Experience of BESIII data production with local cluster and distributed computing model

Z.Y. DENG[1], W.D. LI[1], L. LIN[2], H.M. LIU[1], C. NICHOLSON[3], Y.Z. SUN[1], X.M. ZHANG[1], A. ZHEMCHUGOV[4]

[1] Institute of High Energy Physics, China
[2] Soochow University, China
[3] Graduate University of Chinese Academy of Sciences
[4] Joint Institute for Nuclear Research, Russia

E-mail: dengzy@ihep.ac.cn

Abstract. The BES III detector is a new spectrometer which works on the upgraded high-luminosity collider, BEPCII. The BES III experiment studies physics in the tau-charm energy region from 2 GeV to 4.6 GeV . From 2009 to 2011, BEPCII has produced 106M ψ(2S) events, 225M J/ψ events , 2.8 fb^{-1} ψ(3770) data, and 500 pb^{-1} data at 4.01 GeV. All the data samples were processed successfully and many important physics results have been achieved based on these samples. Doing data production correctly and efficiently with limited CPU and storage resources is a big challenge. This paper will describe the implementation of the experiment-specific data production for BESIII in detail, including data calibration with event-level parallel computing model, data reconstruction, inclusive Monte Carlo generation, random trigger background mixing and multi-stream data skimming. Now, with the data sample increasing rapidly, there is a growing demand to move from solely using a local cluster to a more distributed computing model. A distributed computing environment is being set up and expected to go into production use in 2012. The experience of BESIII data production, both with a local cluster and with a distributed computing model, is presented here.

1. Introduction

BESIII [1] (Beijing Electron Spectrometer) is designed to study physics in the tau-charm energy region utilizing the new high luminosity BEPCII(the Beijing Electron-Positron Collider) double ring electron-positron collider. The expected data samples in a calendar year of the BESIII are summarized in Table 1.

Table 1. Expected data samples in a calendar year

States	Energy (GeV)	Peak luminosity (10^{33} $cm^{-2}s^{-1}$)	Physics cross section (nb)	Events/year
J/ψ	3.097	0.6	3,400	1×10^{10}
ψ'	3.686	1.0	640	3×10^{9}
$\tau^+\tau^-$	3.670	1.0	2.4	1.2×10^{7}
D^0D^0	3.770	1.0	3.6	1.8×10^{6}
D^+D^-	3.770	1.0	2.8	1.4×10^{6}
D_sD_s	4.030	0.6	0.32	1×10^{6}

1.1. BESIII Offline Software

The BESIII Offline Software System (BOSS) has been developed using the C++ language and object-oriented techniques within the Scientific Linux CERN operating system. Software production and configuration management is facilitated using the Configuration Management Tool, CMT [2]. The BOSS framework is based on Gaudi [3], which provides standard interfaces for the common software components necessary for data processing and analysis: simulation, calibration, reconstruction and analysis algorithms. Software framework (Figure 1.) employs Gaudi's event data service as the data manager. All the algorithms should not use directly the data objects in the persistency store but instead register or retrieve data from transient data store with the event data service. Converters are responsible for the conversion between persistent data and transient data. By the end of 2010, the BOSS system has successfully been migrated from SLC4 ia32 to SLC5 x86-64 operating system. It took about one year to finish this work. Most of the software packages in BOSS need to be modified due to the update of GCC from 3.4.6 to 4.3.2, and the update of Gaudi from v19r4 to v21r6. Many bugs were fixed to make sure the stability and correctness of the software system. Most of the bugs were caused by illegal using of pointer.

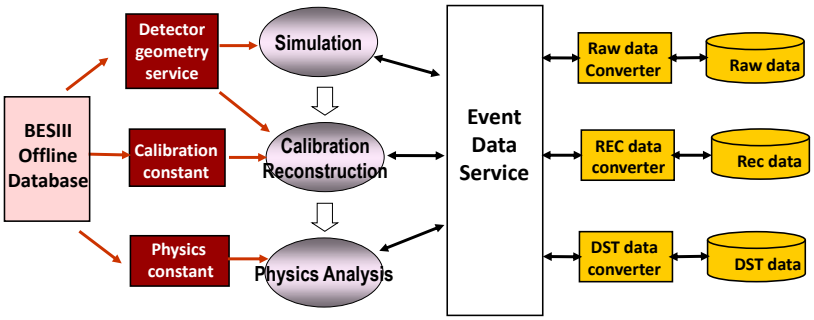

Figure 1. BOSS framework

1.2. BESIII data flow

The data flow of BESIII data processing is shown in Figure 2. The raw data from BESIII detector online DAQ system is saved on Castor[4] system. Offline data managers copy the raw data from Castor to luster[5] file system daily. The raw data includes the original full raw data, random trigger data files, and Bhabha and Dimu data samples which include specific events for detector calibration. Random trigger events are used in Monte Carlo(MC) background mixing. After calibration, the raw data files are reconstructed using the calibration constants. The output of reconstruction job is a DST(Data Summary Tape) file. Physics users skim the DST files with different criteria, 2 prongs, 4 prongs, 6 prongs, DTag, etc. The MC raw data from simulation job will be mixed with random trigger events, and then reconstructed to DST data. Both MC raw data and DST data are in ROOT [6] format.

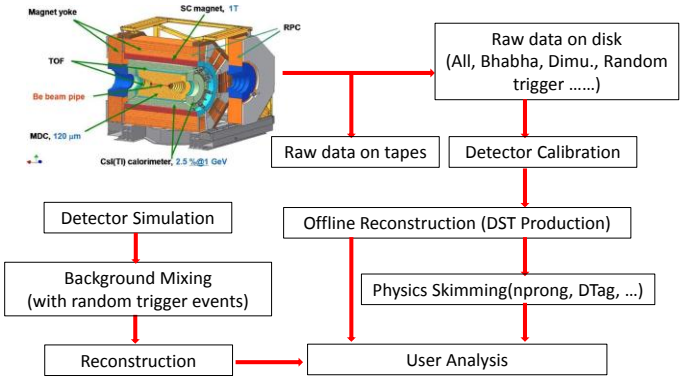

Figure 2. BESIII data flow

2. Local data processing

2.1. Local data management

Local data management system provides interface for management of raw data, calibration constants, and Monte Carlo tuning parameters. Offline data manager uses the interface to import information of raw data files from online database, search data files and create datasets. Calibration managers use this interface to save calibration constants for specific sub-detector, software version, and run range. Interface for users to search specific calibration constants is also provided (Figure 3.). Simulation, calibration, reconstruction and physics analysis jobs are running at local cluster. But as data sample rapidly increase, local cluster could not meet the peak requirements for computing. So there is a growing demand to move from solely using a local cluster to a more distributed computing model.

2.2. Local event-level paralleled data processing system

As data sample rapidly increase, the calibration jobs occupy more and more CPU time. A calibration job with one raw data file as input will last 10 hours, and some specific calibration jobs need iteration for several times. In order to save time for calibration and software validation, a local paralleled data processing system (DistBoss, Figure 4.) is developed to enable paralleled data processing at event level. Ganga [7] is used as user interface. Diane [8] is used to control and manage the running of master and workers. Raw events are distributed to multiple work nodes for event processing. The output DST events are collected from the work nodes and then combined. A job to process one raw data file can be split to several jobs running at different worker nodes. The time for data calibration is much less than before.

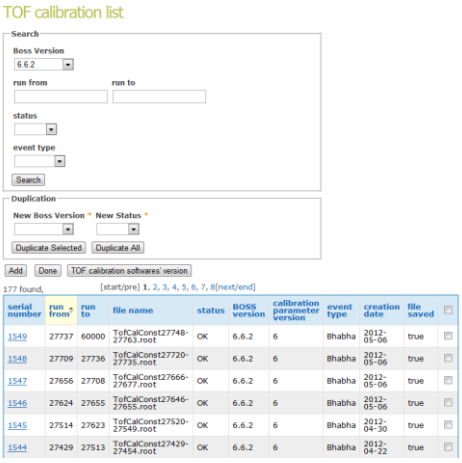

Figure 3. Local data management system

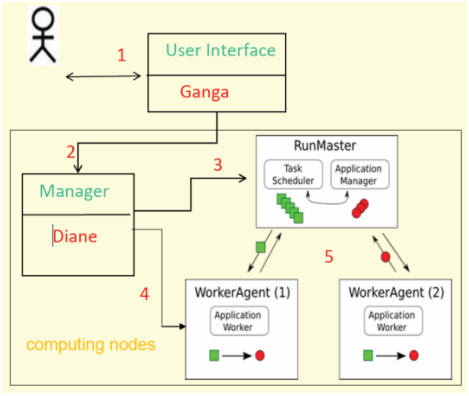

Figure 4. local distributed computing system

2.3. Event navigation

We have designed a new event navigation model (Figure 5.) to be realized in the near future. Event indexes for raw events will be saved into DST data during reconstruction, and the tag generation based on DST data will generate event tag information for each event. The event tag information should include physics tags which are useful for pre event selection, and also the run number, event number, physical location of raw data and DST data. The information is sufficient for event navigation to help to access event in upstream data by event index, and then enable quick reconstruction of selected events, quick analysis, and event display based on TAG data. In the current data processing flow (figure 2), after the reconstruction of raw data files, DST data files were saved permanently for physics study. Besides the full DST data, physics users also skim the DST data with different criteria. These skimmed DST data files also saved for quick analysis. So the events in the skimmed DST data files

exist in two data copies. In the new data flow, there is no need to keep the skimmed DST data, only keep the full DST data, so large disk space can be saved.

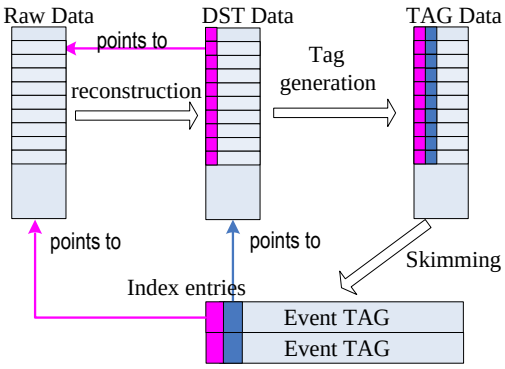

Figure 5. BESIII event navigation

2.4. BESIII data processing

By the end of 2011, BESIII has collected 225M J/ψ, 106M ψ', 2.9fb^{-1} ψ'', 0.5fb^{-1} ψ(4040). For the total data collected by the end of 2011, data size of total raw data is 250TB, random trigger data 9TB, total DST data for real data is 85TB for one software version. MC raw event size is 5-7kB per event. MC DST event size is 18-25kB per event. 225M J/ψ inclusive mc raw data and DST data is 5TB. Total MC data (raw+DST) is 16TB for one software version.

The BESIII detector is planning to collect more J/ψ, ψ', ψ'', and data at higher energies. The BESIII will take about 10 billion J/ψ data and the data collected in other energy points such as ψ', ψ'', ψ(4040), etc will be of equivalent size. The total amount of raw data is estimated to be about 3.6PB. It is supposed the data reconstruction is repeated at least twice a year. The total size of DST for raw data will be about 1.8PB. The total storage capacity for Monte Carlo events should be 1.0 PB.

The BESIII computing environment architecture is shown in Figure 6. Computing resources in IHEP(Institute of High Energy Physics, Beijing, China) are increased year by year. There are 3398 CPU cores, and 1536 cores to be available in the near future. Castor is used for mass storage system. 3PB out of 4PB are available. Lustre is used for local file system, 1000TB out of 2464TB are available.

From the experience of latest production jobs, with 2000 cores, it will take 8 days to produce 1 billion J/ψ inclusive mc DST events; take 7 days to reconstruct 1 billion J/ψ raw data; take 1 day to reconstruct 0.1 billion ψ' raw data; take 13 days to reconstruct 2.9fb^{-1} ψ'' raw data.

With more data accumulated year by year, it's more difficult for IHEP to provide all the computing resources for both raw data processing and MC production. We hope to use the computing resources from other institutes or universities in BESIII collaboration. Distributed computing is necessary.

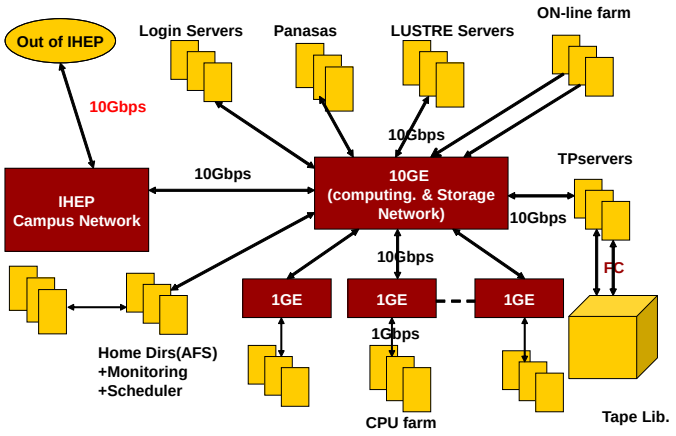

Figure 6. BESIII computing environment architecture

3. Distributed computing

3.1. Computing model

Considering that most sites in BESIII collaboration are in small scale, and lack experts on grid computing, and limitation of network speed between IHEP and other sites, the design principle of BESIII distributed computing is to make it as simple as possible for sites to join and for users to use, and use existing software wherever possible. The computing model is IHEP will be responsible for processing and storage of all the real raw data, MC production and analysis jobs are distributed among a number of sites with enough computing resources. The basic requirement for each site is 100 CPU cores and 100TB storage space.

3.2. Main components

For the main components of BESIII distributed computing (Figure 7.), BES VO and VOMS are ready. Ganga is used for job submission and management. GangaBoss plug-in has been developed to support BESIII software. DIRAC [9] for running distributed computing jobs. DIRAC server is running at IHEP with clients at remote sites. DIRAC File Catalogue is used for data management. CVMFS [10] is used for deploying BOSS software on target sites. Clients running at distributed sites can load BOSS version from server at IHEP.

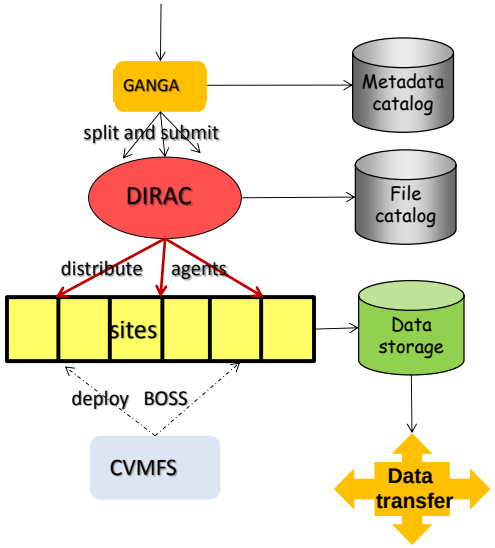

Figure 7. Components of BESIII distributed computing

3.3. MC production job splitting

During the MC production job splitting, random seed is got from database, and the updated random seed is reset to database after job splitting, to make sure each job has a unique random seed. Jobs sent to different sites will have corresponding database configuration. SQLITE [11] is also an option. After the simulation job finished, the succeeding reconstruction job will begin at the same work node. Random trigger data files are needed for background mixing in the reconstruction job. There is a copy of all the random trigger files in the SE at each site. Each reconstruction job then downloads the respective random trigger data files to do background mixing. For each job, 2~3 data files to be downloaded, the total size is 800MB~3GB.

3.4. Performance test

Currently a test prototype for BESIII distributed computing is set up. There are two kinds of sites involved. Two PBS sites (GUCAS: 80 cores, IHEP-PBS: 96 cores) and two LCG sites (JINR: 220 cores, IHEP-LCG: 8 cores). One DPM SE with 200 TB storage capacity is available in IHEP for storage of all the MC data, and also used for the buffer for transferring data between local Lustre system and remote SE. Another SE is dCashe in JINR site with 3.5TB served for BESIII test jobs.

BOSS 6.6.0 is successfully deployed to 4 sites with CVMFS. About 3000 BESIII production jobs have been split and submitted to DIRAC, 50M MC events produced. The output files registered in DIRAC File Catalogue automatically with the designed hierarchy, such as /bes/File/jpsi/660/mc/ rhopi/exp1/stream001. File level metadata registered after job finished. Validation between local and distributed computing finished, results are consistent. Small number of jobs sent to remote sites failed with software libraries not found, Job stalled, pilot not running, and some other reasons.

3.5. Next plan

What we need to do in the next step is to analysis the failed production jobs and find the reason, and try to reduce the failure rate, and test physics analysis job at distributed sites. Since all the raw data and the corresponding DST data are saved in local Lustre file system, we need to add a SRM server to it to enable file registered in the DIRAC File Catalog to be accessed by jobs. IHEP already has bbftp-based tool for data transfer between sites. But it can't meet the requirements for data transfer between SE. Selection and integration of other data transfer tool is an important task.

We need to integrate more sites to BESIII grid and continue cooperation with DIRAC developers to make DIRAC more suitable for BESIII experiment. Volunteer computing with BOINC[12] + CernVM is also being considered to sufficiently use the volunteer PCs in IHEP, which will be a supplement of BESIII distributed computing.

4. Summary

Large scale data samples from BESIII have been successfully processed. With the data sample increasing rapidly, local computing resources can't meet the requirements to process all the raw data and produce MC data. DIRAC based distributed computing system is set up and the performance test based on prototype system shows it can work. More work is needed before it comes into use.

References

[1] M. Ablikim et al. Design and Construction of the BESIII Detector Nuclear Instruments and Methods in Physics Research Section A, Volume 614, Issue 3, p. 345-399.
[2] http://www.cmtsite.org/
[3] G. Barrand et al. GAUDI – a software architecture and framework for building HEP data processing applications, Computer Physics Communications, 140 (2001), p. 45
[4] http://castor.web.cern.ch/
[5] http://wiki.lustre.org/index.php/Main_Page
[6] R. Brun, F. Rademakers ROOT- an object oriented data analysis framework, Nucl. Instrum. Methods A, 389(1997), p.81
[7] J.T.Moscicki et al. GANGA: A tool for computational-task management and easy access to Grid resources, Computer Physics Communications Volume 180, Issue 11, November 2009, Pages 2303-2316
[8] http://it-proj-diane.web.cern.ch/it-proj-diane/
[9] A Tsaregorodtsev et al. DIRAC: A community Grid solution, J. Phys.: Conf. Ser., 119 (2008), p. 062048
[10] http://cernvm.cern.ch/portal/filesystem
[11] http://www.sqlite.org/
[12] http://boinc.berkeley.edu/index.php