Journal of the Korean Physical Society, Vol. 59, No. 4, October 2011, pp. 2695∼2701

The Embedment of a Metadata System at Grid Farms at the Belle II Experiment

S. Ahn, J. H. Kim, T. Huh, S. Hwang, K. Cho,* H. Jang, B. K. Kim, H. Yoon and J. Yu
Korea Institute of Science and Technology Information, Daejeon 305-806, Korea

Z. Drasal
Charles University, Prague 116 36, Czech

T. Hara, Y. Iida, R. Itoh, G. Iwai, N. Katayama, Y. Kawai, S. Nishida, T. Sasaki and Y. Watase
High Energy Accelerator Research Organization (KEK), Tsukuba 305-0801, Japan

T. Uglov
Institute for Theoretical and Experimental Physics, Moscow 117218, Russia

R. Frühwirth and W. Mitaroff
Institute of High Energy Physics, Austrian Academy of Science, Vienna A-1050, Austria

R. Grzymkowski, M. Sitarz and M. Zdybal
H. Niewodniczanski Institute of Nuclear Physics, Krakow PL-31-342, Poland

M. Heck, T. Kuhr and M. Röhrken
Institut für Experimentelle Kernphysik, Universität Karlsruhe, Karlsruhe 76128, Germany

M. Bračko, R. Pestotnik, R. Petrič, L. Šantelj and M. Starič
J. Stefan Institute, Ljubljana 1000, Slovenia

S. Lee
Korea University, Seoul 136-701, Korea

C. Kiesling, S. Koblitz, A. Moll and K. Prothmann
Max-Planck-Institut für Physik, München 80805, Germany

H. Nakazawa
National Central University, Chung-li 32001, Taiwan

T. Fifield and M. E. Sevior
University of Melbourne, School of Physics, Victoria 3010, Australia

S. Stanič
University of Nova Gorica, Nova Gorica SI-5000, Slovenia

(Received 28 June 2011, in final form 18 August 2011)

In order to search for new physics beyond the standard model, the next generation of B-factory experiment, Belle II will collect a huge data sample that is a challenge for computing systems. The Belle II experiment, which should commence data collection in 2015, expects data rates 50 times greater than that of Belle. In order to handle this amount of data, we need a new data handling

system based on a new computing model, which is a distributed computing model including grid farms as opposed to the central computing model using clusters at the Belle experiment. We have constructed a metadata system and embedded the system in the grid farms of the Belle II experiment. We have tested the system using grid farms. Results show good performance in handling such a huge amount of data.

PACS numbers: 29.85.+c, 07.05.-t, 07.05.Bx
Keywords: e-Science, High-energy physics data grid, Grid computing, Data processing
DOI: 10.3938/jkps.59.2695

I. INTRODUCTION

The B factory experiments, Belle at the KEKB (KEK B-Factory collider) at KEK [1] and BaBar at the PEP (positron electron project) II collider at SLAC (Stanford linear accelerator center) [2], were designed to measure the large mixing-induced CP (charge-parity) violation in the B^0 system predicted by the theory of Kobayashi and Maskawa [3]. Although the experimental results are in agreement with the standard model, the goal of Belle II is to search for new physics beyond the standard model [4].

This search is performed via B and charm decay measurements with unprecedented precision. To be sensitive to new physics effects, much larger data samples than those recorded by the B factories are required. The search for new physics via precision flavor physics measurements is complementary to the direct search for new particles at the LHC (large hadron collider) [5].

The aim is to collect a data sample of 50 ab^{-1} with the upgraded KEKB accelerator. This huge data volume is a challenge for the computing system. If computing processing is to be performed at the required scale, data grid technology is a strong requirement [6]. The amazing advance in IT (information technology), such as Moore's law, and the widespread use of IT help computing processing [7]. The objective of a HEP (high-energy physics) data grid is to construct a system to manage and process HEP data and to support the high-energy physics community [8]. To provide the required computing resources, we adopted a distributed computing system based on existing grid technologies [9].

To handle data on the required scale, we have developed a metadata service for the Belle II experiment, which provides a solution for the performance, scalability and durability compared to that of Belle experiment [10]. We have embedded a metadata system in the grid farms. This paper describes the embedment of a metadata system in the grid farms at the Belle II experiment.

*E-mail: cho@kisti.re.kr; Fax: +82-42-869-0789

II. COMPUTING MODEL AT THE BELLE II EXPERIMENT

1. Introduction

The Belle II experiment will begin in 2015. Belle II computing needs to include raw data reconstruction, data reduction, event simulation, and user analysis [5]. The Belle II experiment will have a data sample of about 50 times greater than the size collected by the Belle experiment. The collider will cause the computing requirement for data analysis and MC (Monte-Carlo) production to grow larger than the available CPU resources [5]. Therefore, we need to use a new concept of e-Science in this area [4]. In order to meet future demand, the Belle II experiment has examined the possibility of using shared computing resources. Like the LHC experiments, the Belle II experiment has adopted the distributed computing model with several computing processing systems [5]. Moreover, we will profit from the experience of the LHC experiments.

2. Computing Model

The main infrastructure of data processing is the processing of raw data. The main challenge is a data rate of up to 1.87 GB/s [5]. According to our baseline computing model, all the raw data are deemed to be stored and processed in KEK rather than copied and processed at remote grid sites [5].

The second main computing task is the production of generic MC samples. Like the raw data processing, the generic MC production is done in many ways. Because no input data are needed, it is very well suited for a distributed environment, in particular, for cloud computing [5].

Finally, the computing model has to support data analyses by physicists. The input to the analyses is the output of the raw data processing and MC production in the mini-DST format (mDST). If a distributed data analysis is to be enabled, the real and the MC mDST data samples are distributed to grid sites and replicated, if needed [5]. Because the mDST files are accessed in an uncoordinated, random way, they should be stored on disk. Figure 1 shows an overview of the Belle II computing model. With only three logical layers – the main

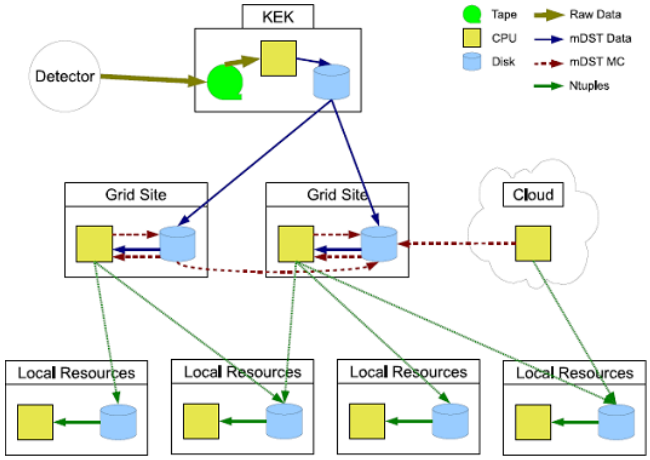

Fig. 1. (Color online) Overview of Belle II computing.

center at KEK, grid sites, and local resources – and centralized raw data processing, the model is simpler than the WLCG (world LHC computing grid) model.

III. METADATA SYSTEM FOR THE BELLE II EXPERIMENT

1. Introduction

Though computing technology improves according to Moore's Law, the processing time for the Belle II experiment will still be significantly greater than that of the Belle experiment. In addition, we need about ∼253 PB disk space for real and MC data in 2019. This means that we have very large disk space requirements and potentially unworkably long analysis times. Therefore, we suggest a metasystem at the event-level to meet both requirements. If we have good information at the metasystem level, we can reduce the CPU time required for analysis and save disk space [10].

2. Metadata System at the Belle II Experiment

In the current Belle experiment, we use a metadata scheme that employs a simple "index" file. This is a mechanism to locate events within a file based on predetermined analysis criteria. The index file is simply the location of interesting events within a larger data file. All these data files are stored on a large central server located at the KEK laboratory. However, for the Belle II experiment, this will not be sufficient as we will distribute the data to grid sites located around the world. Therefore, we use the AMGA (Arda Metadata Catalog for Grid Application) [11] as the metadata service for the Belle II experiment.

Figure 2 shows a scheme for the Belle II data handling system for metadata. First, a user sends a metadata

Table 1. Comparison of the Belle and the Belle II data cache systems.

Contents	Belle	Belle II
Type	PostgreSQL	Text (compression)
Size	8 Byte/event	2.14 Byte/event
Grid	Not available	Available
Metadata Catalog	Event driven	Condition driven

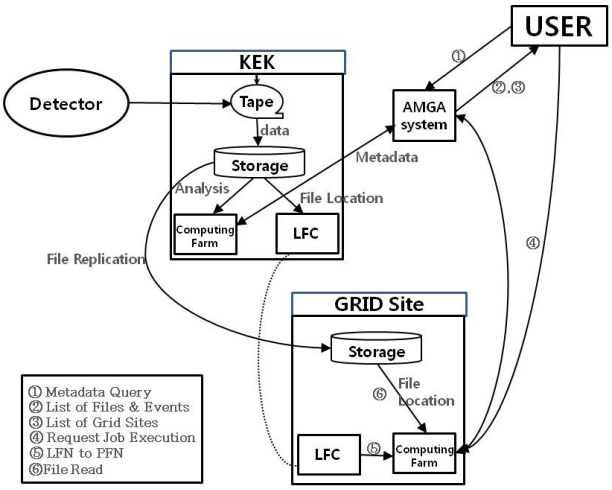

Fig. 2. Scheme of the Belle II data handling system.

query to AMGA server. Second, the AMGA server gives back a list of files and events. Third, AMGA sever also may give a list of grid sites. Fourth, the user requests job execution at grid sites. Fifth, the LFC (logical file catalog) maps a LFN (logical file name) into a set of PFN (physical file names). Finally, the computing farms at the grid site read the requested physical file.

To help users to efficiently select the data files for the analysis, we store metadata about files in the AMGA database. The estimated size of the file-level metadata for Belle II is about 60 GB [5]. While this is perfectly manageable, the estimated size of the event-level metadata is about 20 TB, which makes the realization of this kind of metadata service questionable [5]. To deal with this issue, we designed an advanced metadata service system based on AMGA, which provides efficient and scalable metadata searching [12]. We have built test-bed sites to test the correctness, performance and scalability of the advanced metadata service system, and it has been proven to be able to provide efficient metadata searching for the Belle II experiment [12].

3. Performance of the Metadata System

As shown in Table 1, we have developed a simple data tool that is not based on a data base. The new metadata system significantly reduces the CPU use in the bench

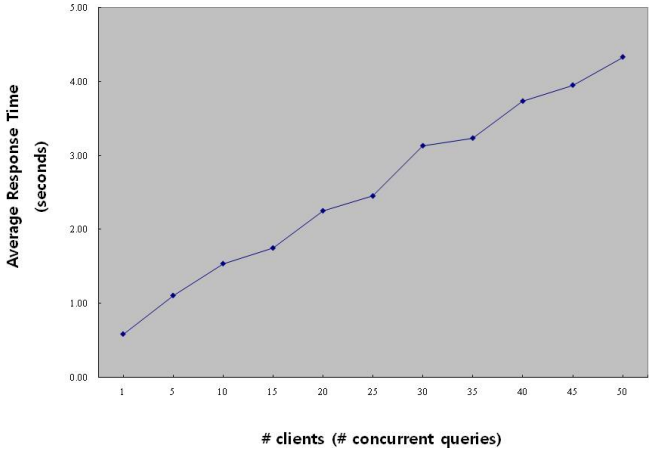

Fig. 3. (Color online) Average response time for the worst-case query of file-level metadata searching.

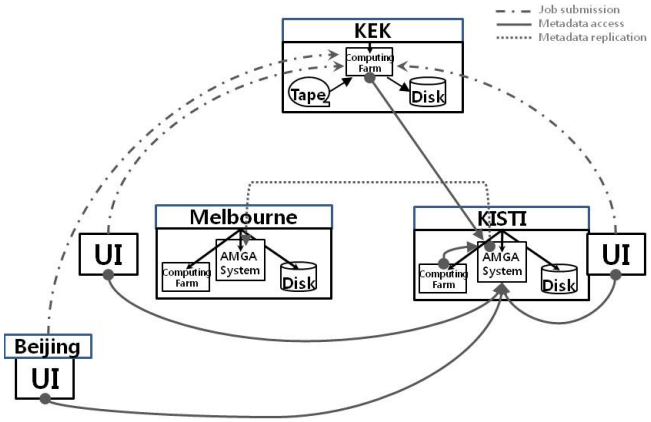

Fig. 4. Scheme of the metadata system between the main center and the KEK grid sites of Melbourne and KISTI.

mark analysis [10]. Since Belle II data are not yet available, we have instead made metadata from Belle data and random generation data for the Belle II experiment. Using 8,000 files, we have tested whether AMGA ensures a proper response time for file-level metadata searching. We have tried to measure the worst-case response time taken to retrieve all file-level metadata. Figure 3 shows the measured average response time taken per query for increasing number of concurrent queries [10]. Our prototype of the metadata service system for the Belle II experiment shows good performance. The replication of redirection of AMGA allows the Belle II metadata service to provide good scalability [10].

IV. EMBEDMENT OF A METADATA SYSTEM AT GRID FARMS

1. Grid Farms for the Belle II Experiment

The data size will be a few PB data per year. In order to handle this amount of data, we use distributed computing of both grid technology and cloud computing. For this work, we use grid farms at the main center at KEK and grid sites for the Belle II experiment. The LCG is the worldwide infrastructure where all the computations relevant to the analysis of the data coming out of the four LHC experiments are taking place [13]. As shown in the session II, Belle II computing hierarchy is simpler than that of the LCG organization, which involves a hierarchy of computing centers from CERN, labeled Tier-1, Tier-2, and Tier-3 [14].

We have built Belle II grid farms at the main center at KEK and the grid sites of Melbourne and KISTI (Korea Institute of Science and Technology Information). User's activity patterns are CPU intensive and involve large data file transport. These activity patterns have required the Belle II computing model to be changed from clusters to a grid to meet the required hardware resources. Dedicated Linux clusters on a PBS (portable batch system) were used when Belle launched in 1999. However, the Belle II cluster has been moved from a PBS to a grid-based implementation. Now, we have adapted and converted to a workflow to the grid. The goal of movement to a grid at Belle II experiment is a worldwide trend for HEP experiments. We need to take advantage of global innovations and resources because Belle II will have many data to be analyzed.

Belle II will use several systems of gLite and Cloud computing. The DIRAC (distributed infrastructure with remote agent control) system is allowed to change the underlying batch systems without changing the user interface.

2. Embedment of a Metadata System at Grid Farms for the Belle II Experiment

To test the metadata system, we have constructed grid farms, as shown in Fig. 4. Among three logical layers (the main center at KEK, grid sites, and local resources), we use two layers of the main center at KEK and grid sites (KISTI, Melbourne, *etc.*). We have installed an AMGA server for the data handling systems at grid sites of KISTI and Melbourne. The master node of the AMGA system is located at the KISTI site, and a slave node is located at the Melbourne site.

V. RESULTS

1. Scheme of Metadata Test at Grid Farms

We use the user interface machines at Beijing, Melbourne and KISTI, as shown in Fig. 4, and grid farms, as listed in Table 2. Figure 5 shows that jobs at the grid farms work well. Using AMGA servers at KISTI and

Table 2. Grid farms with Belle VO.

Site	Grid farms			AMGA server
	Name	Physical CPU	Logical CPU	
KEK site	JP-KEK-CRC-02	57	456	N/A
KISTI site	KR-KISTI-GCRT-01	146	584	Master node
Melbourne site	AU-PPS	1	4	Slave node
Poland site	CYFRONET-LCG2	1580	8712	N/A
Karlsruhe site	FZK-LCG2	2562	13757	N/A
Slovenia site	SiGNET	1445	2272	N/A
Czech site	Prague_cesnet_lcg2	20	80	N/A

Table 3. Authentication and authorization for the user setting.

Users	Description
belle2admin	belle2 Administrator
belle2	a belle2 user users not registered explicitly in AMGA, but having belle2 VOMS attributes in the certificate
belle	a belle user users nt registered explicitly in AMGA, but having belle VOMS attributes in the certificate

Table 4. Authentication and authorization for the group setting.

Groups	Members (account)
belle2admingroup	belle2admin (*Not applied yet)
belle2group	belle2,... (*Not applied yet)

Melbourne, we tested the metadata system at the grid farms.

2. Authentication and Authorization in the Belle II Metadata Service

For metadata service, we have set authentication and authorization for the user setting, the group setting and the ACL (Access Control List), as shown in Tables 3, 4, and 5.

3. How to Access the Metadata Service with a Certificate

Before submitting jobs, we need to submit a certificate and register it with Belle VO and Belle II VO by using the following steps:

Table 5. Authentication and authorization for the ACL setting.

Directory	Permission	Allowance
/belle2/*	write	belle2admingroup (*Not applied yet)
/belle2/*	read	belle2group (*Not applied yet)
/belle2/user/*	write	owner of a directory

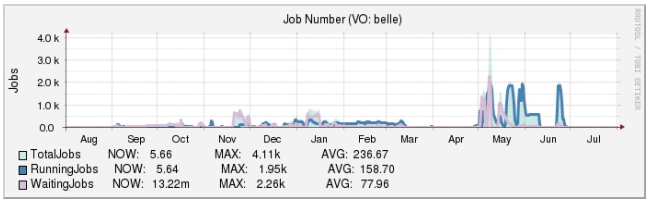

Fig. 5. (Color online) Monitoring system showing that jobs at the grid farms work well.

(1) To initialize user's proxy

```
$ voms-proxy-init – voms belle2
```

(2) To edit 'mdclient.config' file at your current directory.

If there is no 'mdclient.config' file, we need to copy one from '/opt/glite/etc/mdclient.config'.

(2-1) If we are a registered user at AMGA, then we set the 'Login' name to the registered account.

```
$ cat mdclient.comfig
Host = belle.amga.server.address
Login = belle2_admin
...
UseSSL = 1
AuthenticateWithCertificate = 1
UseGridProxy = 1
VerifyServerCert = 0
CertFile=/home/asunil/.globus/uwercert.pem
KeyFile=/home/asunil/.globus/userkey.pem
TrustedCertDir = /etc/grid-securit/certificates
...
```

(2-2) If we are not a registered user at AMGA, then we set the Login name to NULL. It will match us to a

default user (belle or belle2) if we are involved in the Belle VO or Belle II VO.

```
$ cat mdclient.comfig
Host = belle.amga.server.address
Login = NULL
...
```

(3) Now access AMGA with some tools: belle_amga_access, mdcli, mdclient and *etc.*

```
$ mdclient
Query > whoami
belle2
Query > exit
$
$ belle2_amga_access – mc –stream 0 – start_run 1000
–end_run 1200 –type uds 7

process_event /bdata/mcprod/dat/e000007/evtgen/
uds/00/all/0807/on_resonance/10/evtgen-uds-00-all-
e000007r001000-b20030807_1600.mdst 0

process_event /bdata/mcprod/dat/e000007/evtgen/
uds/00/all/0807/on_resonance/10/evtgen-uds-00-all-
e000007r001002-b20030807_1600.mdst 0

process_event /bdata/mcprod/dat/e000007/evtgen/
uds/00/all/0807/on_resonance/10/evtgen-uds-00-all-
e000007r001002-b20030807_1600.mdst 0
...
```

4. Check Results

In order to make sure that metadata accesses have been performed successfully, we check the log files at the AMGA server. The log files show that the performance has been performed successfully. An example of an output log file is as follows:

```
Fri Jan 7 09:26:48 2011 Authentication succeeded for 'C
= AU, O = APACGrid, OU = The University of Mel-
bourne, CN = *** *******': user belle in VO 'belle'
with groups '/belle' Fri Jan 7 13:35:12 2011 Authenti-
cation succeeded for 'C = AU, O = APACGrid, OU =
The University of Melbourne, CN = *** *******': user
belle in VO 'belle' with groups '/belle'
...
```

5. Summary

We have embedded the metadata system in grid farms and supported users. We have submitted our jobs from UI (user interface) machines at KISTI, Melbourne and Beijing. The results show our jobs have been finished successfully without any errors. Figure 6 shows jobs submitted from UI at Melbourne, and Fig. 7 shows jobs from Beijing. Both results show that the jobs at the metadata system in the grid farms provided output successfully.

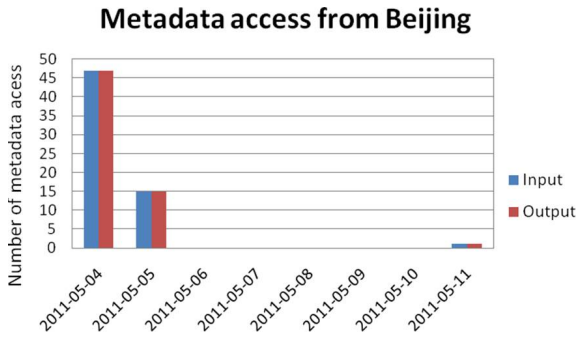

Fig. 6. (Color online) Metadata access from the Melbourne site.

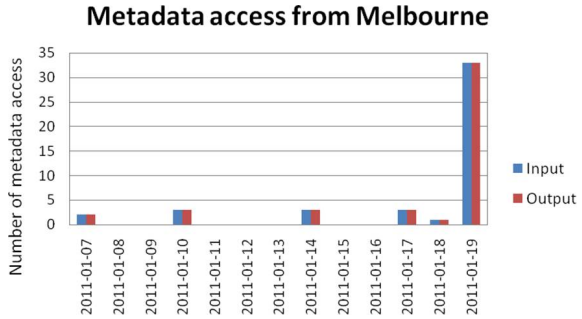

Fig. 7. (Color online) Metadata access from the Beijing site.

VI. CONCLUSION

In order to search for new physics beyond the standard model, the next generation of B-factory experiment Belle II will collect a huge data sample, which is a challenge for computing systems. In order to provide the required computing resources, we adopted a distributed computing model based on existing grid technologies. Using the main center at KEK and grid sites (KISTI, Melbourne and *etc.*), we have embedded a metadata system in grid farms and tested the system. The results show good performance in handling the metadata system in grid farms.

ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0027430).

REFERENCES

[1] A. Abashian *et al.*, Nucl. Instrum. Methods Phys. Res., Sect. A **479**, 117 (2002).
[2] B. Aubert *et al.*, (BaBar Collaboration) Nucl. Instrum. Methods Phys. Res., Sect. A **479**, 1 (2002).

[3] M. Kobayashi and T. Maskawa, Prog. Theor. Phys. **49**, 652 (1973).
[4] K. Cho and H. W. Kim, J. Korean Phys. Soc. **55**, 2045 (2009).
[5] T. Kuhr, in *Proceeding of Conference on Computing in High Energy Physics 2010* (Taipei, Taiwan, 2010).
[6] I. Foster, C. Kesselman and S. Tuecke, Int. J. High Perform. Comput. Appl. **15**, 200 (2001).
[7] K. Cho, Comput. Phys. Commun. **177**, 247 (2007).
[8] M. Jeung, H. W. Kim, K. Cho and O-H. Byeon, J. Korean Phys. Soc. **55**, 2067 (2009).
[9] K. Cho, J. Korean Phys. Soc. **53**, 1187 (2008).
[10] J. H. Kim *et al.*, Comput. Phys. Commun. **182**, 270 (2011).
[11] See http://cern.ch/amga.
[12] S. Ahn *et al.*, J. Korean Phys. Soc. **57**, 175 (2010).
[13] K. Cho, H. W. Kim and M. Jeung, J. Phys. Conf. Ser. **219**, 072032 (2010).
[14] K. Cho, Int. J. Comp. Sci. Network Sec. **7**, 49 (2007).