Письма в ЭЧАЯ. 2016. Т. 13, № 5(203). С. 1084–1088

КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ В ФИЗИКЕ

BES-III DISTRIBUTED COMPUTING STATUS

S. D. Belov [a], *Z. Y. Deng* [b], *V. V. Korenkov* [a,c], *W. D. Li* [b], *T. Lin* [b], *Z. T. Ma* [b], *C. Nicholson* [b], *I. S. Pelevanyuk* [a,1], *B. Suo* [b], *V. V. Trofimov* [a], *A. U. Tsaregorodtsev* [d], *A. V. Uzhinskiy* [a], *T. Yan* [b], *X. F. Yan* [b], *X. M. Zhang* [b], *A. S. Zhemchugov* [a]

[a] Joint Institute for Nuclear Research, Dubna

[b] Institute of High Energy Physics, Chinese Academy of Sciences, Beijing

[c] Plekhanov Russian University of Economics, Moscow

[d] Particle Physics Center of Marseille, Marseille, France

The BES-III experiment at the Institute of High Energy Physics (Beijing, China) is aimed at the precision measurements in e^+e^- annihilation in the energy range from 2.0 to 4.6 GeV. The world's largest samples of J/ψ and ψ' events and unique samples of XYZ data have been already collected. The expected increase of the data volume in the coming years required a significant evolution of the computing model, namely, shift from a centralized data processing to a distributed one. This report summarizes the current design of the BES-III distributed computing system, some of key decisions and experience gained during two years of operations.

PACS: 89.20.Ff

INTRODUCTION

The distributed computing system has been expanded rapidly for the last 15 years [1]. Now these systems are capable to solve complex scientific problems in such areas as physics, biology, cosmology, astrophysics, etc. The distributed computing infrastructure BES-III experiment at BEPC-II collider in Beijing, China, can be considered as an example of successful organization of distributed data processing.

BES-III experiment started in 2009 after an extensive upgrade of Beijing Electron–Positron Collider (BEPC) and particle detector BES at the Institute of High Energy Physics of the Chinese Academy of Sciences in Beijing [2]. This project is run by more than 400 scientists from 52 institutes in 12 countries including the Joint Institute for Nuclear Research (Dubna, Russia). The main goal of the BES-III experiment is the study of properties of charmonium, charmed particles, and τ-leptons. After six years of operation, BES-III detector has become one of the best sources of experimental data in τ-charm domain. Despite the fact that BES-III is producing much smaller amount of data than experiments at the LHC [3], it also requires a

[1] E-mail: pelevanyuk@jinr.ru

substantial computing facility in order to process collecting data. It is in order of 1 PB of data collected per year now and the data volume is continuously growing. The lack of computing resources in IHEP became the reason to build a distributed computing infrastructure using the resources of institutes working within the BES-III collaboration [4].

The organization of the distributed computing infrastructure for the purposes of the BES-III experiment is a challenging task due to the following constraints:

1. Lack of grid experience among communities.
2. Storage elements not affordable to all the sites.
3. Weak network connection.
4. Lack of manpower to maintain sites.

Initially, all the data were kept and processed in the computing center of IHEP. The cluster has 4500 CPU cores, 3 PB disk storage, and 4 PB tape storage. One of the first tasks was the designing of the data access model and the data transfer model. The decision was taken to store and reconstruct experiment data from BES-III locally in IHEP. At the same time, CPU-consuming simulation and data analysis can be performed at the remote computing centers.

The following operation models were proposed depending on the capabilities and priorities of each site:

1. Monte Carlo simulation runs at remote sites. The resulting data are copied back to IHEP and then Monte Carlo reconstruction runs there. This operation model is appropriate for sites without storage element.

2. Monte Carlo simulation and reconstruction run at remote sites. This requires big Random Trigger dataset to be present on a remote site.

3. DSTs are copied from IHEP or other sites and then analyzed using local resources.

To realize these scenarios, the following software tools should be prepared: authorization and authentication, job management, data management and data transfers, distribution of the experiment software, information system and monitoring.

Of course, these entire tasks were solved before in frames of the WLCG project. A number of software have been already developed after the realization of European EDG, EGEE, and EMI projects: EMI/gLite [5]. Similar tasks were solved in the USA within OSG [6] project. However, these general-purpose software are not appropriate to cover specific requirements of LHC experiments. That is why all LHC experiments develop their own tools and components according to specifics of the experiments. BigPanDA [7], Rucio [8], DIRAC [9] are the examples of software like these. Some of these products at some point began to possess the functionality that supplements or completely replaces the functionality of EMI/gLite and OSG services.

However, it is difficult to use WLCG tools for the BES-III purposes. The main problem is that the WLCG software was designed to handle the enormous amount of data and jobs and it requires a high-level technical support all the time. WLCG tools do not consider some specific features of BES-III experiment. This will eventually require a qualified developers team in order to optimize or develop some of components.

An alternative to the use of complete WLCG installation is the using of a system based on a custom set of various components from WLCG project. Unfortunately, this decision would require efforts which are comparable with the use of WLCG tool itself. However, it turned out that most of required functionality is already provided or easily realized on the basis of DIRAC interware [9] (Distributed Infrastructure with Remote Agent Control). This is the

reason why the DIRAC system was chosen as a basis of the distributed computing system for the BES-III experiment.

DIRAC INTERWARE

DIRAC is an open-source system that provides a complete grid solution for both workload and data management, and it is designed to minimize the effort of local sites and system maintainers. It was initially developed for the purposes of LHCb experiment but evolved later in an independent and general-purpose platform for distributed computing. Besides, LHCb DIRAC software is already used at CTA [10], Belle-II [11], ILC, and some other projects [12].

DIRAC is written in Python and consists of a set of modules. It is quite easy to modify and develop new modules and components when needed. One of the key components of DIRAC is a web-interface which allows one to perform most administrative and user actions through the web page.

There are several general services in DIRAC, and they fully cover all the BES-III basic requirements. Some additional development was required in order to fulfill new tasks. BES-DIRAC is a set of extensions for BES specific modules. BES-DIRAC unites several modules: some of them are written from scratch, some of them just enhance the existing DIRAC functionality. Data Management System was upgraded. A Monitoring system and a Task Management System were developed for the purposes of BES-III.

USE OF CLOUDS

A useful feature of DIRAC is an interface to cloud resources, both private and commercial, such as Amazon EC2 [13], and volunteers computing platform BOINC [14]. Nowadays, BES-III experiment extensively uses the cloud resources based on OpenNebula and OpenStack. Approximately one third of all BES-III jobs is processed using cloud resources. The biggest cloud providers are IHEP and INFN-Torino, but a lot of tests were done in other institutes, in order to allow one to add cloud resources fast when it is necessary.

That is why preliminary study has been carried out to know the performance loss of simulation, reconstruction, and analysis jobs. The initial tests of KVM showed performance loss on the level of 10% that was caused mostly by CPU overhead. I/O only occupy 1–10% of total processing time and appeared to be insignificant.

Some elements have been studied on operation system level to reduce the loss: expose the properties of Advanced Vector Extensions and XSAVE on hosts to VMs on it to help speed up floating point operations, inherit NUMA feature from host, turn on disk preallocation, use virtio disk drive, raw format image, cache mode:writethrough. The above configurations have been tuned and better performance has been gained: tests have shown that performance loss has been reduced below 3% after optimizations.

MONITORING SYSTEM

The monitoring system is another example of custom development. The distributed infrastructure with a limited technical support requires a detailed centralized monitoring of the resource centers, all services, and the whole infrastructure, because the local administrators

cannot take part in this responsibility on themselves due to many reasons. This functionality is missing in DIRAC, so a custom monitoring system should be developed for BES-III experiment. During 2014–2015 the BES-III grid monitoring system has been developed and deployed. Main features of the monitoring system are robustness and modularity. Monitoring system uses three sources of information:

1. DIRAC internal database (for job monitoring).
2. Output of CLI commands (for network and data transfer monitoring).
3. Special jobs which run on a remote hosts and check the functionality there (usually checks accessibility of CVMFS and ability of work-node to solve simple physics task).

The system automatically detects and tests sites gathering all necessary information from DIRAC Configuration system. A number of tests were implemented to provide information about the most important metrics of the BES-III grid: network ping test, WMS test (sending simple job), simple BOSS job (full simulation of 50 events), combined test of CVMFS, environment and resources availability, queue information, CPU limit test, network, SE status, and so on. All the information available via the web page is integrated into the BES-III DIRAC web portal. For some tests there is a possibility to get historical information about their statuses. The system allows the operator to control sites reliability and to identify problematic nodes promptly.

MULTI-VO SUPPORT

Recently, inspired by the success of BES-DIRAC project, some other experiments carried out by IHEP are willing to use DIRAC for their solution to distributed computing. BES-III DIRAC installation was configured in order to accept jobs related to the Circular Electron Positron Collider (CEPC) [15] and the Jiangmen Underground Neutrino Observatory (JUNO) [16]. Each experiment's community naturally forms a virtual organization (VO). They have various heterogeneous computing and storage resources located at geographically distributed universities or institutions worldwide. They wish to integrate those resources and supply their VOs users, especially at an early stage of the experiments.

The distributed computing environment for the BES-III experiment at IHEP has been extended to support multi-VO usage scenario [17]. The middleware DIRAC has been made as a service for several experiments in IHEP.

SUMMARY

Currently, the BES-III distributed infrastructure includes 12 resource centers from the People's Republic of China, USA, Italy, and JINR (Russia), providing access to about 3000 CPU cores and 0.5 PB of disk space. More than 500 000 jobs were executed for the last year. A distributed data analysis has been successfully tested but for the moment it is not widely used yet. Currently, the BES-III distributed computing system works reliably and delivers a significant fraction of the computing power to process the experimental data.

The multi-VO support allows other experiments to use the BES-III computing infrastructure. New virtual organizations could joint in future. Using of multi-VO increases resource utilization and decreases total amount of work required for administrating several different installations of DIRAC.

The authors express their hope that the experience received by the developers of the computing infrastructure for BES-III will be interesting and useful to other projects of a comparable scale in which the distributed data processing is foreseen.

Acknowledgements. This work is supported in part by the joint RFBR–NSFC project No.14-07-91152 and NSFC project No. 11375221.

REFERENCES

1. *Foster I., Kesselman C., Tuecke S.* The Anatomy of the Grid: Enabling Scalable Virtual Organizations // Intern. J. High Perf. Comp. Appl. 2001. V. 15, No. 3. P. 200–222.
2. *BES-III Collab.* Design and Construction of the BES-III Detector // Nucl. Instr. Meth. A. 2010. V. 614. P. 345–399.
3. *Bird I.* Computing for the Large Hadron Collider // Ann. Rev. Nucl. Part. Sci. 2011. V. 61. P. 99–118; http://wlcg-public.web.cern.ch/.
4. *Zhang X. M. et al.* BES-III Production with Distributed Computing // J. Phys.: Conf. Ser. V. 664. P. 032036.
5. EMI. http://www.eu-emi.eu/.
6. OSG. http://www.opensciencegrid.org.
7. *Maeno T. et al. (the ATLAS Collab.).* Evolution of the ATLAS PanDA Workload Management System for Exascale Computational Science // J. Phys.: Conf. Ser. 2014. V. 513. P. 032062; http://pandawms.org/.
8. Rucio. http://rucio.cern.ch.
9. *Tsaregorodtsev A. et al.* DIRAC: A Community Grid Solution // J. Phys. Conf. Ser. 2008. V. 119. P. 062048; http://diracgrid.org/.
10. *Arrabito L. et al.* Application of the DIRAC Framework to CTA: First Evaluation // J. Phys.: Conf. Ser. V. 396. P. 032007.
11. *Graciani R. et al.* Belle-DIRAC Setup for Using Amazon Elastic Compute Cloud // J. Grid Comp. 2011. V. 9, No. 1. P. 65–79.
12. *Mendez V. et al.* Powering Distributed Applications with DIRAC Engine // The Intern. Symp. on Grids and Clouds (ISGC), Academia Sinica, Taipei, Taiwan, March 23–28, 2014.
13. Amazon EC2. http://aws.amazon.com/ru/ec2/.
14. BOINC. https://boinc.berkeley.edu/.
15. CEPC. http://cepc.ihep.ac.cn/.
16. JUNO. http://english.ihep.cas.cn/rs/fs/juno0815/.
17. *Yan T. et al.* Multi-VO Support in IHEP's Distributed Computing Environment // J. Phys.: Conf. Ser. V. 664. P. 062068.