PROCEEDINGS OF SCIENCE

Overview of the Belle II computing

Yuji Kato[*a] **on behalf of the Belle II computing group**[b]

[a] *Kobayashi-Maskawa Institute for the Origin of Particles and the Universe, Nagoya University, Chikusa-ku Furo-cho, Nagoya, Japan*

[b] *https://confluence.desy.de/download/attachments/35005658/AuthorList4Belle2Computing2016v4.tex*

E-mail: kato@hepl.phys.nagoya-u.ac.jp

In the Belle II experiment, a data sample of 50 ab^{-1} will be collected. The Belle II computing system is expected to manage the processing of massive raw data, production of copious simulation as well as many concurrent user analysis jobs. We adopted the distributed computing model to realize it. In this contribution, the overview and highlights of the Belle II computing system and activity of KMI are presented.

The 3rd International Symposium on " Quest for the Origin of Particles and the Universe"
5-7 January 2017
Nagoya University, Japan

*Speaker.

PoS(KMI2017)024

1. Introduction

Belle II is a next-generation B-factory experiment at the SuperKEKB accelerator in Japan, which aims to find new physics beyond the Standard Model from precise measurements of the decay of B-meson, τ lepton, charm meson and so on. The final target for the peak luminosity of the SuperKEKB accelerator is 8×10^{35} /cm^2/s, which corresponds to about 40 times higher than that of the KEKB, the predecessor of the SuperKEKB. The detail of the Belle II experiment can be found in [1]. The physics run without vertex detector (Phase II) will start at the beginning of 2018 and the one with full Belle II detector (Phase III) will start in late 2018. A data sample of 50 ab^{-1} will be accumulated by the end of the data taking. The computing resource required to process and store such a large data sample are 1MHS06 CPU and more than 100 PB storage. The Belle II experiment has adopted a distributed computing model to provide such a huge computing resource. In this paper, overview of the Belle II computing and contribution of KMI are described.

2. Overview of the Belle II distributed computing system.

The computing resources are provided by collaborative institutes around the world. The role of a computing site depends on the scale of the site as follows:

- *'Raw Data Center'*, which stores raw data and process it to produce 'mDST', containing all necessary information for physics analysis. KEK, where the experiment is performed, plays this role. In addition, big computing sites in each region i.e., North America, Europe and Asia will contribute to store a copy of the raw data. In the first 3 years of the experiment, PNNL (Pacific Northwest National Laboratory) will keep the whole copy. From the 4th year, the copy will be distributed to other big computing sites.
- *'Regional Data Center'*, where mDST files for data are distributed.
- *'MC production site'*, which is relatively small site, performs the Monte-Carlo (MC) data production and user skim to produce Ntuple.

The schematic view of the Belle II computing model is shown in Fig. 1. We have been developing the computing system based on existing technologies. We choose DIRAC[2] as a workload and data management system. It provides a common interface to a number of heterogeneous resources such as grids with different middlewares, cloud and local computing clusters. A file catalog is provided by LCG File Catalog[3]. A meta data catalog is provided by ARDA Metadata Catalog (AMGA)[4]. The Belle II software is distributed by CernVM-FS (CVMFS)[5]. Transfer of data is performed by FTS3 (File Transfer Service, version 3.0)[6]. In addition to these existing technologies, we have been developing the extension of DIRAC, BelleDIRAC, to meet the requirement of Belle II experiment. In particular, a large fraction of the effort has been spent for the development of 'Production System', which is responsible to produce the Monte-Carlo events, to process raw data and distribute the output automatically based on the definition of the production. The Production System is consist of several parts:

- 'Fabrication System' is responsible for defining jobs based on production definition and submitting them to DIRAC. It is also responsible for re-defining failed jobs and verifying output

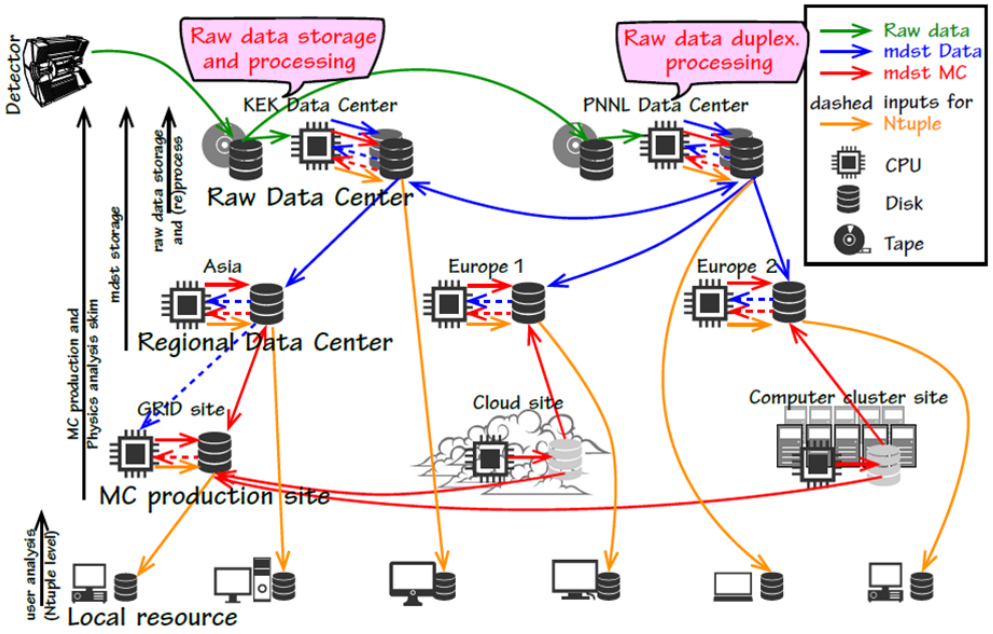

Figure 1: Schematic view of the Belle II computing model.

files. At this moment, the output files are stored in the temporally storage close to the production site.

- 'Distributed Data Management System (DDM)' is responsible for gathering the output files from temporally storage into the major storage (destination storage). It chooses a destination and defines the transfers after checking the availability of the destination storage and submits them on the DIRAC.

- 'Monitoring System' is responsible for monitoring the whole production activities. The detection of the issues in computing sites and monitoring of the progress of the production.

3. Monte-Carlo productions campaigns

In order to provide the simulation sample for physics sensitivity test and at the same time to understand the bottleneck of the computing system, the Belle II computing group has performed Monte-Carlo (MC) production campaigns seven times. The history of the MC production campaigns are shown in Fig. 2. At the beginning, only a limited number of jobs can be handled due to low performance of AMGA or failure of job when downloading (uploading) input (output) data. We have gradually improved the sytem to obtain the higher throughput[7, 8] and number of jobs have been increased. In addition, we have been developing the production system in order to automate the MC production. Until the third MC campaign, jobs were submitted manually by computing group members. The Fabrication System was implemented to automate job submission and has been in use from the fourth campaign. We have implemented the monitoring system that detects issues at the sites and lists them on a web-page in order to automate manual investigation

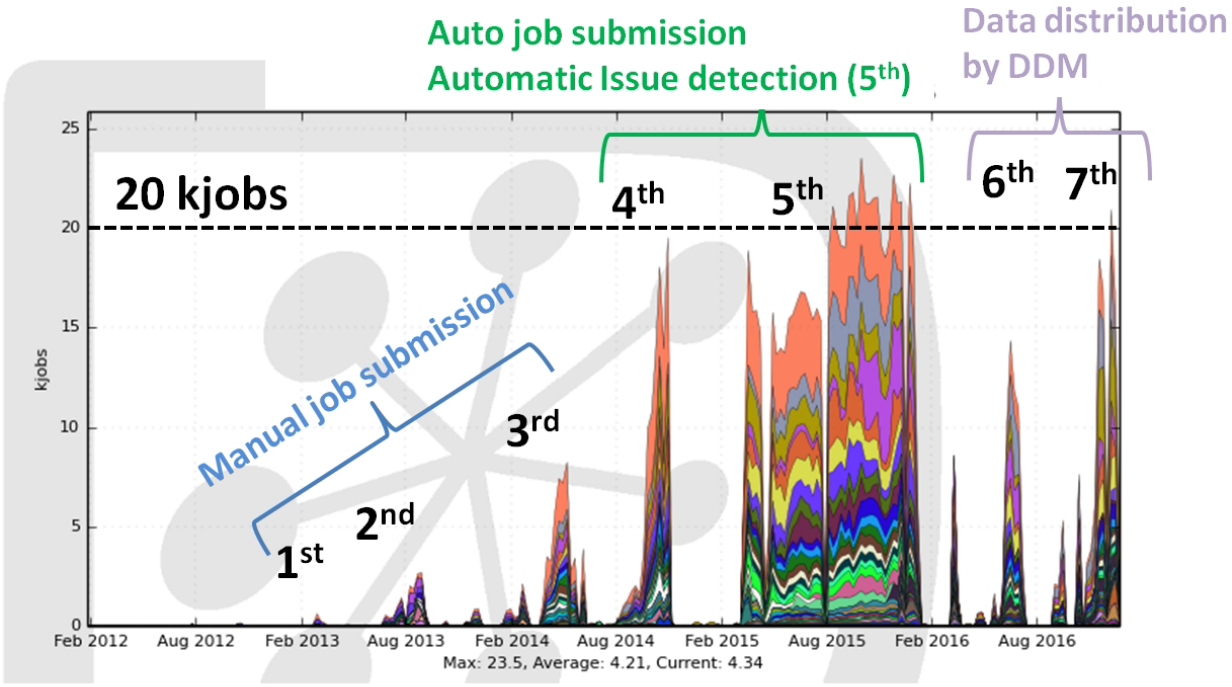

Figure 2: The history of number of jobs running. Different colors indicate different computing sites.

work. It has been used from the fifth campaign and has reduced the human cost largely. From the sixth campaign, the distribution of the output data to the destination storage has been handled by DDM. Currently, we can handle more than 20k concurrent jobs and transfer more than 3000 files per hour.

4. KMI contributions

The Tau-Lepton Data Analysis Laboratory in the KMI has the computing resources dedicated to Belle II computing from 2013. For the CPU resource, at the beginning, we had 160 core CPU resources, corresponding to around 2 kHEPSpec06. System has been upgraded gradually. We have added another 200 core CPU, corresponding to 2.5 kHEPSpec06 in October 2014, and added another 96 core CPU, corresponding to 2 kHEPSpec06 in January 2017. These CPUs are working quite stably and accounting around 4% of the whole Belle II computing resource in the latest MC production campaign. The history of the KMI CPU power is shown in Fig. 3 (left). The storage in KMI has been serving as a destination storage and more than 70 TB of the output files are collected (fourth most-used) in the last MC campaign as shown in Fig. 3 (right).

We are also responsible for the development of the Monitoring System. As written in section 3, we have developed monitoring tools to detect site issues and show them automatically on a web. More detail of the monitoring tools can be found elsewhere[9, 10]. Figure 4 shows the tool to check the progress of the production, which enables us to confirm how many events are planned to be produced and how many events are produced already.

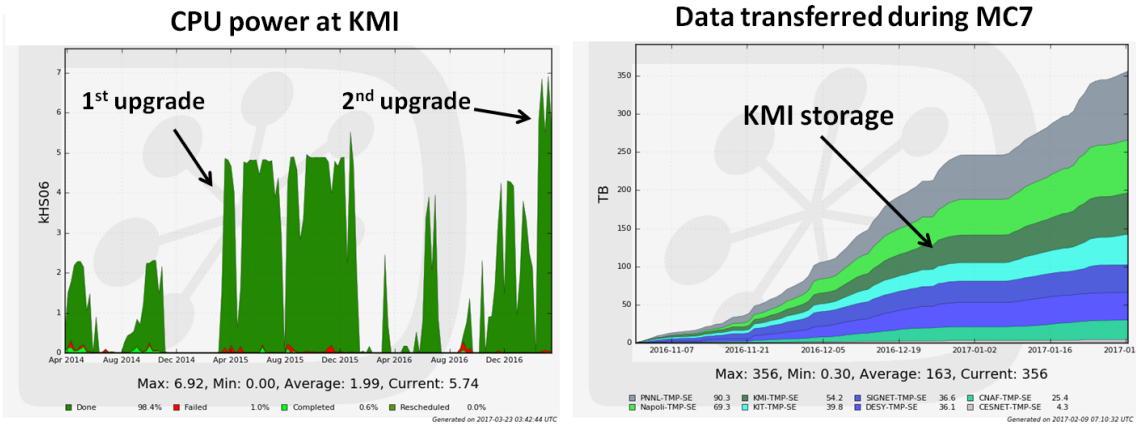

Figure 3: Left: History of CPU power of KMI resource in HEPSpec06. Green (red) histograms correspond to jobs finished successfully (in failure). Right: Data transferred during the 7th MC production campaign. Each color indicates a destination storage. Deep green represents KMI storage.

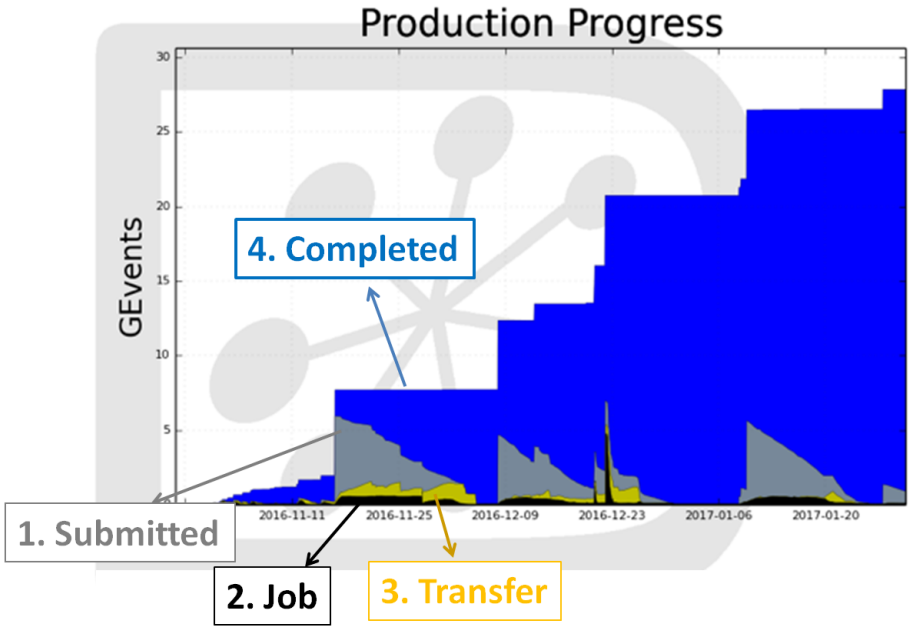

Figure 4: The monitor for the progress of MC production. Gray, black, yellow, and blue histograms show the events submitted, under job processing, under transfer, and finished, respectively.

5. Conclusion

We have been developing the Belle II distributed computing system towards the start of the physics run in 2018. We adopt DIRAC as the workload and data management system and utilize many other existing technologies, and developed the extension, BelleDIRAC. Our Production System can handle more than 20 k concurrent jobs and more than 3000 transfer of the output files per hour automatically. We have been working on the improvement of the Production System in order to increase the throughput and implement more features to automated for the monitoring and daily operation. KMI group has contributed on providing computing resources and developing the monitoring tools. In 2017, we are planning to perform cosmic-ray data processing which is the first use case to try the raw data processing with real data, and system dress rehearsals to confirm the full chain work-flow from raw data to the event skimming process.

6. Acknowledgments

We are grateful for the support and the provision of computing resources by CoEPP in Australia, HEPHY in Austria, Compute Canada (McGill and Victoria), CESNET in the Czech Republic, IPHC computing center, CNRS, in France, DESY, GridKa, LRZ/RZG in Germany, DST and NKN in India, INFN-CNAF, INFN-LFN, INFN-LNL, INFN Pisa, INFN Torino, ReCaS (Univ. & INFN) Napoli in Italy, KEK-CRC, KMI in Japan, KISTI GSDC in Korea, FCFM-UAS in Mexico, Cyfronet, CC1 in Poland, NUSC, SSCC in Russia, SiGNET in Slovenia, ULAKBIM in Turkey, and OSG, NERSC (under DOE Contract No. DE-AC02-05CH11231), PNNL in USA We acknowledge the network service provided by CANARIE, Dante, ESnet, GARR, GEANT, and NII. We thank the DIRAC and AMGA teams for their assistance and CERN for the operation of a CVMFS server for Belle II. This work is supported by a Grant-in-Aid for Scientific Research (S) "Probing New Physics with Tau-Lepton" (No.26220706).

References

[1] T. Abe *et al.* [Belle-II Collaboration], arXiv:1011.0352 [physics.ins-det].

[2] http://diracgrid.org/.

[3] https://twiki.cern.ch/twiki/bin/view/LCG/LfcGeneralDescription.

[4] S. Ahn *et al.* Journal of the korean Physical Society 57 issue 4 715, 2010.

[5] http://cernvm.cern.ch/.

[6] "FTS3: New Data Movement Service For WLCG", A. A. Ayllon *et al.*, 2014 J. Phys.: Conf. Ser. 513 032081.

[7] "Improvement of AMGA Python Client Library for the Belle II Experiment", J. H. Kwak, *et al.*, J. Phys.: Conf. Ser. 664 042041.

[8] "Directory Search Performance Optimization of AMGA for the Belle II Experiment", G. Park, *et al.*, 2015 J. Phys.: Conf. Ser. 664 042030.

[9] "Monitoring system for the Belle II distributed computing", K. Hayasaka *et al.*, 2015 J. Phys.: Conf. Ser. 664 062020.

[10] "Job monitoring on DIRAC for Belle II distributed computing", Y. Kato *et al.*, 2015 J. Phys.: Conf. Ser. 664 062023.