Design and operation of the BES-III distributed computing system

Belov S. ¹, Suo B. ², Deng Z.Y. ², Korenkov V. ¹³, Li W.D. ², Lin T. ², Ma Z.T. ², Nicholson C. ², Pelevanyuk I. ¹, Trofimov V. ¹, Uzhinskiy A. ¹, Yan T. ², Yan X.F. ², Zhang X.M. ², Zhemchugov A. ¹
¹Joint Institute for Nuclear Research, Dubna, Russia
²Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, China
³Plekhanov Russian University of Economics, Moscow, Russia

Abstract
The paper summarizes general information, the latest results and development plans of the BES-III distributed computing system. The BES-III experiment at the Institute of High Energy Physics (Beijing, China) is aimed at the precision measurements in e⁺e⁻ annihilation at 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. To process the BES-III experiment’s data the distributed computing infrastructure based on DIRAC interware has been setup and became operational since 2013. The work was done in order to tune and enhance DIRAC itself for BES-III purposes: monitoring system and data transfer system were developed. Experience of the BES-III development team may be very useful for medium scale HEP experiments aiming to apply distributed data processing in their computing model.

Keywords: Distributed computing, grid, BES-III, DIRAC

1 Introduction

Distributed computing using heterogeneous resources developed rapidly during last 15 years from pioneer R&D [1] to a deployment of full-scale production systems capable to solve various sophisticated scientific problems.

The most remarkable among systems like these is a Worldwide LHC Computing Grid (WLCG) designed to process data acquired at the Large Hadron Collider at CERN. WLCG unified computing power of more than 170 research centers in 40 countries and already allowed to process about 100 PB of LHC data [2]. While distributed computing is mostly widespread in high-energy physics, it is not the only field of its application. There are many tasks with significant computing demands in astronomy
(data flow from the modern telescopes reaches tens of terabytes per day and comparable with the one from an average HEP experiment), biomedicine, genetics, pharmacology, seismology and others [3-7]. Computing hardware price reduction and big improvement in networking during recent decade made distributed computing available not only for large projects like LHC but also for research collaboration with much more modest budget, resources and manpower. The key decision for such collaboration is a choice of the middleware that will allow building the computing system and organizing distributed data processing with high efficiency and reliability at reasonably low cost. The BES-III experiment at the BEPC-II collider in China can be considered as a typical example of successful solution of this problem.

2 BES-III experiment

The BES-III experiment started operation in 2009 at the Institute of High Energy Physics of Chinese Academy of Sciences (Beijing, P.R.C.) after major upgrade of BEPC accelerator and BES spectrometer [8]. The experiment is run by an international collaboration of more than 400 scientists and engineers from 53 institutions in 13 countries. The main goal of the experiment is a precision study of charmonium, charmed particles and tau-leptons. After six years of operation BES-III detector proved out to be the best source of experimental data in tau-charm domain. BES-III installation is a source of much smaller volume of data than experiments on the Large Hadron Collider, nevertheless, the volume of the received experimental data is about one petabyte now and it is expected to increase several times within the next 5-7 years. While the computing center of IHEP CAS takes the main burden of data processing, the shortage of the computing resources there resulted in a research of the ways to involve external computing power. Distributed infrastructure based on computer centers of participating institutes appears to be the most attractive alternative.

The data processing model of BES-III experiment is typical for High Energy Physics as well as for some other areas. The experimental data containing information on events is recorded on tape storages during the data taking. The maximum data rate is about 40 MB/s. After collecting the data, reconstruction of events is carried out followed by the data analysis. Typical signal is much smaller than the background from the more frequent and well known physics processes. That is why the simulation of the detector response to the signal and background events is of a great importance for the data analysis. The amount of the simulated data usually is equal or larger than the one of real data. Simulation is done by Monte-Carlo method and it is extremely expensive operation demanding significant computing resources. There is a specific feature of BES-III simulation, namely the use of real random trigger data taken continuously along with the physics data to reproduce the noise due to the machine background.

3 Distributed processing of the BES-III data

Initially the data processing model assumed that all operations of reconstruction, simulation and the data analysis would be carried out on a computing cluster of IHEP CAS. The cluster has 4500 CPU cores, disk storage with a capacity of 3 PB and tape storage with a capacity of 4 PB. Priority task on the design stage of the distributed data processing system was to assess the data flow between the resource centers and to define the operations that would perform on remote resources. It is the networking connectivity what is the limiting factor for the large-scale data transfer from IHEP CAS to remote centers. So it was decided to continue to store and reconstruct real experimental data directly at local computer center. At the same time, simulation which requires significant CPU time, and data analysis can be executed at the remote resource centers. Finally, three main scenarios of the distributed data processing were elaborated:
1. Simulation runs at remote centers, followed by the transfer of simulated data to IHEP CAS. This scenario allows to integrate to the distributed computing system the resource centers without storage element (SE) and capable to keep temporarily the limited volume of data only.

2. Simulation runs at remote centers and simulated data are stored at remote SE.

3. Reconstructed physics data transferred from IHEP CAS to remote centers for analysis.

To be able to implement the scenarios listed above the distributed infrastructure for BES-III software should be able to cope with the following tasks:

a) authorization and authentication;
b) job management;
c) data management and data transfers;
d) distribution of the experiment software;
e) information system and monitoring.

Of course, all of these tasks were solved before within the WLCG project. The software of EMI/gLite [9], which provides all necessary components for the organization of the distributed infrastructure, was developed after the realization of European EDG, EGEE and EMI projects. Along with it the similar problems were solved in the USA within the OSG [10] project. However, the general-purpose software like these does not consider any specific features of LHC experiments. That is why all LHC experiments had to develop additional software that coordinates general purpose components according to specific data model of experiments and provides missing functionality. BigPanDA [11], Rucio [12], DIRAC [13] are the examples of this kind of software. Some of these products at some point began to possess the functionality that supplement or completely replace the functionality of EMI/gLite and OSG services.

The natural idea to use the WLCG tools for BES-III tasks is difficult to implement in practice because of several reasons. The main problem is that the WLCG software is designed to handle the extremely big data volumes and numerous jobs and requires rather high level of technical support. It is not easy for the BES-III experiment to find specialists in this software at all of the distributed resource centers. For an ordinary system administrator installation and configuration of a complete set of WLCG services is quite difficult task. The organization of the effective work of several resource centers without allocation of constantly operating highly skilled technical support team is almost impossible.

WLCG tools do not consider specific features of the BES-III experiment, the same as for LHC experiments. Therefore, development of the addition software on top of WLCG services which will optimize job and data management under specific needs of experiment is required. The experienced team of developers and time are necessary to solve this task.

Alternatively, to the deployment of the complete WLCG installation it is possible to organize a distributed processing system based on the independent components taken from various projects including WLCG. For example, authentication and authorization of users can be realized by VOMS [14] package from the EMI tools. To organize the distribution of the experiment software CVMFS [15] can be used. For job management one can deploy HTCondor [16], ARC [17] or BigPanDA [11]. For data management and data storage there are packages like gridFTP [18], BeStMan [19], dCache [20], Xrootd [21], StoRM [22]. Nevertheless, association of all these services into an integrated system will require the development of additional software that should take care of component interaction and would consider specific features of BES-III experiment. At first sight, this task is not much simpler, than the use of the WLCG tools. However, it turns out that the considerable part of the required functionality can be realized rather easily on the basis of DIRAC (Distributed Infrastructure with Remote Agent Control) interware [10]. This is the reason why DIRAC was chosen as a basis of the distributed processing system for BES-III experiment.
4 DIRAC as a key factor

DIRAC was developed in scope of the LHCb experiment at LHC but later it has evolved into an independent platform for distributed computing. DIRAC based on service architecture and can be adapted for different tasks. Besides LHCb DIRAC software is already used at CTA [23], Belle-II [24], ILC and some other projects [25].

DIRAC is written in Python and organized as a set of modules and extensions. DIRAC framework allows to develop new modules and integrate DIRAC with third-party application. Large part of administration and user tasks is available via highly configurable web-interface.

There are several general services in DIRAC.

1. Authorization and Authentication service based on VOMS from EMI/gLite tools
2. Information system based on DIRAC Configuration Service. BDII from EMI/gLite tools can be used in addition to IS.
3. Job management based on DIRAC Workload Management Service (WMS) with interfaces to computing elements like gLite WMS, CREAM CE, ARC CE and resource managers like Torque, Condor, SLURM, LSF. DIRAC WMS can be used to submit jobs without any resource managers through ssh-tunnel directly to the computing element.
4. DIRAC Data Management Service (DMS) based on DIRAC File Catalog, File Transfer Service (FTS) and storage elements with SRM interfaces. LFC from EMI/gLite tools can be used in addition to DFC.

An interesting feature of DIRAC is an interface to cloud resources, both private and commercial such as Amazon EC2, and volunteers computing platform BOINC. DIRAC user community is growing rapidly and current level of support from the DIRAC developers and the user's community is quite high.

5 Challenges and results

At the same time, DIRAC has a number of deficiencies that were identified while the distributed infrastructure for BES-III experiment has been developed. Although the basic installation of DIRAC is quite simple and does not require much effort, expert help is required when complex services like WMS are configured. DIRAC continues to evolve quickly enough so the frequent, every few months, upgrade of central services is desirable. While automated update procedure is in place, it often leads to a misconfiguration of the system and manual intervention is required. The same problem is evenly inherent in every distributed computing platform, though.

The DIRAC monitoring system is primarily focused on the control of the central system services. Distributed infrastructure with limited technical support requires detailed centralized monitoring of the resource centers, all services and the whole infrastructure. This functionality is missing in DIRAC so a custom monitoring system had to be developed for BES-III experiment. While 2014-2015 the BES-III grid monitoring system has been developed and deployed by JINR team. Main features of the monitoring system is robustness and modularity. Monitoring system use three sources of information: DIRAC internal database (for job monitoring), output of CLI commands (for network and data transfer monitoring) and special jobs which runs on a remote hosts and checks the functionality (usually checks accessibility of CVMFS and ability of work-nodes to solve simple physics task). The system gathers all necessary information from DIRAC Configuration system and automatically detects and tests sites. Number of tests is implemented to provide an 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, CPUtilimit test, network, SE status and so on. All the information available via the web page integrated into the BES-III DIRAC web portal. For some tests there are possibility to get historical view of status changes. The system allows to control sites reliability and identify problematic nodes.
DIRAC data management system can effectively work with the individual files but it is not sufficient, when the number of files becomes too large and the data has complex structure. In this case logical collection of files in datasets allowing group operation like copy, move, delete, etc are very convenient. The latest DIRAC version has option to consolidate the files into dataset, but the tools to work with them are scarce. The DIRAC file catalog need some optimization and tuning to work with datasets [25]. Data integrity control is quite important mechanism that is not developed enough in current DIRAC version.

Despite the deficiencies described above DIRAC interware was successfully used to organize distributed computing of the BES-III experiment. Extended version of GANGA which includes modules to run BES-III specific jobs is used as a distributed job submission and job management tool [26]. Several key services were developed from scratch, like data transfer service [27] and task management service. The former allows copying of data over network via SRM operations or FTS service from EMI/gLite toolkit. The latter allows combining of individual jobs into tasks thus simplifying the job management. A system to monitor status of remote centers has been developed and integrated into DIRAC platform. Automatic distribution of experiment's software is done using CVMFS. A lightweight zero-configuration version of the storage element StoRM has been developed to automate installation of SE at the remote resource centers. VMDIRAC extension was installed and configured to submit BES-III jobs to a cloud infrastructure based on OpenStack and OpenNebula platforms.

6 Summary and future plans

Current BES-III distributed infrastructure includes 12 resource centers from 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 400 000 simulation jobs were executed since 2013 when a prototype of the system was deployed. 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 significant fraction of the computing power to process the experimental data.

Developers team continues to improve monitoring and integration of the job management and data management systems. The team works closely with DIRAC development group in order to provide reliable high-quality tools for monitoring. There are plans to merge functionality of BES-III DIRAC and DIRAC.

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

Acknowledgement

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

References

8. EMI: http://www.eu-emi.eu/
9. OSG: http://www.opensciencegrid.org
11. Rucio: http://rucio.cern.ch
12. VOMS: https://github.com/italiangrid/voms
13. CVMFS: http://cernvm.cern.ch/portal/filesystem
15. ARC: http://www.nordugrid.org/arc/
17. BeStMan:https://twiki.opensciencegrid.org/bin/view/Documentation/Release3/InstallOSGBes
tmanSE
18. dCache: http://www.dcache.org
19. Xrootd: http://www.xrootd.org
20. StoRM: http://italiangrid.github.io/storm/