BES-III DISTRIBUTED COMPUTING

Z.Y. Deng[1], W.D. Li[1], L. Lin[2], C. Nicholson[3], X.M. Zhang[1], A. Zhemchugov[4]
1) Institute of High Energy Physics, 100049 Beijing, China
2) Soochow University, 215000 Suzhou, China
3) Graduate University of Chinese Academy of Sciences, 100049 Beijing, China
4) Joint Institute for Nuclear Research, 141980, Dubna, Russia

Introduction

The BES-III experiment [1] started to take data in 2009 after a major upgrade of the electron-positron collider BEPC-II at the Institute of High Energy Physics (Beijing, China). The experiment is run by an international collaboration of more than 400 active members from 52 institutes in 12 countries from around the world. The main physics goals of the experiment are precision measurements in the τ-charm domain in the energy range of 2 – 4.6 GeV. The BES-III experiment has already taken the world's largest data samples of J/ψ ($1.2x10^9$ events) and ψ' decays ($0.3x10^9$ events), as well as a large amount of ψ(3770) data and a unique sample of ψ(4040) data. The total volume of experimental data is about 0.5 PB, of which about 120 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 power of the experiment and speed up data analysis.

The BES-III computing model

Experimental data are taken from the BES-III detector and stored as raw 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 100 TB. Both inclusive and exclusive Monte-Carlo simulation (MC) is made for each data sample as well. Experimental data taken with 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 20 TB now. The BES-III offline software is based on the Gaudi framework and runs on Scientific Linux CERN operating system. Almost all data processing and user analysis are carried out at IHEP local computing farm so far.

The BES-III distributed computing system

After successful deployment in the LHC experiments, Grid computing became a routine tool for data processing in high energy physics. However, the main difficulty for 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 which was designed for LHC experiments for use in a medium scale experiment, and because of limited manpower, it is even more difficult to maintain. For BES-III the situation is even worse, because very few participating sites are members of WLCG; there are therefore few experienced Grid users and developers and little corresponding computing infrastructure already installed. 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 experimental and simulated data is done at IHEP as before. If this is the case, 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. For the moment there are no plans to develop a distributed analysis system.

Several components are necessary to implement these models in BES-III: authentication and authorization system, production job management system, data management system and information and monitoring system. Authentication and authorization is based on use of X.509 certificates and on membership in the virtual organization 'bes' which is managed by VOMS from the gLite software stack [2]. For the information and monitoring system, custom tools will be developed, reusing components of the CERN Dashboard [3]. The most challenging parts of the BES-III distributed computing system are the job management and data management systems.

Job management system

The DIRAC (Distributed Infrastructure with Remote Agent Control) project is the most advanced and complete Grid solution for medium-scale high energy physics experiments today [4]. This solution was designed originally for the LHCb experiment, but was later developed as a generic product which could be used to access distributed computing resources in various communities of users. The key point of DIRAC is its workload management system, based on generic pilot jobs.

DIRAC is adopted as a central part of the BES-III job management system. A prototype installation of DIRAC has already been set up for BES-III, with five remote sites and the DIRAC server running at the IHEP central site (Beijing). Users of DIRAC can also benefit from use of the GANGA tool to manage production and analysis jobs [5]. The main problem is that not all LRMS used at BES-III remote sites, like Condor, are supported by DIRAC yet, so new DIRAC plugins need to be developed.

CVMFS (CERN VM File System) [6] is deployed to centrally manage the experiment software BOSS and distribute it to the target sites.

Data management system

The data management system is a key issue when building the BES-III distributed computing system. Of course, DIRAC has certain data management functionality, but it is not sufficient for BES-III, taking into account that network is not stable between most of the BES-III remote sites. There are several issues to be solved before BES-III data management becomes operational.

The first one concerns reliability of data transfer. A number of services to provide reliable file transfer between Grid sites already exist. FTS from the EMI/gLite software stack is adopted as a data transfer service for BES-III. The BES-III FTS server is installed at JINR, providing reliable data transfer between IHEP and remote sites via both SRM and GridFTP protocols. Certain modifications are made, though, to avoid using BDII and to optimize performance of data transfer.

The second issue is related to the file and metadata catalog. There are two solutions – one is AMGA from the gLite software [7] and the other is DFC from DIRAC. Of course, the latter fits better because integration with other pieces of the BES-III Grid is easy. Several tests were made to assure that the performance of DFC meets BES-III requirements. The BES-III metadata schema was implemented in both AMGA and DFC catalogs, using current data (~$2x10^5$ files) and a MySQL backend. Configuration was optimized both for AMGA and DFC (8 DFC instances, max. 50 threads / instance; for AMGA maximum 140 processes allowed). The tests have shown that with a low number of clients AMGA queries are ~10x faster than DFC, but with a high number of clients query times become approximately equal (see Fig.1). At the same time, DFC CPU usage rises more slowly with number of concurrent clients (see Fig.2). As a result, both AMGA and DFC give acceptable performance, but DFC meets more of BES-III requirements.

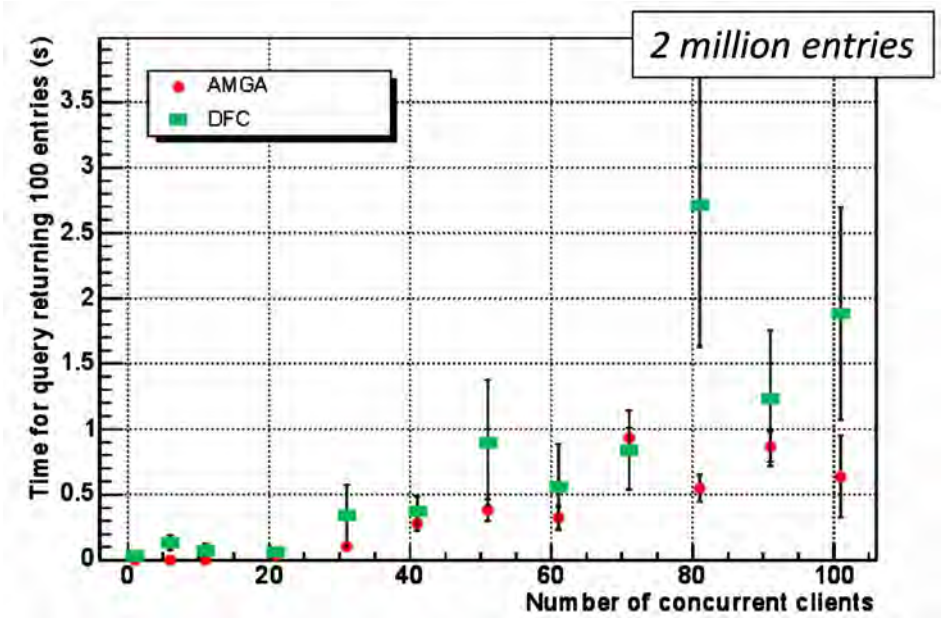

Fig. 1: Performance of AMGA and DFC versus the number of concurrent clients

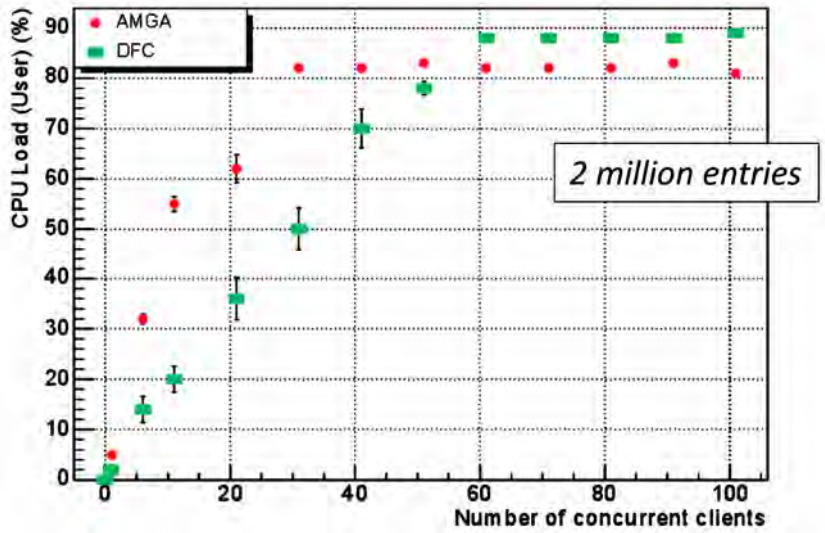

Fig. 2: CPU load created by AMGA and DFC versus the number of concurrent clients

The last issue concerns management of datasets. The BES-III experiment collects different types of data, so using datasets as containers of files and metadata is very convenient. DIRAC allows DFC queries or 'meta-sets' to be used, which can be considered as dynamically changing datasets. At the same time, most of the physics tasks at BES-III require reproducibility of results of these queries, because the total number of events is important in the data analysis. Datasets at BES-III should therefore be static in the sense that their content is always the same. For the moment it is assumed that dynamic datasets from DIRAC can be used provided extra instruments to assure the dataset constancy are implemented.

Of course, to glue all these pieces into a working system one has to develop other tools which take into account BES-III specific issues, provide a user interface and API for the job management system etc. These tools are under development as the BADGER (BES-III Advanced Data manaGER) project.

Summary

The BES-III experiment has been running since 2008 and is currently the best source of data in the τ-charm domain. The amount of data is increasingly high so using distributed computing is an attractive option to go beyond the limits of the computing power available at BES-III now. A BES-III Grid is being constructed, based on the DIRAC infrastructure combined with experiment-specific data management. A working prototype has already been set up which unites the computing resources of IHEP CAS, GUCAS, Peking University, USTC, the University of Minnesota and JINR, with several more sites planning to join. In conclusion, it is worth mentioning that Grid computing is becoming widely used not only in big projects like the LHC experiments, but also in many medium-scale experiments in high energy physics and even beyond. Experience gained in the BES-III Grid is valuable to better design a yet missing universal Grid solution for medium scale projects like these.

References

[1] M.Ablikim et al., Design and construction of the BESIII detector, Nucl. Instrum. Meth A614 (2010) 345.
[2] http://glite.cern.ch/
[3] http://dashboard.cern.ch/
[4] http://diracgrid.org/
[5] J.T.Moscicki et al., Ganga: a tool for computational-task management and easy access to Grid resources; Comp. Phys. Comm. Vol 180, Issue 11, (2009) ;arXiv:0902.2685; see also http://cern.ch/ganga
[6] http://cernvm.cern.ch/portal/filesystem
[7] http://amga.web.cern.ch/amga/