular experiment. Some of these add-ons at some point began to have functionality that completely replaced individual EMI services.

Using the WLCG toolkit for solving problems seems natural and logical, but in practice a number of difficulties arise, the most striking of which is the lack of a qualified development team and specialists in general. A way out of this situation can be the creation of a grid system based on separate independent components taken from various projects, including the WLCG. For example, user authentication and authorization can be implemented using the virtual organization user management service from the EMI toolkit, and the distribution of experimental software can be organized using the CernVM File System [3]. However, the integration of services into a unified system entails the development of additional software, which will have to take over the functions of coordinating task launch and data transfer.

To solve this problem, developers have put into operation the software called Interware. Interware tools provide the ability to combine disparate computing resources, further developing grid technologies to a new level.

One of such software is the Production and Distributed Analysis (PanDA) platform [4, 5], which ensures the transparency of the process of processing, storing and managing data for applications with large information flows. The PanDA system has been developed to meet the ATLAS production and analysis requirements for a data-driven workload management system capable of operating at the LHC data processing scale. PanDA scalability has been demonstrated in ATLAS through the rapid increase in usage over the last decade. PanDA was designed to have the flexibility to adapt to emerging computing technologies in processing, storage, networking, and distributed computing middleware. This platform is actively developing towards combining various systems of distributed and parallel computing to work with clusters, cloud environments and supercomputers. This development makes it possible to expand and diversify the distributed infrastructure to solve the largest problems in various fields of science and business.

Another option is the Distributed Infrastructure with Remote Agent Control (DIRAC) Interware [6, 7], a distributed computing software environment that provides a complete solution for one or more user communities requiring access to distributed resources. DIRAC was originally developed for the LHCb experiment. In 2009, the core DIRAC development team decided to generalize the software to make it suitable for any user community. The results of this work allow offering DIRAC as a general-purpose distributed computing framework. DIRAC provides all the necessary components to build ad-hoc grid infrastructures interconnecting computing resources of different types (such as computational grids, clouds or clusters), allowing interoperability and simplifying interfaces. This allows speaking about the DIRAC interware. DIRAC creates a layer between users and resources, offering a common interface for a number of heterogeneous providers, seamlessly integrating them, ensuring compatibility with optimization, the transparent and reliable use of resources. The pilot load control system presented in the DIRAC project is now widely used in various grid infrastructures.

DIRAC and PanDA distributed as software packages, used by different experiments with pilot workload management, have the modular organization of the system, central task queues, web and command line interfaces, an accounting system, support for various data transfer protocols, the ability to use clusters and high-performance resources, to aggregate computing resources from different sources and nature into a unified system transparently for end users.

The distinctive features of the PanDA and DIRAC interware are presented in Table 1.

CONCLUSIONS

A study of existing solutions was carried out, and a particular model was selected.

DIRAC was chosen for the following reasons: it provides all the necessary functionality, including job and data management; easier deployment and maintenance of services compared to other platforms with similar functionality; the modular organization of the system allows one to quickly adapt and expand DIRAC for use in various tasks; presence of a fully functional and extensible web interface, with which one can access the platform's capabilities and administer the system in a regular browser; ability to aggregate computing resources from different sources and nature into a unified system, such as computer networks, clouds or clusters, transparently for end users; availability of interfaces to clouds based on Amazon EC2, OpenStack, OpenNebula. Also the use of the DIRAC Interware made it possible to successfully

Table 1. Interware features

	PanDA	DIRAC
Data management	Additional component Rusio; uses the rucio storage element	Storage element abstraction with client implementation for each access protocol (SRM, XROOTD, gfal2 based, etc.); each SE is seen by clients as a logical entity
File Catalog	Additional component Rusio	DIRAC File System
Authentication and authorization	X.509 certificates, VOMS	X.509 certificates, VOMS; can work without VOMS using DIRAC groups
Ability to use cloud computing infrastructures	GCI API (Google Compute Cloud) or Kubernetes (Amazon EC2)	Amazon EC2, OpenStack, OpenNebula

organize and put into operation a distributed information and computing infrastructure of the organizations of the JINR Member States, based on cloud resources deployed using the OpenNebula platform [8].

The next steps are to create a CERN-JINR-INP-KazNU data center using OpenNebula [9, 10] as a cloud platform and Ceph [11] as a storage for images of cloud virtual machines, as well as to integrate its resources into the JINR DICE for data storage, processing and analysis.

FUNDING

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (grant no. BR10965191 "Complex research in nuclear and radiation physics, high-energy physics and cosmology for development of the competitive technologies").

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. WLCG. https://wlcg.web.cern.ch/. Accessed November 2, 2021.
2. EMI. https://en.wikipedia.org/wiki/European_Middleware_Initiative. Accessed November 2, 2021.
3. CernVM-FS. https://cernvm.cern.ch/fs/. Accessed November 2, 2021.
4. PanDA. https://twiki.cern.ch/twiki/bin/view/PanDA/PanDA. Accessed November 2, 2021.
5. T. Maeno, A. Petrosyan, et al. (ATLAS Collab.), "Evolution of the ATLAS PanDA workload management system for exascale computational science," J. Phys.: Conf. Ser. **513**, 032062 (2014). https://doi.org/10.1088/1742-6596/513/3/032062
6. DIRAC. https://dirac.readthedocs.io/en/latest/. Accessed November 2, 2021.
7. V. Gergel, V. Korenkov, I. Pelevanyuk, M. Sapunov, A. Tsaregorodtsev, and P. Zrelov, "Hybrid distributed computing service based on the DIRAC interware," Commun. Comp. Inform. Sci. **706** (2017). https://doi.org/10.1007/978-3-319-57135-5_8
8. N. A. Balashov, N. A. Kutovskiy, A. N. Makhalkin, Y. Mazhitova, I. S. Pelevanyuk, and R. N. Semenov, "Distributed information and computing infrastructure of JINR member states' organizations," AIP Conf. Proc. **2377**, 040001 (2021). https://doi.org/10.1063/5.0063809
9. OpenNebula. http://opennebula.org. Accessed November 2, 2021.
10. N. Balashov, R. Kuchumov, N. Kutovskiy, I. Pelevanyuk, V. Petrunin, and A. Tsaregorodtsev, "Cloud integration within the DIRAC interware," CEUR Workshop Proc. **2507**, 256 (2019); CEUR Workshop Proc. **2507**, 260 (2019).
11. Ceph. https://ceph.io/en/. Accessed November 2, 2021.