Proceedings of the Asia-Pacific Advanced Network ***2014*** *v. 38, p. xx-zz.*
http://dx.doi.org/10.7125/APAN.38.x
ISSN 2227-3026

Belle II Computing and requirements of the network

Takanori Hara[1]*
on behalf of the Belle II computing group
[1] High Energy Accelerator Research Organization
1-1, Oho
Tsukuba, Japan

E-mail: takanori.hara@kek.jp
* Author to whom correspondence should be addressed; Tel.: +81-29-879-6209

Abstract:

Toward the start of physics run 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 run since the last year 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.

Keywords: Belle II, particle physics, flavor physics, DIRAC, AMGA, distributed computing, network

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 $L=8\times10^{35}$ $cm^{-2}s^{-1}$ after roughly five years operation and collect a total of $50ab^{-1}$ data by 2023, which corresponds to 50 times more data than the ten-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.

In this paper we will report the current model of the Belle II computing, the estimation of required resources including CPU, storage and network, then the current activities such as mass production of simulation events and the network data 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 required resource estimation for the Belle II computing system shows a similar evolution curve of the resource pledges in LHC experiments as shown in Figure 1. Eventually, we have to handle several tens of Peta byte of beam data per year.

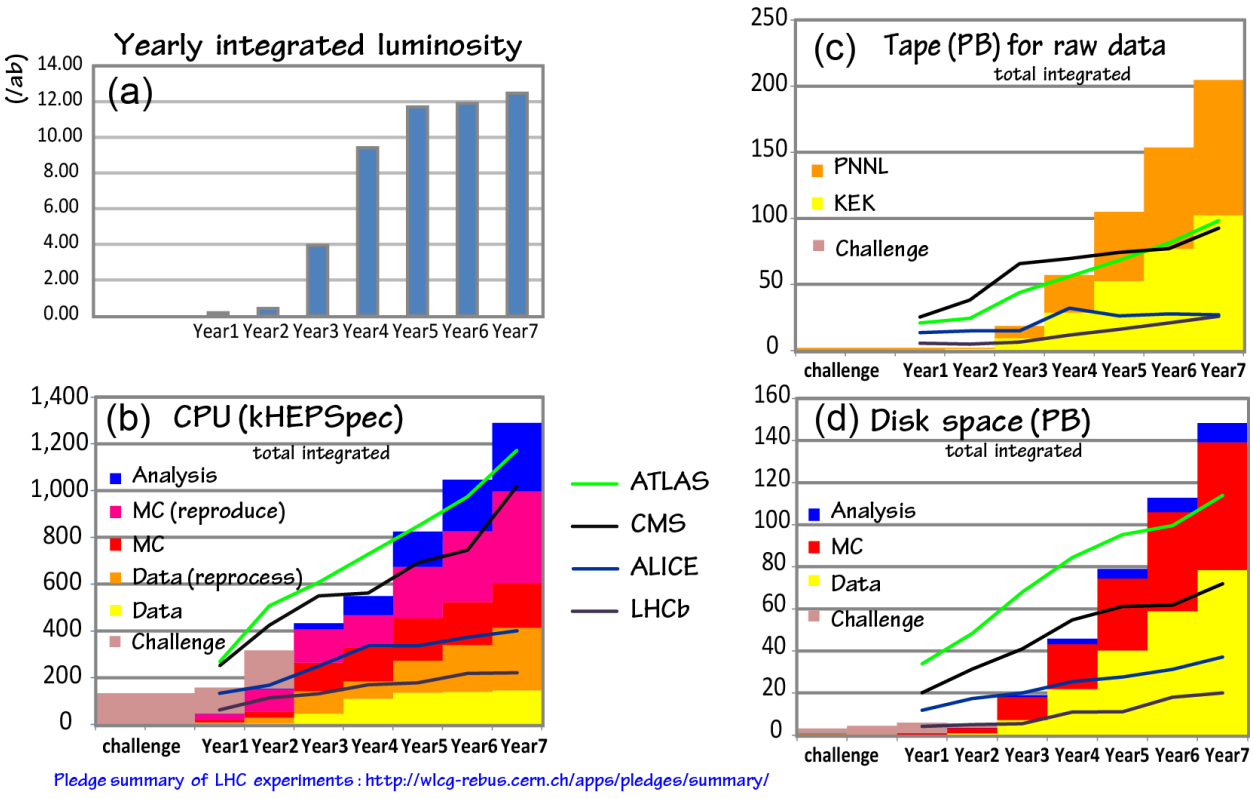

Figure 1. (a) Yearly integrated luminosity prospect of the SuperKEKB accelerator. **(b,c,d)** Required resources for the CPU power **(b)**, tape **(c)** and disk **(d)** storage space. Also the resource pledge summary of LHC experiments (ATLAS, CMS, ALICE, LHCb) is superimposed as line graphs with different 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[6] as a workload and data management system and AMGA[7] as a metadata service. In particular, DIRAC provides us an interoperability of heterogeneous computing systems such as grids with different middleware, academic/commercial clouds and local computing clusters.

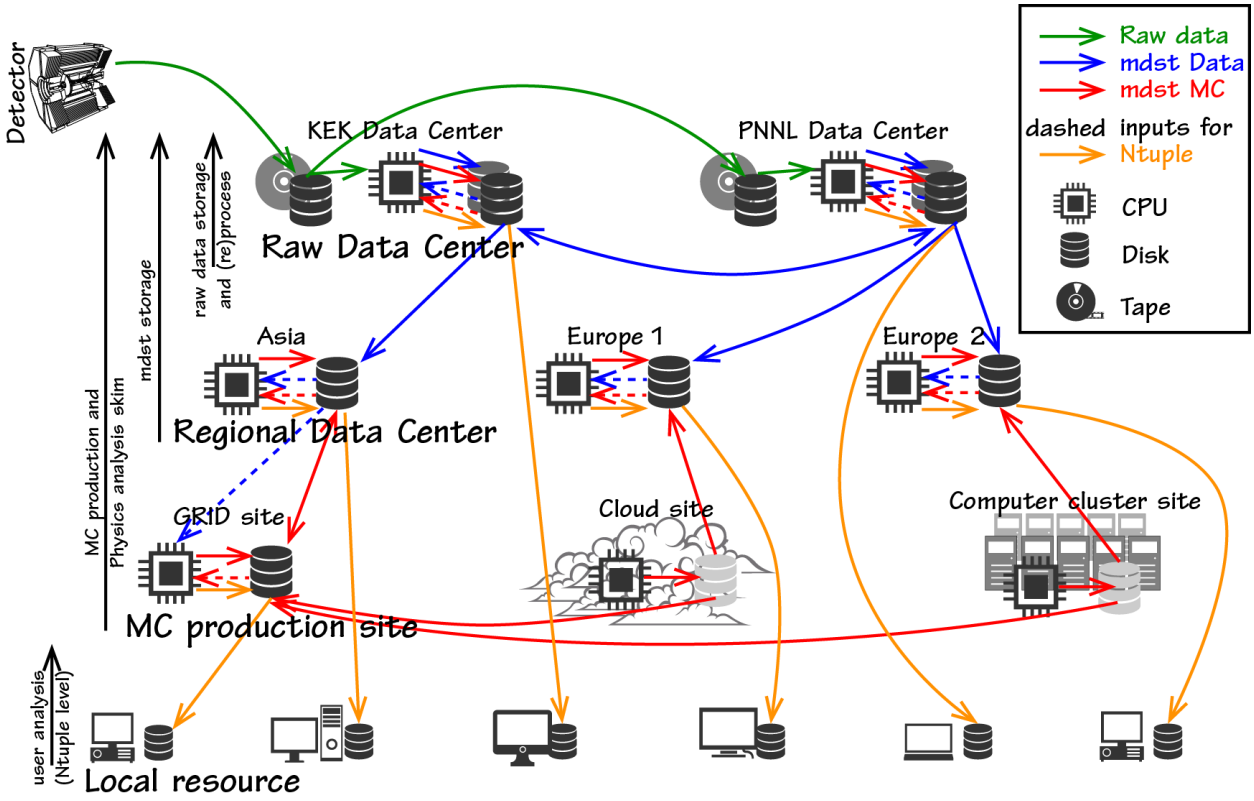

Figure 2. The Belle II computing model (up to the 3rd year of operation)

The Belle II computing has a hierarchical structure based on the data processing and analysis paradigm as shown in Figure 2, which is similar to the Worldwide LHC Computing Grid (WLCG). 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) belongs to this category. This must ensure the backup copy of the raw data and accelerate the reprocessing process which will happen later with new 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, KISTI in Korea. The computing site where a proportional share of the MC (Monte-Carlo simulation) production/reconstruction and physics analysis is performed

is defined as "*MC Production Site*". According to used technology, the sites are divided into three types, the "*GRID*": a site operated with a standard GRID middleware (e.g. EMI, OSG), "*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 has 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 USA from the 4th year of the operation.

3. Requirements on the network

The change of the raw data management strategy requires a reconsideration of the data distribution scheme through the network from KEK to other sites. Up to the 3rd year, we estimate the bandwidth for out-bound data transfer from KEK to be roughly 6 Gbit/s in total, of which a large part comes from the requirement on the data transfer between KEK and PNNL. After the 4th year of experiment, we will start distributing the second raw data copy to PNNL (30%), Germany (20%), Italy (20%), Korea, India and Canada (10% each). In addition, the output mDST files in each *Raw Data Center* has to be copied to two other major regions (Asia, North America, and Europe) so that we can have one full dataset of the whole amount of mDST in each region. Depending on the network path especially on those between KEK (Japan) to sites in Europe, we evaluate two scenarios.

3.1. Scenario 1 (direct data transfer from KEK to Europe)

Assuming that a high-speed direct network connecting Japan to Europe will be constructed in the near future to reinforce the scientific collaboration in both regions, we will be able to distribute the raw data from KEK to sites in Germany and Italy directly. Figure 3 (a) shows an estimation of the required network bandwidth as a function of the operation year of the experiment. The requirement of the bandwidth between KEK and each site reaches several Gbit/s at the 7th year corresponding to the end of the Belle II experiment.

3.2. Scenario 2 (two-step data transfer from KEK to Europe via USA)

In case that we cannot expect the direct high-speed network connection between Japan and Europe by the 3rd year of the experiment, it may be faster to copy the second raw data from KEK to sites in Europe through USA. Currently three 10Gbps links are available between Japan and USA, but SINET plans to upgrade this international line to 100Gbps in April 2016 [8]. Furthermore, the new 100Gbps transatlantic network is being deployed. Figure 3 (b) shows the estimated inbound and outbound bandwidth for each site. As 80% of the second raw data copy will be sent from KEK to USA, the inbound traffic to PNNL would reach almost 14Gbps in the 7th year.

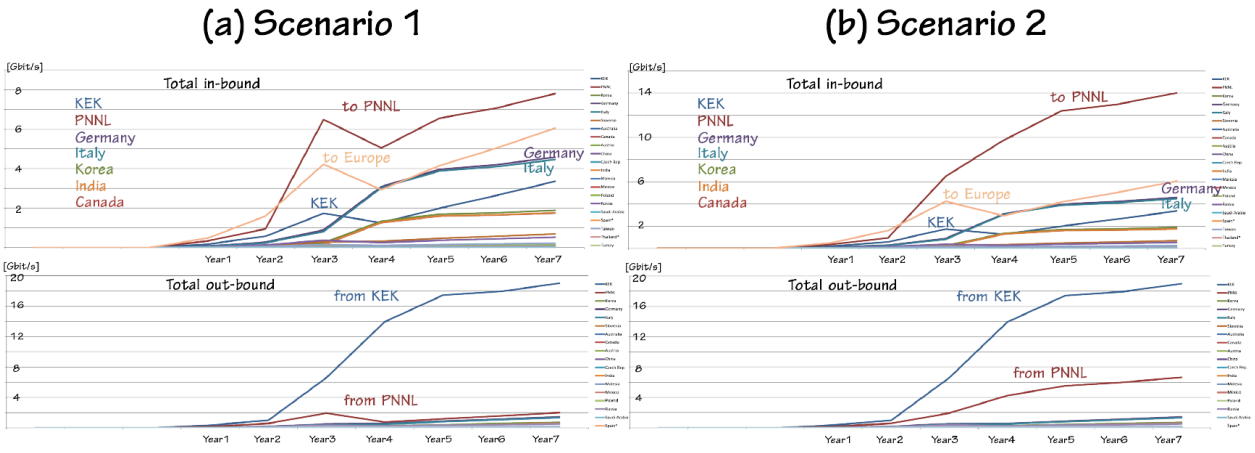

Figure 3. (a) Required inbound and outbound network bandwidth for each site under an assumption that we can distribute the raw data directly from Japan to sites in Europe. (b) the same estimation under another assumption that we will distribute the raw data from Japan to sites in Europe via PNNL in USA. Up to the 3rd year, the estimation is the same for both cases because a 100% second raw data is copied from KEK to only PNNL.

4. MC production campaigns and data transfer challenges

4.1. MC mass production campaign

In order to evaluate the Belle II computing model, we have repeated three 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. Then finally, we could reach a level of 1% job failure rate at the second MC production campaign. As shown in Figure 4, the number of produced events in each campaign got increased drastically. In particular, at the latest campaign run in April and May 2014, 29 computing sites in 15 countries/region joined and provided the CPU power, ~70 kHepSPEC at maximum. Owing to this contribution, we could produce more than 6 billion events corresponding to about 0.8 ab^{-1} experimental data in total.

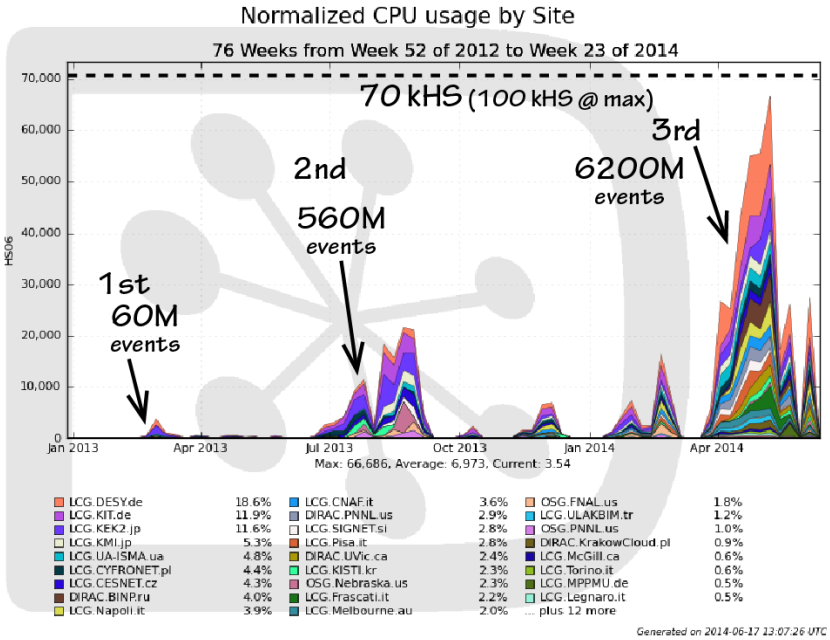

Figure 4 : Normalized CPU power (in kHS06 unit) used for the MC production campaigns.

4.2. Data transfer challenge

As written in Section 3, the high-speed network connection between each site is essential not only for the raw data transfer but also for the MC production and user physics analysis. To understand the present status, we performed a couple of data transfer challenges so far. In 2013, we tested the existing transpacific and transatlantic network. 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 observed a significantly slow bandwidth for the transatlantic network, between PNNL to sites in Europe. In June 2014, thanks to helps from GEANT, DFN, GARR and ESnet, we had chances to exercise the transatlantic data transfer with ANA-100G. We uses "traceroute", "iperf" and "gridftp" with a FTS3 server at GridKa. The results with iperf reached ~9.6 Gbps, which fulfills 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. In addition, during this test, we faced some difficulties in the setting up of the configuration of the local network apparatus and in the optimization of FTS3 parameters. This test was a good experience for us to establish the smooth and reliable network configuration for the Belle II computing. And this also led us to the conclusion that the Belle II prefers to have a closed network like LHCONE [9]

4. Conclusions

The Belle II, the next-generation flavor factory experiment, will start the physics run in 2017 and collect 50 times more data that the previous experiment, Belle. The total size of data will be similar or larger level than that of a current LHC experiment. 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. In particular, to establish the fast, smooth and secure transpacific and transatlantic network connections are essential for the intercontinental raw data transfer.

Acknowledgments

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 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 and Notes

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

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

3. A. Abashian et al. [the Belle Collaboration] "Measurement of the *CP* violation parameter $\sin 2\phi_1$ in B_d^0 meson decays". *Phys. Rev. Lett.* 86, 2509-2514, **2001**.

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

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

7. S. Ahn 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**

8. T. Kozono “SINET Updates”. Network Engineering Workshop APAN 38th Meeting. This proceedings. http://www.jp.apan.net/meetings/1408-TW/SINET-updates_APAN38_20140814_kozono.pdf

9. LHCONE https://twiki.cern.ch/twiki/bin/view/LHCONE/WebHome (accessed on 27 October, 2014).