КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 469−473

Ки&М

СЕКЦИОННЫЕ ДОКЛАДЫ

УДК: 004.75, 004.052.2, 004.052.32

BES-III distributed computing status

S. Belov[1], Z. Deng[2], W. Li[2], T. Lin[2], I. Pelevanyuk[1], V. Trofimov[1], A. Uzhinskiy[1], T. Yan[2], X. Yan[2], G. Zhang[2], X. Zhao[2], X. Zhang[2], A. Zhemchugov[1]

[1] Joint institute for nuclear researches, Laboratory of Information Technologies, Joliot-Curie, 6, Moscow reg., Dubna, 141980, Russia
[2] Institute of High Energy Physics, Chinese Academy of Sciences, 19B YuquanLu, Shijingshan District, Beijing, 100049, China

Received September 30, 2014

The BES-III experiment at the IHEP CAS, Beijing, is running at the high-luminosity e+e- collider BEPC-II to study physics of charm quarks and tau leptons. The world largest samples of J/psi and psi' events are already collected, a number of unique data samples in the energy range 2.5–4.6 GeV have been taken. The data volume is expected to increase by an order of magnitude in the coming years. This requires to move from a centralized computing system to a distributed computing environment, thus allowing the use of computing resources from remote sites — members of the BES-III Collaboration. In this report the general information, latest results and development plans of the BES-III distributed computing system are presented.

Keywords: BES-III, distributed computing, grid systems, DIRAC Interware, data processing

Распределенные вычисления для эксперимента BES-III

С. Белов[1], Ц. Ден[2], В. Ли[2], Т. Линь[2], И. Пелеванюк[1], В. Трофимов[1], А. Ужинский[1], Т. Янь[2], С. Янь[2], Г. Чжан[2], С. Чжао[2], С. Чжан[2], А. Жемчугов[1]

[1] Лаборатория информационных технологий, Объединенный институт ядерных исследований Россия, 141980, г. Дубна, ул. Жолио-Кюри, д. 6
[2]Институт физики высоких энергий, Китайской академии наук, Китай, 100049, г. Пекин, ЮкуаньЛу 19Б

В 2009 году в Пекине заработал детектор BES-III (Beijing Spectrometer) [1] ускорителя BEPC-II (Beijing Electron–Positron Collider). Запущенный еще в 1989 году BEPC за время своей работы предоставил данные для целого ряда открытий в области физики очарованных частиц. В свою очередь на BES-III удалось получить крупнейшие наборы данных для J/ ψ, ψ ' и ψ частиц при энергии ускорителя 2.5–4.6 ГэВ. Объемы данных с эксперимента (более 1 ПБ) достаточно велики, чтобы задуматься об их распределенной обработке. В данной статье представлена общая информация, результаты и планы развития проекта распределенной обработки данных эксперимента BES-III

Ключевые слова: BES-III, распределённый компьютинг, грид системы, DIRAC Interware, обработка данных

This work is supported in part by the joint RFBR-NSFC project No.14-07-91152 and NSFC projects 11179020 and 11375221.

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 3, pp. 469–473.

Introduction

The BES-III experiment at the BEPC-II collider located in the Institute of High Energy Physics (Beijing, China) started data taking in 2009 after a major upgrade of already existing accelerator BEPC and detector BES-II [Ablikim M. et al., 2010]. The experiment is run by an international collaboration of more than 400 members from 52 institutes in 12 countries from around the world. The main physics goals of the experiment are precision measurements in the tau-charm domain. The BES-III experiment has already taken the world's largest data samples of J/ψ ($1.2 \cdot 10^9$ events) and ψ' decays ($0.3 \cdot 10^9$ events), as well as a large amount of ψ (3770) data and a number of unique samples of data in the energy range 2.5–4.6 GeV. The total volume of experimental data is already about 0.9 PB, of which about 300 TB is event summary data for physics analysis (DSTs). This amount of data is rather large to be processed in a single computing center. Use of distributed computing looks like an attractive option to increase the computing resources of the experiment and to speed up the data analysis.

The BES-III computing model

Raw experimental data are taken from the BES-III detector and stored to the tape storage managed by CASTOR. The maximum data rate is about 40 MB/s. After reconstruction DSTs are produced and used in further physics analysis. DSTs are stored in a disk pool managed by Lustre and can be accessed only from internal IHEP network. The total amount of DSTs currently is about 300 TB. Both inclusive and exclusive Monte-Carlo simulation (MC) is made for each data sample as well. Experimental data taken with a random trigger are used in the simulation to reproduce noise and machine background individually for each run. The total amount of MC DSTs is more than 50 TB now. The BES-III offline software is based on the Gaudi framework [CERN Web site] and runs on Scientific Linux OS.

The BES-III distributed computing system

Grid computing became a routine tool for data processing in high energy physics after successful deployment in the LHC experiments. However, the main difficulty for a widespread use of the grid tools developed in the WLCG project is their large scale and complexity. It is not easy to adapt the distributed computing software that was designed for LHC experiments for use in a medium scale experiment. Limited manpower makes it even more difficult to maintain. For BES-III, the situation is even worse, because very few participating sites are members of WLCG. As a result there are few experienced grid users and developers and there is lack of grid computing infrastructure. Another problem is that network connectivity between institutes participating in the BES-III experiment is typically low. All these considerations motivate the following approach to the BES-III distributed computing model.

It is assumed that remote sites participate only in MC production and physics analysis, while all reconstruction of real experimental data is done at IHEP. Three operation models are considered, depending on the capabilities and priorities of each site:

a) MC simulation runs at remote sites. The resulting data are copied back to IHEP and then MC reconstruction runs there. (This model is convenient for sites with no SE or with only a small one);

b) MC simulation and reconstruction runs at remote sites. The resulting data are copied back to IHEP;

c) DSTs are copied from IHEP and other sites and analyzed using local resources.

Distributed analysis is postponed for later stage of the project.

BES-III grid solution

The DIRAC (Distributed Infrastructure with Remote Agent Control) software [DIRAC Web site] is chosen to be the main BES-III grid solution. DIRAC was designed originally for the LHCb experi-

ment, but with time it evolved into a generic product which could be used to access distributed computing resources in various communities of users. The main reasons why DIRAC is suitable for BES-III needs are the following:

- DIRAC provides all the necessary components to build ad-hoc distributed computing infrastructures interconnecting resources of different types, allowing interoperability and simplifying interfaces.
- DIRAC provides job management, data management, information system, monitoring, security system.
- DIRAC is rather easy to install, configure and maintain.
- DIRAC supports grids based on different middleware (gLite, EGI, VDT, ARC, etc).
- DIRAC requires no grid middleware installation on site. Remote hosts can be accessed through an SSH tunnel and the application runs via local resource management system.

DIRAC is adopted as a core part of the BES-III grid system. A production installation of DIRAC has been set up for BES-III, with nine remote sites and the DIRAC server running at the IHEP central site in Beijing. DIRAC job submission was tuned to fit to BES-III needs. Computing elements like gLite-CREAM and SSH-CE are used on the BES-III sites.

Interesting feature of DIRAC job management system is a capability to use cloud resources. Cloud computing becomes very popular technology nowadays, thanks to its flexibility and universality. For BES-III community cloud computing looks attractive because it allows to compensate peak overrun of resources and to use existing resources more effectively. VMDIRAC is an extension for DIRAC which allows to submit jobs to the clouds. Two servers for OpenStack and for OpenNebula have been set up at IHEP. Virtual resources from University of Turin, JINR and Soochow University are used for BES-III job processing.

Data management system

Data management in BES-III includes data storage, data transfer and catalogs. Main data storage in IHEP is managed by Lustre. and available from internal network only. To access data from outside a bridge connecting Lustre storage and the external network is needed. The problem was solved by introducing an extension to the dCache system, which allows to connect Lustre storage to the dCache server at IHEP as a disk pool and to synchronize the namespace. Using this bridge the data can be reached from internal IHEP network using all dCache-supported protocols. The BES-III transfer system based on FTS, provides reliable data transfer between IHEP and remote sites via both SRM and GridFTP protocols. SRM-capable storage elements dCache, Bestman and Storm are used at the BES-III sites. Catalogs are based on the DIRAC FileCatalog (DFC) with the MySQL backend. Variety of physics tasks of the experiment requires high granularity of data. Dynamic datasets based on metadata queries are used as containers of files. DIRAC provides a mechanism of the dataset management but more functionality needs to be developed to meet the experiment needs.

The BES-III Monitoring

A monitoring system is necessary to ensure the reliable data production using the BES-III distributed computing and to simplify maintenance and troubleshooting of the system. Regretfully, there is no low-level or high-level monitoring system provided by DIRAC, except the monitoring of DIRAC services themselves. Information about failure of the jobs or unavailability of the resources appears after several days after the event. Development of the monitoring system is required to provide enough input to decrease the number of failed jobs, to understand the failure reasons, to show system malfunction before failure occurs, to control overall status of the grid, to optimize data transfers, to check storage availability etc. By the end of 2013 the first prototype of BES-III grid monitoring system has been

developed and deployed. Simple jobs are submitted by a monitoring agent hourly via DIRAC job management system, both running the standard BES-III applications and providing system tests followed by sending the information back to the system. This information is collected, analyzed and available via the web page integrated into the BES-III DIRAC web portal. The 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, CPUlimit test, network, and SE latency test. Analysis includes site reliability estimation and identification of problematic hosts. An example of the monitoring page is shown in Fig. 1.

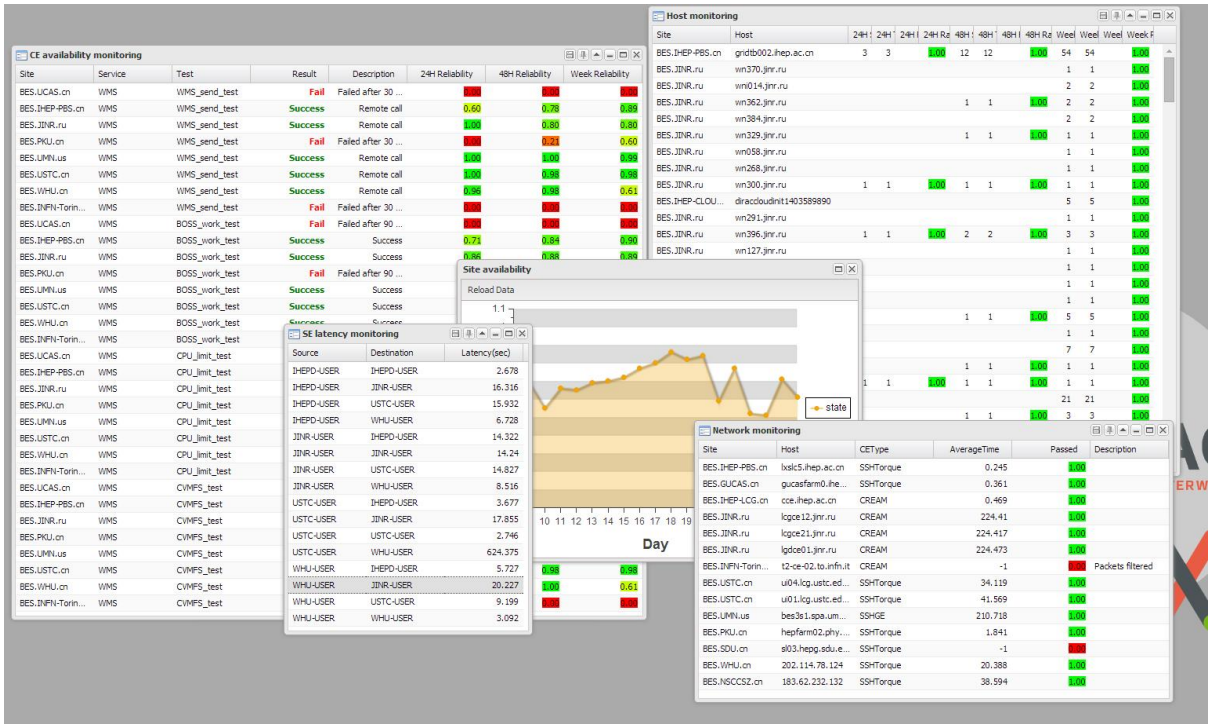

Fig. 1. Examples of the reports at BES-III monitoring system

Currently the existing monitoring system is being moved to be part of the DIRAC Resource Status System (RSS) service is in progress. RSS service is a new part of DIRAC that provides scheduling mechanism to collect and keep information about computing resources and to take decisions on use of these resources based on their availability. This mechanism suits very well to carry out functional tests and to perform the monitoring Being implemented in scope of the RSS framework the monitoring system can be used not only for BES-III experiment but as a generic solution for all DIRAC projects.

Summary

The BES-III distributed computing is operational since 2013. Since then more than 350000 jobs were executed and about 250 TB of disk space are managed by the system. While the basic infrastructure is built and the system is already put in production, more development is necessary to improve the dataset management, to integrate job management and data management systems, to implement fully functional monitoring & accounting system and to use clouds resources effectively. Experience and approaches to the organization of distributed computing gained by the BES-III collaboration may be interesting and useful for other medium scale experiments willing to use grid for their data processing.

References

Ablikim M. et al. “Design and Construction of the BESIII Detector’’ Nucl. Instrum. Meth. A614 (2010) 345–399.
CERN Web site. http://www.cern.ch/gaudi
DIRAC Web site. http://diracgrid.org/