Computing at the Belle II experiment

2015 J. Phys.: Conf. Ser. 664 012002

(http://iopscience.iop.org/1742-6596/664/1/012002)

Computing at the Belle II experiment

Takanori HARA[1] on behalf of the Belle II computing group[2]

[1]High Energy Accelerator Research Organization, 1-1, Oho, Tsukuba, Japan
[2]https://belle2.cc.kek.jp/~twiki/pub/Public/ComputingPublic/AuthorList4Belle2Computing.tex

E-mail: takanori.hara@kek.jp

Abstract. Toward the start of data taking with the beam provided by the SuperKEKB accelerator in 2017, the Belle II, the next-generation flavor factory experiment in Japan, is establishing the computing system based on a distributed computing technologies. The system was examined in the periodical simulation mass production campaigns and was improved according to constructive feedback. In parallel, the data transfer challenges were performed with the transpacific and transatlantic network which plays an essential role in success of the Belle II experiment in the next decade. In this paper, the Belle II computing model and recent activities are presented.

1. Introduction

The results from B-factories in 2000s, Belle[1] and BaBar[2], confirmed the existence of large *CP* asymmetry in the *b*-quark system[3,4] as predicted in the Kobayashi-Maskawa theory[5]. However, the matter-antimatter unbalance in the universe we live in cannot be explained by the theory alone. It implies that as-yet undiscovered new physics is there to be found.

The Belle II experiment is the next-generation flavor factory experiment at the SuperKEKB asymmetric energy e^+e^- collider in Tsukuba, Japan. The first physics run will take place in 2017, then we plan to increase the luminosity gradually. We will reach the world's highest luminosity L=8x10^{35} cm^{-2}s^{-1} after roughly five years operation and collect a total of 50ab^{-1} data by 2024, which corresponds to 50 times more data than the 11-year operation of the Belle experiment. Thanks to such a huge amount of data, we can explore the new physics possibilities through a large variety of analyses in quark sectors as well as tau physics and deepen understanding of nature. In this exploration, the computing system is essential as well as the Belle II detector and the SuperKEKB accelerator.

Now, the SuperKEKB accelerator construction is finishing up and starts the commissioning run ("phase 1") in early 2016. The integration of the sub-detectors to the Belle II detector is ongoing. The stringing wires of the CDC (Central Drift Chamber) was finished and the Barrel and Endcap part of the KLM (K_L and muon detector) was completed last year. All the sub-detectors except for the vertex detectors will be installed in 2016 and the so-called "phase 2" beam run will take place in the middle of 2017. After the installation of the vertex detectors in early 2018, we will start the physics run ("phase 3") with the full Belle II detector from October 2018. In terms of the Belle II distributed computing, we need to start the stable operation before the beam data taking.

In this paper we will introduce the current model of the Belle II distributed computing system. Then we will present the highlights of the recent achievements such as mass production of simulation events and the data transfer challenges.

2. The Belle II Computing Model and Resource Estimation

The Belle II computing system is expected to manage the process of massive raw data, production of copious simulation as well as many concurrent user analysis jobs. The estimation of the resource requirements shows a similar yearly profile to the pledged resources for the LHC experiments as shown in figure 1. Eventually, we have to handle several tens of Peta bytes of beam data per year.

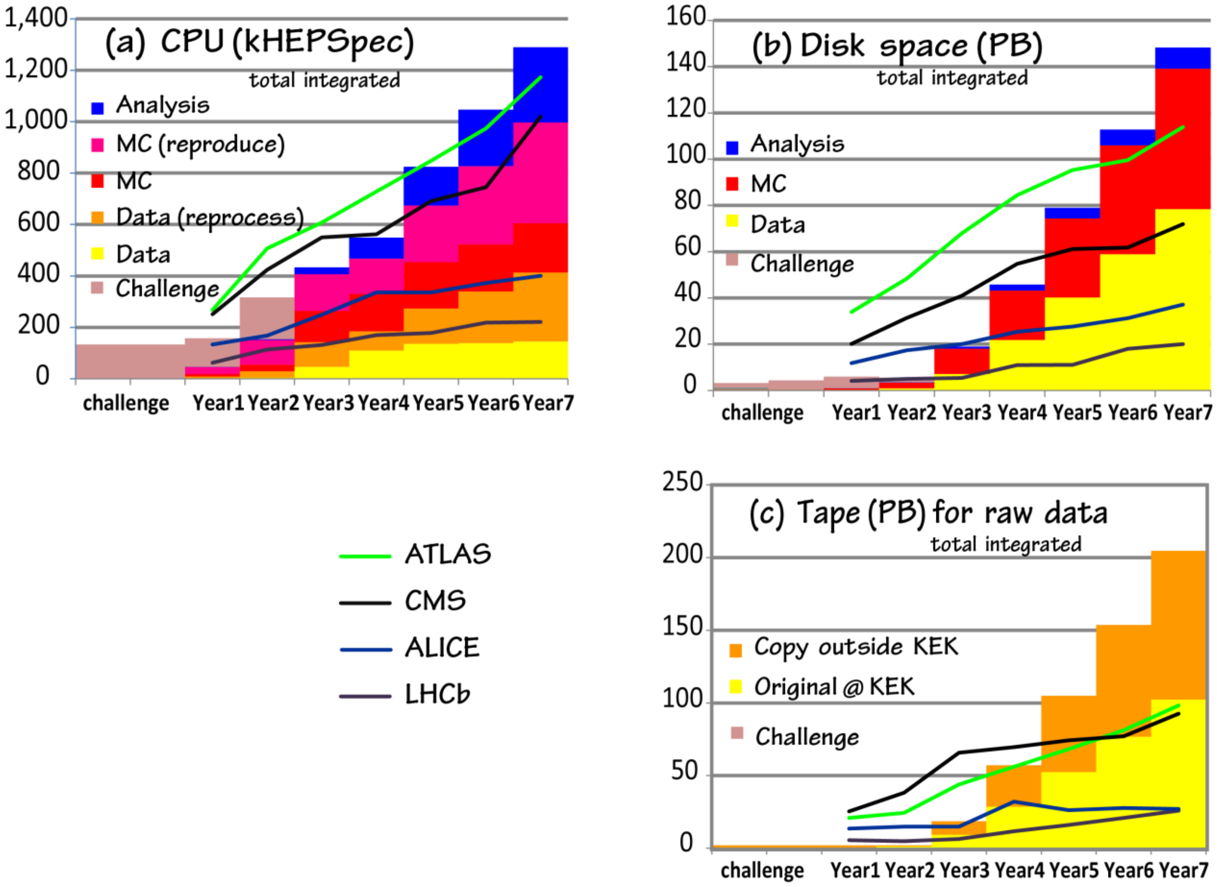

Figure 1. Required resources for the CPU power (a), disk (b) and tape (c) storage space. The real resource capacity and usage can be different. Also the resource pledge summary of LHC experiments (ATLAS, CMS, ALICE, LHCb), based on the published information [6], is superimposed as line graphs with different colors.

Here, the Belle II is a worldwide collaboration of about 600 scientists working in 23 countries and region. It is natural to adopt a distributed computing model based on existing technologies. We chose DIRAC[7] as a workload and data management system and AMGA[8] as a metadata service. For the file replica catalog, we use LFC (LCG File Catalog)[9]. In particular, DIRAC provides us an interoperability of heterogeneous computing systems such as grids with different middleware, academic/commercial clouds [10,11] and local computing clusters as shown in figure 2.

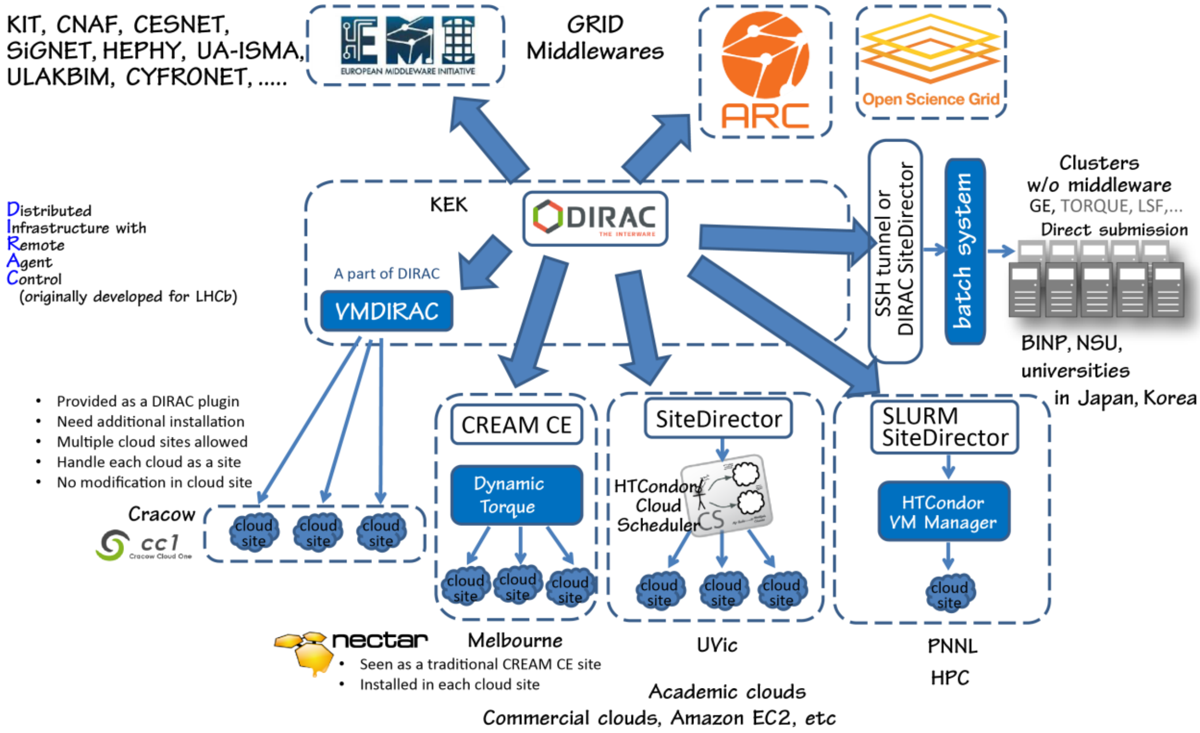
KIT, CNAF, CESNET, SiGNET, HEPHY, UA-ISMA, ULAKBIM, CYFRONET, ...
EUROPEAN MIDDLEWARE INITIATIVE
GRID Middlewares
ARC
Open Science Grid
Distributed Infrastructure with Remote Agent Control
(originally developed for LHCb)
KEK
DIRAC
THE INTERWARE
A part of DIRAC
VMDIRAC
SSH tunnel or DIRAC SiteDirector
batch system
Clusters w/o middleware
GE, TORQUE, LSF,...
Direct submission
BINP, NSU, universities in Japan, Korea
Provided as a DIRAC plugin
Need additional installation
Multiple cloud sites allowed
Handle each cloud as a site
No modification in cloud site
Cracow
cc1
Cracow Cloud One
cloud site
CREAM CE
Dynamic Torque
SiteDirector
HTCondor Cloud Scheduler
CS
SLURM SiteDirector
HTCondor VM Manager
nectar
Melbourne
Seen as a traditional CREAM CE site
Installed in each cloud site
UVic
PNNL
HPC
Academic clouds
Commercial clouds, Amazon EC2, etc

Figure 2. Interoperability of heterogeneous computing systems in the Belle II distributed computing.

The Belle II computing has a hierarchical structure based on the data processing and analysis paradigm as shown in figure 3, which is similar to the Worldwide LHC Computing Grid (WLCG)[12]. We categorize the computing sites according to the assigned role. The "*Raw Data Center*" is the site where the raw data is recorded and/or processed, KEK and PNNL (Pacific Northwest National Laboratory) belong to this category. This must ensure the backup copy of the raw data and accelerate the reprocessing process which will happen later with newer analysis algorithm and more precise detector calibration constants. The output from the raw data processing is stored in "mDST" root-based format containing all necessary information for physics analyses and is distributed to the "*Regional Data Center*" such as DESY and GridKa in Germany, INFN/CNAF in Italy. The computing site, where a proportional share of the MC (Monte-Carlo simulation) production/reconstruction and physics analysis are performed, is defined as "*MC Production Site*". According to adopted technology, the sites are divided into three types, the "*GRID*": a site operated with a standard GRID middleware (e.g. EMI[13], OSG[14]), "*Cloud*": a site operated with a standard Cloud infrastructure, and "*Computing cluster*": a site is a standalone computer cluster which is accessible with the ssh protocol from the internet and available through a batch system such as LSF, TORQUE. Owing to DIRAC, we can handle these different types of computing resources in the Belle II computing model from the beginning.

After three years of operation, we will have a phase shift in the raw data management because of an increased data volume resulted from a higher instantaneous luminosity. In terms of computing, the data acquisition and raw data archiving/processing have the priority at KEK, where the original raw data should be kept. On the other hand, the second copy of the raw data can be distributed not only to PNNL but also to other big computing sites where the reprocessing can be possibly done. It makes the reprocess speed faster. However, we have to consider the management of the output data distributed around the world seriously. Although we are still working on the detailed design for this challenging data management, we plan to distribute the raw data to several computing centers in Germany, Italy, Korea, India and Canada as well as the USA from the 4th year of the operation.

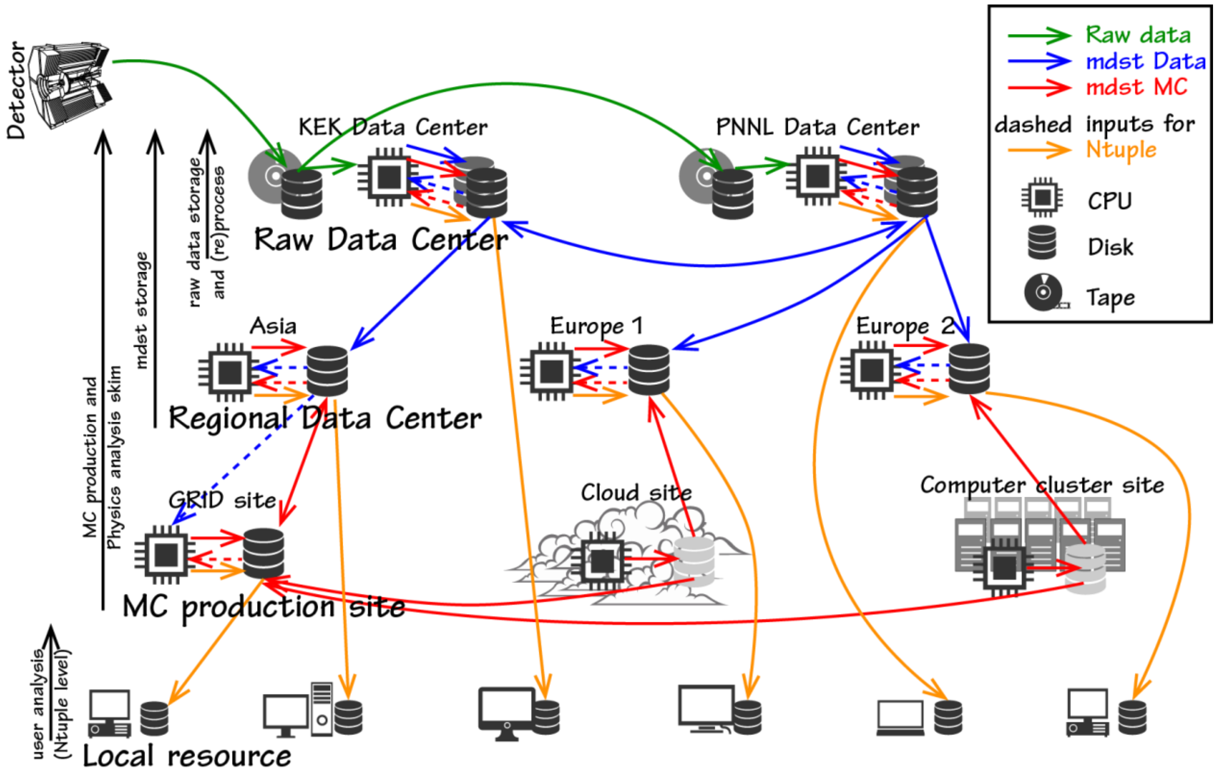

Figure 3. The Belle II computing model (up to the 3rd year of operation)

3. MC production campaigns and data transfer challenge

3.1. MC mass production campaigns

In order to evaluate the Belle II computing model, we have repeated four MC mass production campaigns by now. At the first campaign in February 2013, we observed a job failure rate of roughly 20% caused by failure of the input data download, errors in application software and heavy load during the metadata registration. Through the outcome of this campaign, we could review the computing model by ourselves and fix many glitches in our application and operation software. At the second campaign, we could reach a level of 10% job failure rate, but still we suffered from the failure of the metadata registration and output data upload to the destination SE because of the heavy load. After a consultation of the AMGA developer team, they optimized the metadata query performance by modifying the query search mechanism and the database scheme [15,16]. Concerning the event production procedure, up to the third campaign in March 2014, the MC production shifts submitted the production jobs manually. However, it made some mistakes such as specifying the wrong destination SE accidentally and overwriting the output by submitting the same jobs twice. In order to reduce errors in operation and heavy workload of the expert operators, we decided to implement the "production system", which takes care the production itself once an expert operator defines production parameters, for example, the software version, number of events, and physics modes. The proto-type production system [17] was realized and started operation in the fourth MC production campaign successfully. Thanks to this system, a single person could control roughly 4.7M jobs finally. The monitoring system is also an essential tool to make the production smooth and to utilize the computing resource effectively. We developed two-way monitoring tools, one is "Active monitor" [18] which submits test jobs to check the status of each site including the access to the SE and the necessary network ports, and "Passive monitor" [19] visualizing the time profile of the statistics of the pilot job results and analyzing the job outputs.

In parallel to enormous efforts to improve the Belle II distributed computing system, we had considerable contribution from the computing centers in the world. In particular, at the latest campaign run in later 2014, 31 computing sites in 15 countries/region joined and provided resources. Figure 4 shows the history of the running jobs during the four MC production campaigns in the past three years. The number of produced events in each campaign got increased drastically. Owing to latest many improvements and contributions, we could produce more than 11 billion events corresponding to about 3 *ab*$^{-1}$ experimental data in total. However, there are many things yet to be implemented. One of the important missing features is the data distribution system, which is now being developed.

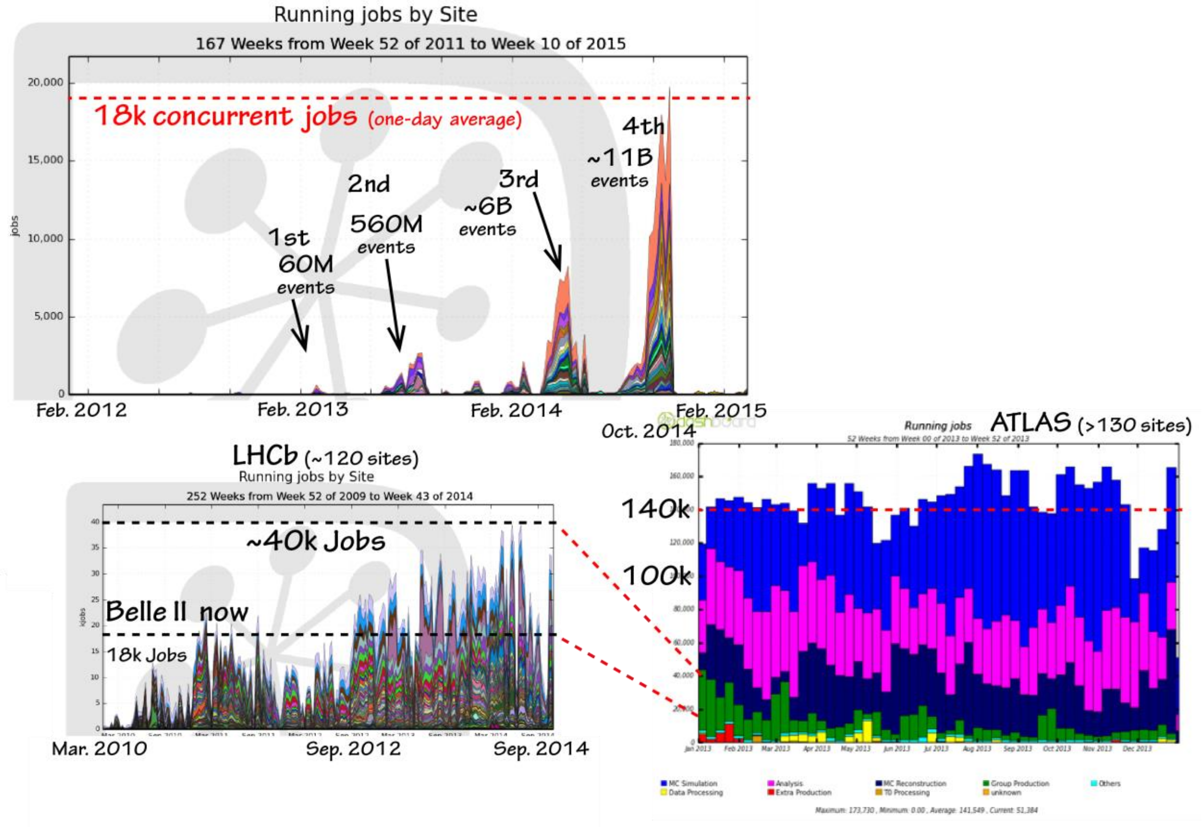

Figure 4. Number of running jobs for the Belle II experiment with a history of the number of produced events in each MC production campaign. As a comparison, the number of running jobs recorded in LHCb (from March 2010 to September 2014) and in ATLAS (from January 2013 to December 2013) are shown.

3.2. Data transfer challenge

The high-speed network connection between each site is indispensable not only for the raw data transfer but also for the MC production and user physics distributed analysis [20]. To understand the present status, we performed a couple of data transfer challenges since 2013. We established a dedicated virtual circuit between KEK and PNNL and tested in 2014. Though we achieved 500MB/s transfer rate from KEK to PNNL almost corresponding to the required network bandwidth in the 3rd year of experiment, this is not enough for the requirement in the 4th year and later. In addition, we experienced that the performance of the storage server at the sender and/or receiver side must degrade the data transfer rate (figure 5). In June 2014, thanks to support from GEANT [21], DFN [22], GARR [23] and ESnet [24], we had chances to exercise the transatlantic data transfer with ANA-100G using FTS3 server at GridKa. The results with iperf reached ~9.6 Gbps, which fulfils the requirement of the network bandwidth between PNNL to *Raw Data Centers* in Europe. However, the data transfer with FTS3 shows a bandwidth of 4Gbps in average. During this test, we faced some difficulties in the setting up of the configuration of the local network apparatus. This test was a good experience for us to recognize the importance of establishment of the smooth and reliable network configuration for the Belle II computing similar to the LHC Open Network Environment (LHCONE)[25]. In September 2014, the Belle II joined the LHCONE as many Belle II computing sites which also support the LHC experiments already joined. Now, we are testing the connectivity between the Belle II LHCONE sites.

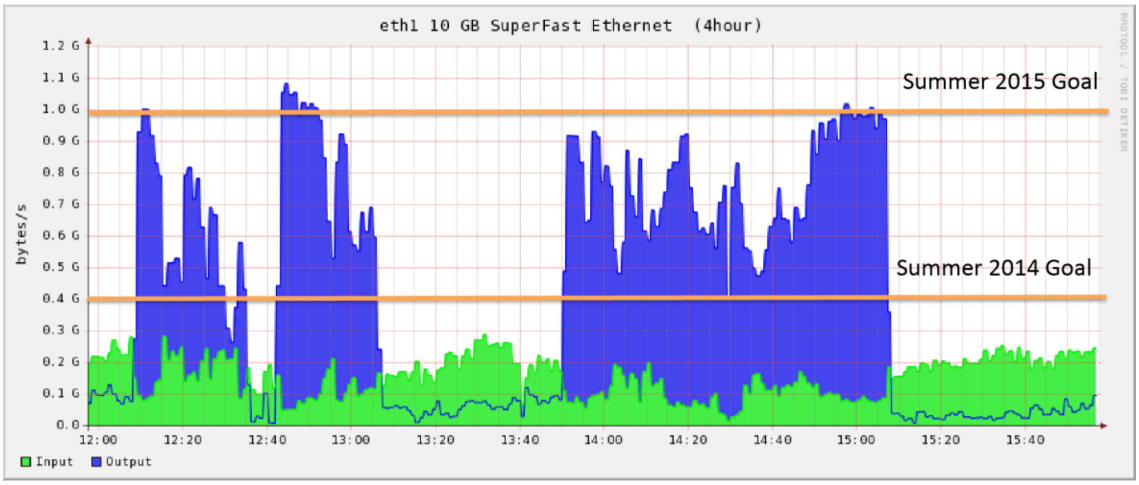

Figure 5. Observed data transfer rate during the transpacific data transfer challenge in 2014. Blue histograms show the transfer rate from PNNL to KEK and the green corresponds to that from KEK to PNNL. We suspect that the asymmetric rate would be caused by the performance of the disk servers at each site.

4. Summary

The Belle II, the next-generation flavor factory experiment, will start the physics run in 2017 and collect 50 times more data than the previous experiment, Belle. The total size of data will be similar or larger level than that of the current LHC experiments. The Belle II computing design is based on a distributed computing architecture with existing technologies. Since the last year, we have started the MC mass production campaigns and data transfer challenges. Through these campaigns and challenges, we could assess the computing system including the application and operation software and the network environment.

5. Acknowledgements

We are grateful for the support and the provision of computing resources by CoEPP in Australia, HEPHY in Austria, McGill HPC in Canada, CESNET in the Czech Republic, DESY, GridKa, LRZ/RZG in Germany, 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, Cyfronet, CC1 in Poland, NUSC, SSCC in Russia, SiGNET in Slovenia, ULAKBIM in Turkey, UA-ISMA in Ukraine, and OSG, PNNL in USA. We acknowledge the 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.

References

[1] Abashian A *et al.* [the Belle Collaboration] "The Belle Detector". *Nucl. Instrum. Meth.* A479, 117-232, **2002.**

[2] Aubert B *et al.* [the BaBar Collaboration] "The BABAR detector". *Nucl. Instrum. Meth.* A479, 1-116, **2002**.

[3] Abashian A *et al.* [the Belle Collaboration] "Measurement of the *CP* violation parameter sin2ϕ_1 in B_d^0 meson decays". *Phys. Rev. Lett.* 86, 2509-2514, **2001**.

[4] Aubert B *et al.* [the BaBar Collaboration] "Measurement of *CP* violating asymmetries in B^0 decays to *CP* eigenstates". *Phys. Rev. Lett.* 86, 2515-2522, **2001.**

[5] Kobayshi M and Maskawa T, "*CP* Violation in the Renormalizable Theory of Weak Interaction". *Prog. Theor. Phys.*, 49, 652-657, **1973**.

[6] http://wlcg-rebus.cern.ch/apps/pledges/summary/

[7] Casajus A, Graciani R, Pateson S, Tsaregorodtsev A and the LHCb DIRAC team, "DIRAC pilot framework and the dirac workload management system". *J. Phys. Conf. Ser.* 219, 062049, **2010.**

[8] Ahn S *et al.* "Design of the Advanced Metadata Service System with AMGA for the Belle II Experiment", *Journal of the Korean Physics Society* 57 issue 4 , 715-724, **2010**

[9] "LCG File Catalog", https://twiki.cern.ch/twiki/bin/view/LCG/LfcWlcg

[10] Sobie R, "Utilizing cloud for Belle II", to be published in CHEP2015 proceedings

[11] Grzymkowski R, "Belle II public and private clouds management in VMDIRAC system", to be published in CHEP2015 proceedings

[12] "Worldwide LHC Computing Grid", http://wlcg.web.cern.ch/.

[13] "European Middleware Initiative", http://www.egi.eu/.

[14] "The Open Science Grid", https://www.opensciencegrid.org/.

[15] Park G, "Directory Search Performance Optimization of AMGA for the Belle II Experiment", to be published in CHEP2015 proceedings

[16] Kwak J, "Improvement of AMGA Python Client Library for the Belle II Experiment" , to be published in CHEP2015 proceedings

[17] Miyake H, "Belle II production system", to be published in CHEP2015 proceedings

[18] Hayasaka K, "Monitoring system for the Belle II distributed computing", to be published in CHEP2015 proceedings

[19] Kato Y, "Job monitoring on DIRAC for Belle II distributed computing", to be published in CHEP2015 proceedings

[20] Hsu C, "The Belle II analysis on Grid", to be published in CHEP2015 proceedings

[21] "The GÉANT Project Home", http://www.geant.net/Pages/default.aspx .

[22] "DFN", The German National Research and Education Network, https://www.dfn.de/.

[23] "GARR", The Italian Academic & Research Network, http://www.garr.it/.

[24] "ESNet, Energy Sciences Network", https://www.es.net/.

[25] "LHCONE : LHC Open Network Environment", http://lhcone.net/.