Journal of Physics: Conference Series **664** (2015) 032036 doi:10.1088/1742-6596/664/3/032036

BESIII production with distributed computing

X M Zhang[1], T Yan[1], X H Zhao[1], Z T Ma[1], X F Yan[3], T Lin[1], Z Y Deng[1], W D Li[1], S Belov[2], I Pelevanyuk[2], A Zhemchugov[2], H Cai[3]

1 Institute of High Energy Physics, 19B Yuquan Road, Beijing 100049, R. P. China
2 Joint Institute for Nuclear Research, Joliot-Curie 6, 141980 Dubna, Moscow region, Russia
3 Department of Physics, Wuhan University, 299 Bayi Road, Wuhan 430072, P. R. China
E-mail: zhangxm@ihep.ac.cn

Abstract. Distributed computing is necessary nowadays for high energy physics experiments to organize heterogeneous computing resources all over the world to process enormous amounts of data. The BESIII experiment in China, has established its own distributed computing system, based on DIRAC, as a supplement to local clusters, collecting cluster, grid, desktop and cloud resources from collaborating member institutes around the world. The system consists of workload management and data management to deal with the BESIII Monte Carlo production workflow in a distributed environment. A dataset-based data transfer system has been developed to support data movements among sites. File and metadata management tools and a job submission frontend have been developed to provide a virtual layer for BESIII physicists to use distributed resources. Moreover, the paper shows the experience to cope with lack of grid experience and low manpower among the BESIII community.

1. Introduction

The BESIII experiment [1], at the BEPCII accelerator at the Institute for High Energy Physics (IHEP), Beijing, studies electron-positron collisions in the tau-charm threshold region. The data volume of BESIII, which has aggregated about 3 PB of data over the last 5 years, is 100 times larger than that of BESII. With the increasing data volume, the local cluster in IHEP eventually can't meet all the resource requirements of data processing. Peek needs for resource was happening frequently after data taking periods and before important physics conferences as shown in figure 1. Therefore, it is necessary for the BESIII experiment to explore opportunistic resources for supplement. The available resources to explore can come from local resources from BESIII collaboration members, cloud and volunteering computing.

Grid technology is still not widely used among BESIII collaboration especially in China although it is quite mature in European countries. T2_CN_Beijing is the only WLCG T2 site in China to support CMS and Atlas experiments. Only few collaboration members have experience to deploy a grid site and lack of man power for good maintenance is a common situation. Therefore, it is difficult for us to adopt the same rules from WLCG to integrate resources completely based on grid technology. It is important to avoid too much effort from the local group with central control and management.

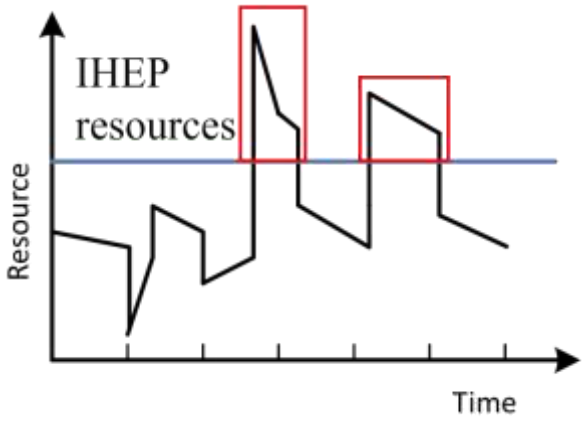

Figure 1 BESIII resource requirements with time

2. BESIII distributed computing

Based on the manpower situation and grid environment mentioned above, the design principle of BESIII distributed computing system is:

- Make it as simple as possible for sites with various computing facility to join, especially local clusters since 80% of BESIII collaborations are using local clusters.
- The whole system need to be set up, maintained and supported in an easy way. Take full use of existing and mature technology and tools wherever possible.
- Make distributed environment transparent to physics users and provide unique and friendly user interface.
- The system should well meet the requirements of BESIII data processing workflow.

2.1 Computing model

Currently BESIII distributed computing is mainly used for the Monte Carlo production, and will be extended to support physics analysis soon. The Monte Carlo production includes official production from the central production group and private production from BESIII individual users. The complete Monte Carlo production includes two processes -- simulation and reconstruction. The random trigger data is necessary input for the reconstruction process. Therefore, random trigger data need to be accessible to all the sites. But not all the sites can provide SE to hold random trigger data. Therefore, three types of sites have been introduced in the BESIII distributed computing model: central site, "big" site and "small" site, as shown in figure 2.

Data is being taken at IHEP. IHEP acts as a central site, mainly responsible for the processing and storage of all the raw data as well as bulk reconstruction and analysis. The central sites have central storage element (SE) storing all the raw data, random trigger data and DST data. Since "big" sites can provide SE, the random trigger and part of DST data can be transferred from central SE to these sites for the complete Monte Carlo jobs and physics analysis jobs. The "small" site, which don't have SE, can do simulation jobs, also do reconstruction jobs in which random trigger data is accessed from "big" site and central site depending on network situation. All the output files from simulation and reconstruction jobs in "big" sites and "small" sites are transferred back to the central SE directly after jobs finish.

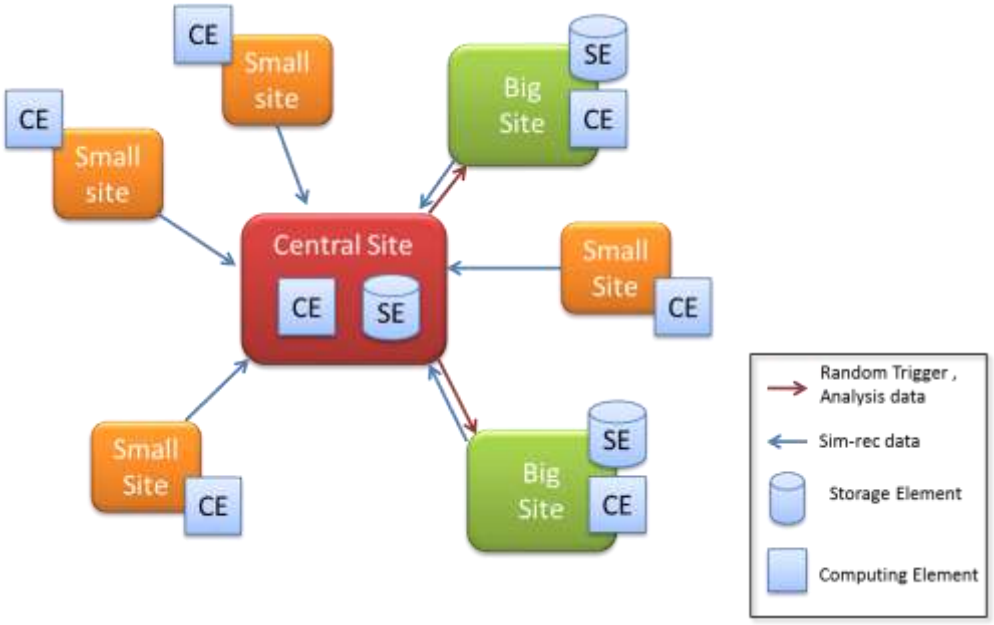

Figure 2 Computing model of BESIII distributed computing

2.2 Workload management

Resource types provided by BESIII community are quite different. Local cluster is the most common resource type and the rest 20% of total resources comes from private cloud and grid. To take advantage of these possible distributed resources, while hiding complexity and heterogeneity from end users, the workload management is built up based on DIRAC (Distributed Infrastructure with Remote Agent Control) [2] , which provides a middle layer between end users and heterogeneous resources. As a pilot-based distributed resources management framework, DIRAC can be easily extended to support a new resource type. Currently the DIRAC system can provide a single scheduling mechanism to distribute jobs to clusters, clouds and grids, even desktop resources, as shown in figure 3. Also DIRAC provides other functionalities as a general purpose distributed computing framework, which is necessary for BESIII distributed computing, including accounting, user and group management, resource management, priority control and job monitoring.

To make massive job submission and management convenient, the BESIII distributed computing job frontend called GangaBOSS has been developed based on Ganga [3]. Closely combined with the BESIII software, the frontend can automatically take care of BESIII job life cycle for BESIII physics users in distributed computing environment, including splitting, submissions, run-time scripts generating, log retrieval, rescheduling, status monitoring, workflow arrangement, dataset query and register. CVMFS (CERNVM File System) is adopted to deploy the BESIII offline software system (BOSS) among remote sites. Two CVMFS servers have been set up. One is located at CERN, and the other is located in IHEP. The collaboration member who can't have a good connection to Europe can use the IHEP CVMFS server instead.

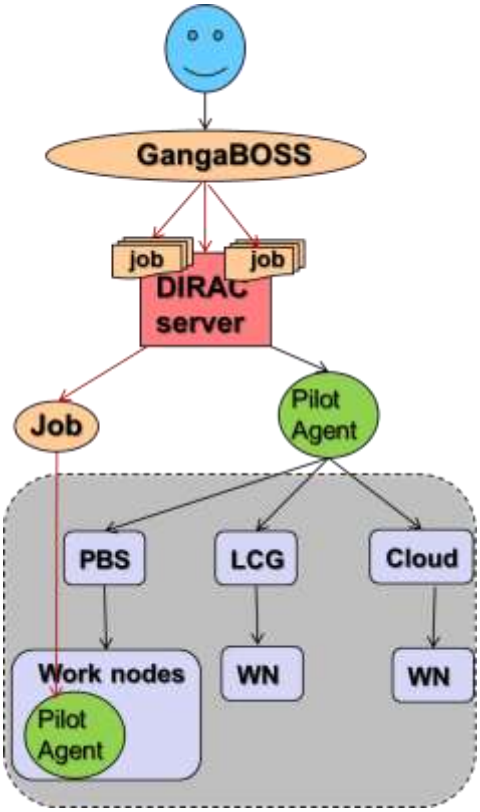

Figure 3 The architecture of workload management system

2.3 Data Management

To manage experiment data files in distributed environment and provide a unique interface for users and administrators for query, a book-keeping system called BADGER (BESIII Advanced Data manaGER) is established based on DIRAC File Catalogue (DFC) [4], whose data structure are organized based on BESIII physics requirements. BADGER includes three Catalogues:

- Replica Catalogue - maps logical file names (LFNs) to physical file names (PFNs) at different sites
- Metadata Catalogue - maps LFNs to file-level metadata
- Dataset Catalogue - maps between datasets and LFNs

With proper conditions given to the catalog, the user can easily access to a batch of files required called dataset for processing or transferring. According to the requirements from physics users, the support of static dataset which is queried by dataset names is also added.

The central storage is to share DST and random trigger Data with sites, accept and save MC output from remote sites. All the experiment data are located in the local file system Lustre. To expose the data to remote sites, the copy from Lustre to the central SE is not automatic if no combinations between Lustre and SE were done. To make exchange of local data and grid data more efficient and convenient, two solutions have been studied and compared. One is based on dCache, metadata synchronization and data movement between Cache pool and Lustre has been developed to enable such combination. The other is based on StoRM (Storage Resource Manager) [5], which can provide grid interface to POSIX file system. The comparison result is shown in Table 1. We can see the transfer speed gained from both solutions is similar, but the solution of StoRM is simpler to be realized and maintained.

Table 1 Comparison of the dCache and StoRM solutions

	dCache + Lustre	StoRM + Lustre
Hardware	Need extra disk array as cache pool	Just mount Lustre
Extra supports for combination	Need metadata synchronization between dCache and Lustre	no
SE Transfer speed	83.5 MB/s	80.9 MB/s
Data movement	Cache pool <-> Lustre	no
Security	Grid authentication	Grid authentication

The dataset-based transfer system has been developed to support the transfers of random trigger data and DST data between the central SE and the remote SE. As shown in figure 4, the transfer system is composed of 4 parts: request service to accept transfer request, Database to store status and log of transfers, transfer agent to arrange transfers, and the works to control real transfers. The maximum speed can reach 10TB per day with the bandwidth of 2 Gigabit per second.

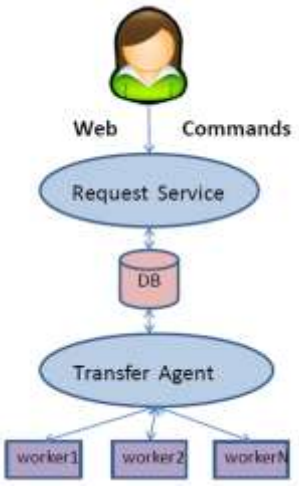

Figure 4 Architecture of data transfer system

2.4 Cloud integration

To integrate cloud resources, the extension VMDIRAC[6] is added to the original DIRAC set-up. In VMDIRAC extension, the VM scheduler as shown in figure 5 is introduced to support elastic startup and shutdown of VMs with the job requests in central queue of DIRAC. That is to say, virtual machines are created if there are job requests, and destroyed if there are no more jobs running. The usage of cloud resources is transparent to the end user.

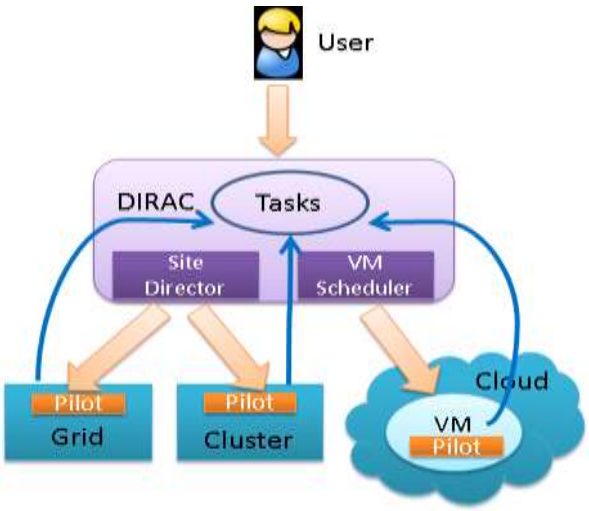

Figure 5 Cloud integration based on pilot scheme

3. Monte Carlo production with the system

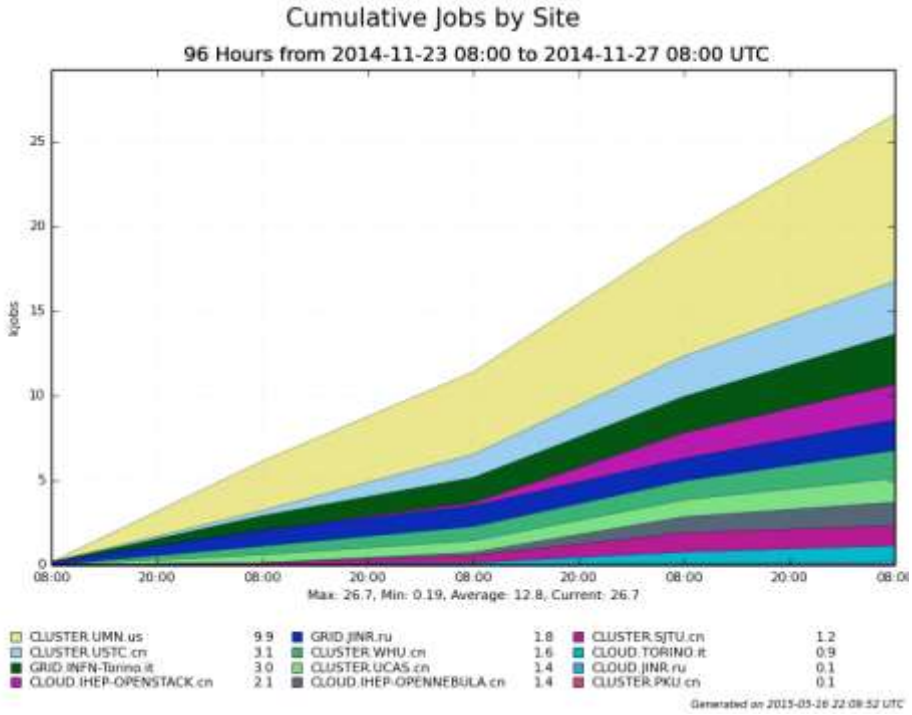

Figure 6 Cumulative jobs by site

In 2012, the BESIII distributed computing system was being built up. The BESIII distributed computing system has integrated more than 3000 CPU cores across more than 10 sites, with about 400TB storage. Three large-scale official production tasks have been completed, with more than 150,000 jobs completed successfully, 1550 Million Psi(3770) events and 800 Million Jpsi Inclusive events produced. The figure 6 and 7 shows one recent production task, which produced 620 Million bhabha events. About 26000 jobs are completed with over 98% success rate. More than 10 sites have joined the production.

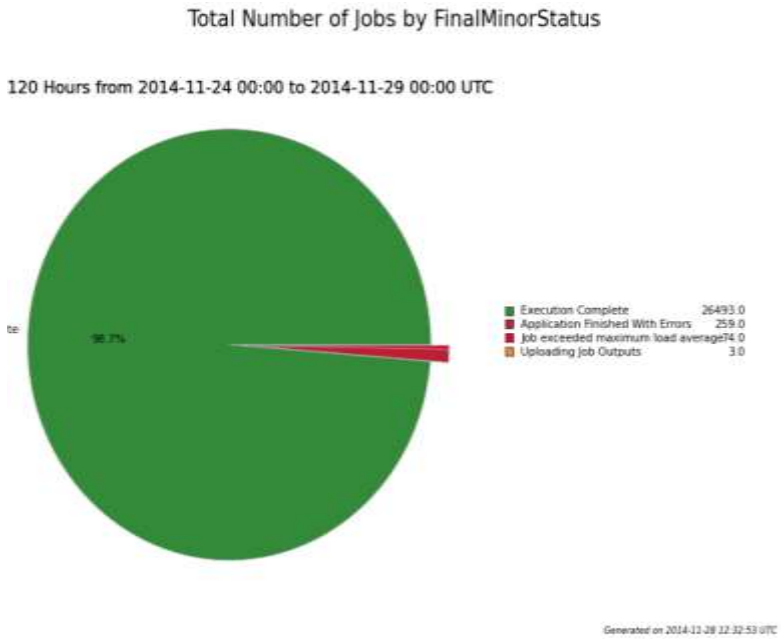

Figure 7 jobs failure rate

4. Conclusions and Future work

The basic elements of BESIII distributed computing system including data management and workload management are ready. The system is working well during current production. But to cope with lack of manpower for maintenance in sites, more measures still need to take to cope with the instabilities of sites. Monitoring and automatic measures based on monitoring is a key part in the near future to make the whole system more robust. The access way of random trigger data in the BESIII Monte Carlo reconstruction process need to be improved since current system is difficult to meet the requirements of a large amount of random trigger files to be accessed in one job as what local clusters normally do. Cloud storage is being explored to hold common data among sites to reduce data movements.

Acknowledgments

The authors would like to thank Andrei Tsaregorodtsev and Ricardo Graciani for their help on DIRAC configurations and many discussions, Victor Mendez and Victor Fernandez for their help on VMDIRAC, and colleagues at the IHEP computing centre for their support. This work was funded by the National Natural Science Foundation of China (NSFC) under grant no. 11375221, Joint Funds of NSFC under grant no. U1232201 and U1232109, and joint RFBR-NSFC project no.14-07-91152.

References

[1] Design and Construction of the BESIII Detector, Nucl.Instrum. Meth. A614 (2010)345-399
[2] A Tsaregorodtsev *et al*. DIRAC: a community Grid solution. 2008 *J. Phys.: Conf. Ser.* **119** 062048
[3] Harrison, K; Lavrijsen, WTLP; Tuli, CE; Mato, P; Soroko, A; Tan, CL; Brook, NH; Jones, RWL. Ganga: A user grid interface for ATLAS and LHCb. Conference for Computing in High-Energy and Nuclear Physics (CHEP 03). 2004. p. 1 - 9.
[4] A Tsaregorodtsev and S Poss. DIRAC File Replica and Metadata Catalog. 2012 J. Phys.: Conf. Ser. 396 032108
[5] https://italiangrid.github.io/storm/[

[6]Tom Fifield *et al*. Integration of cloud, grid and local cluster resources with DIRAC. 2011 *J. Phys.: Conf. Ser.* **331** 062009