Belle II distributing computing

P. Krokovny

Novosibirsk State University and Budker Institute of nuclear Physics, 630090, Novosibirsk, Russia

E-mail: krokovny@inp.nsk.su

Abstract. The next generation B factory experiment Belle II will collect huge data samples which are a challenge for the computing system. To cope with the high data volume and rate, Belle II is setting up a distributed computing system based on existing technologies and infrastructure, plus Belle II specific extensions for workflow abstraction. This paper describes the highlights of the Belle II computing and the current status. We will also present the experience of the latest MC production campaign in 2014.

1. Introduction

The existence of large matter-antimatter asymmetry (CP violation) in the $b\bar{b}$-quark system as predicted in the Kobayashi-Maskawa [1] theory was established by the B-Factory experiments [2, 3]. However, this cannot explain the magnitude of the matter-antimatter asymmetry of the universe we live in today. This indicates undiscovered new physics exists. The Belle II experiment [4], the next generation of the B-Factory, is expected to reveal the new physics by accumulating 50 times more data (50 ab^{-1}) than Belle by 2022. The Belle II computing system has to handle an amount of beam data eventually corresponding to several tens of PetaByte per year under an operation of the SuperKEKB accelerator with a designed instantaneous luminosity 810^{35} $cm^{-2}s^{-1}$. Under this situation, it cannot be expected that one site, KEK, will be able to provide all computing resources for the whole Belle II collaboration including the resources not only for the raw data processing but also for the MC production and physics analysis done by users. In order to solve this problem, Belle II employed a distributed computing system based on DIRAC, which provides the interoperability of heterogeneous computing systems such as grids with different middleware, clouds and the local computing clusters. Since last year, we performed the MC mass production campaign to confirm the feasibility and find out the possible bottleneck of our computing system. In parallel, we also started the data transfer challenge through the transpacific and transatlantic networks.

2. Computing model and distributed computing system

A consequence of the drastically increased data rate and volume compared to Belle is an increase in required computing resources by about two orders of magnitude. Because the requirements on storage, processing power, and network bandwidth will be at a similar scale as for the LHC experiments and because the collaboration has become more and more international, Belle II has decided to adopt a distributed computing model.

It has a similar hierarchical structure to the Worldwide LHC Computing Grid (WLCG), but is a bit simpler. The raw data is recorded and processed at KEK and copied only to one other site,

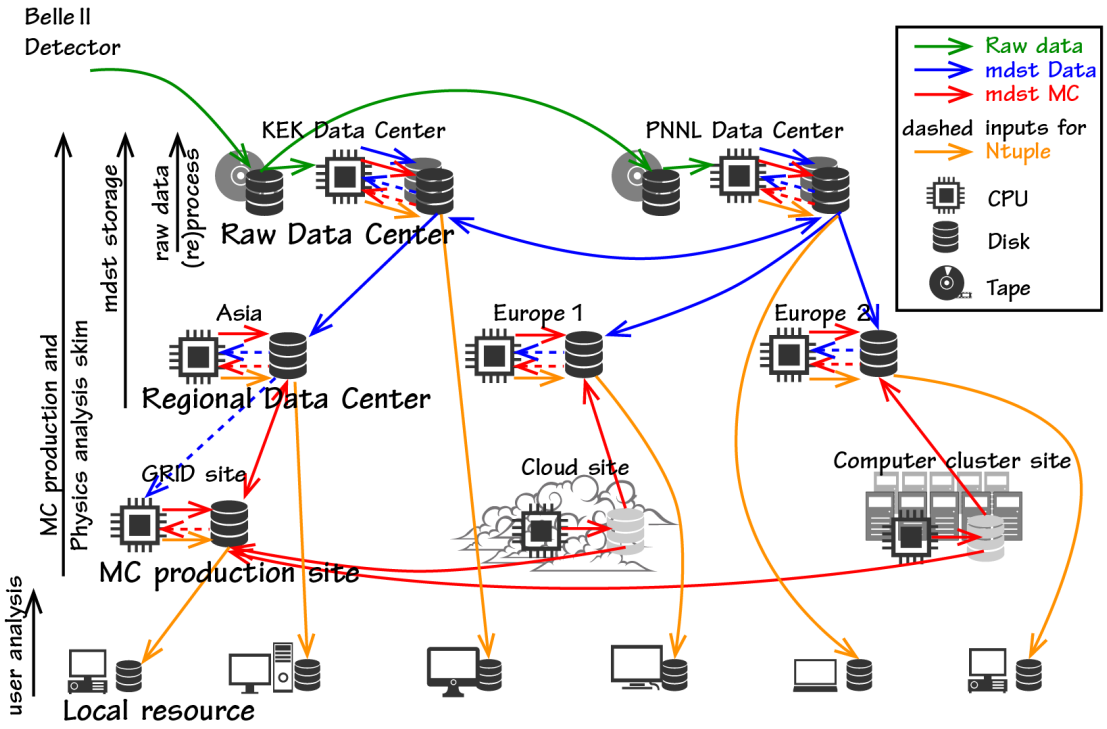

Figure 1. The Belle II computing model.

PNNL (Pacific Northwest National Laboratory). This provides a backup copy and the possibility to distribute reprocessing tasks over both sites. The output of the raw data processing are mDST files, a format containing all relevant information for physics analyses, which are copied to grid sites around the world. These sites also produce simulated data samples (Monte-Carlo, MC) in mDST format. Physicists process the mDST files on grid sites and transfer the output to local resources where the final analysis is performed. The Belle II computing model is illustrated in figure 1.

In the design of our distributed computing system we rely on existing and well-proven infrastructure and software. By choosing DIRAC [5], which was originally developed by LHCb, as the basis for our distributed computing system we can use resources provided via different middleware interfaces, including WLCG, Open Science Grid (OSG), and cloud resources. The location of output files is stored in the Logical File Catalog (LFC). For each file we record metadata in an AMGA database [6]. The installation of offline software releases is done conveniently via CVMFS (CERN Virtual Machine File System) [7], a system for mounting and caching a file system from a remote server, on most sites. The CVMFS stratum-0 server is kindly provided by CERN. A stratum-1 server is running at GridKa. At those sites without CVMFS support, mainly OSG sites, the software is distributed via installation jobs.

On top of these components Belle II has developed a user interface tailored to the needs of the Belle II collaborators. The interface is called gbasf2 and is designed to provide an easy transition from an analysis done with the offline software framework basf2 [8] to an analysis performed on the grid. A guiding design principle was to provide an abstraction of the workflow from single files and jobs to datasets and collections of jobs, called projects. The idea is that the user selects a set of input files based on metadata criteria and then processes them with the same analysis code. Instead of submitting multiple jobs for multiple files, the user starts just one project for the input dataset and then can refer to the collection of individual jobs via a

chosen project name. This project-level management of jobs comprises the submission of jobs, the monitoring of jobs, the cancellation of jobs, and the retrieval of their output sandboxes. The workflow abstraction is also applied to the output files of a project which can be collectively referred to as an output dataset.

3. Production campaigns and data challenges

To evaluate the computing model and the distributed computing system, Belle II has performed a few MC production campaigns. The first campaign was performed in February-March 2013. In a first stage $B\bar{B}$ events were generated with EvtGen [9] and the detector response simulated based on a geometry implemented with Geant4 [10]. The output are raw data files. They are uploaded to storage elements and their metadata information is registered in AMGA. In a second stage the raw data files were the input to reconstruction jobs.

With the experience of the first MC production campaign at hand, the distributed computing and offline software were developed further to address the observed issues. Another assessment of the system was then done in a second MC production campaign from July 23rd to September 8th, 2013. In this campaign the event generation, the detector simulation, and the reconstruction were executed within one job. As the output was stored in mDST format, the ratio of output size to CPU usage is much lower than in the first campaign. A further difference with respect to the first campaign is that background data is mixed with the simulated signal events. This requires the import of background data files at the beginning of each job. The type of jobs executed in the second MC production campaign is similar to the type of jobs that will produce MC data samples for Belle II physics analysis later during the data taking phase. In total 630k jobs consumed 700 kHEPSpec06 days and produced 560M events, corresponding to an output size of 8.5 TB. Compared to the first campaign the used CPU resources could thus be increased by more than an order of magnitude as illustrated in figure 3. At the same time the job failure rate could be reduced considerably. While in the beginning it was at a level of 10% it reached about 1% at the end. Inefficiencies were caused by issues or downtimes of individual sites, the expiration of proxys, overloaded servers, and human errors like the submission of jobs with wrong parameters. It should be noted that no crash of the offline software was observed.

MC 3 and 3.5 were done to provide the first official MC samples to the physics group. It was started by April 3 2014 and finished in May 30 2014. In total 2.5M jobs consumed 2.2M HEPSpec06 days and produced 6000M events, corresponding to an output size of 45 TB. This sample is equivalent to 400 fb^{-1} data. The MC sample equivalent 100 fb^{-1} was open to user analysis.

Several sites around the world participated in the third MC production campaign. These are CoEPP (Australia), HEPHY (Austria), McGill HPC (Canada), CESNET in Czech Republich, DESY, GridKa, LRZ/RZG (Germany), INFN/CNAF (Italy), KEK-CRC, KMI (Japan), KISTI GSDC (Korea), Cyfronet, CC1 (Poland), NUSC, SSCC (Russia), SiGNET (Slovenia), ULAKBIM (Turkey), UA-ISMA (Ukraine), and OSG, PNNL (USA). A summary of the contributed CPU resources is shown in figures 2 and 3.

Like the support of the sites the support of shifters was another essential ingredient for the success of the production campaign. More than twenty volunteers kept production running by submitting jobs, monitoring them, and resubmitting failed jobs. Shifts were assigned in 8 hour blocks so that continuous coverage could be achieved by day shifts in Asia, Europe, and America. Log book information was recorded in a twiki. For communication, in particular at the hand over between shifts, it turned out that a video/audio conference with chat using SeeVogh/EVO provided a very efficient and productive solution.

In the Belle II computing model we rely on high-bandwidth network connections between sites, in particular between KEK and PNNL for the transfer of raw data. But also for the replication of MC datasets network connections between all sites are needed. To assess the

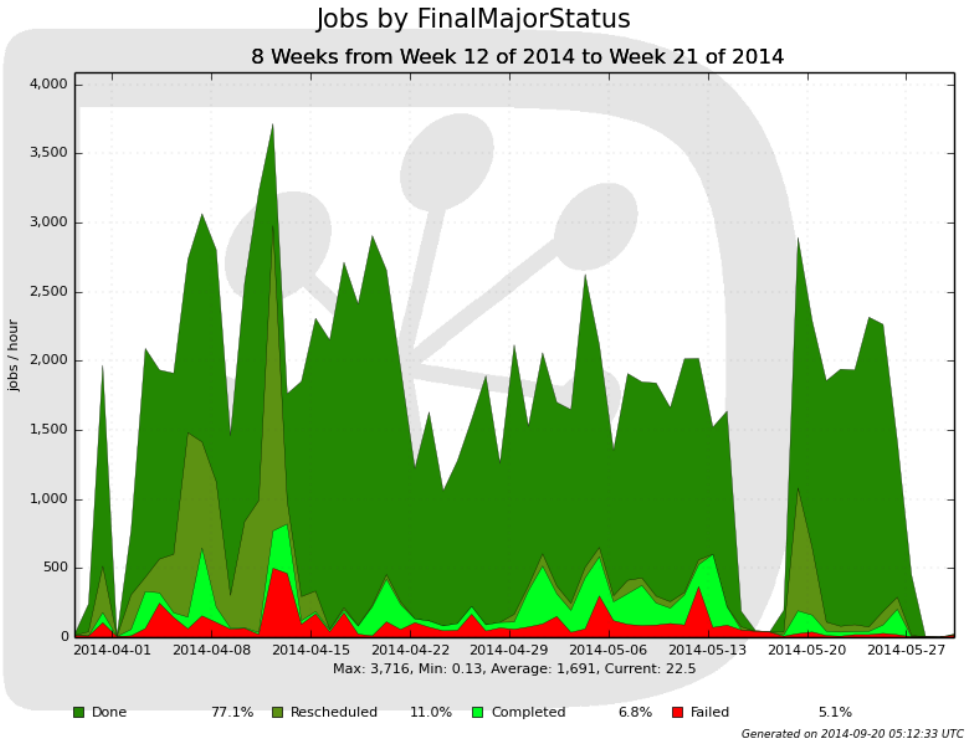

Figure 2. The number of running jobs for the third MC production campaign in April to May. Successfully completed jobs are shown in green and failed jobs in red. Light green indicates that the failover mechanism for the upload of the output was triggered.

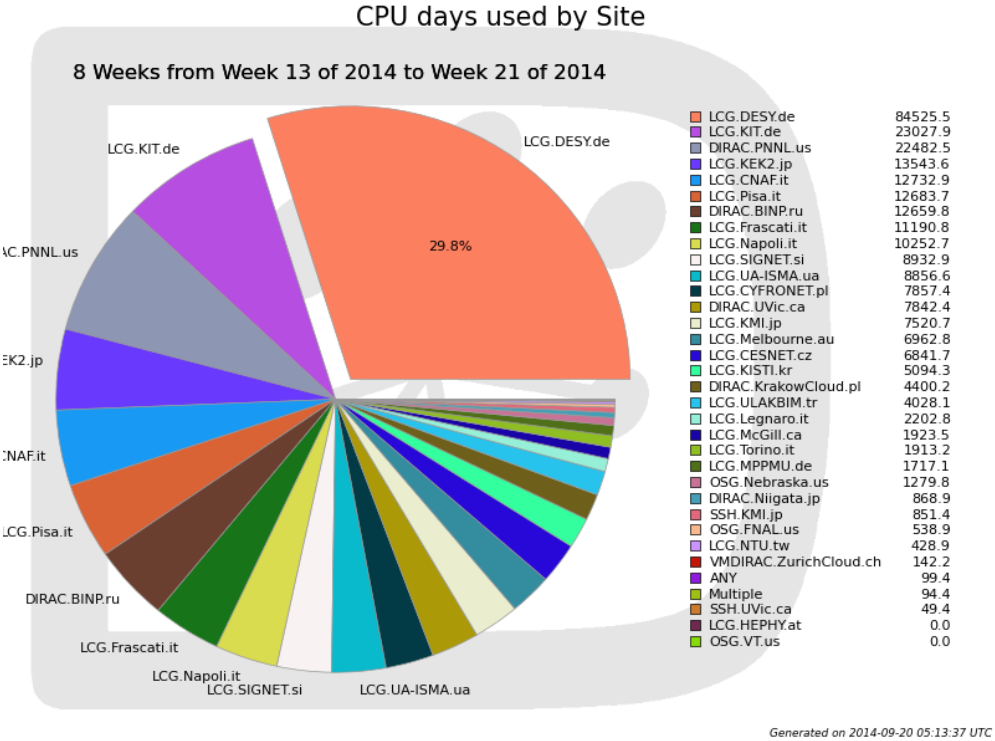

Figure 3. The contributions of sites to the third Belle II MC production campaign.

current status of the network resources available to Belle II, we exercised transfers in a data challenge in May 2013. The transfers were controlled by a FTS2 service running at GridKa. One of the conclusions from the data challenge was that the bandwidth for transfers from Japan to Europe is significantly lower than between Japan and the US.

4. Summary

The computing resources required to process and analyze the data of the next-generation B factory experiment Belle II will be of similar scale as the resources requirements of the LHC experiments. Therefore Belle II has adopted a distributed computing model which enables all collaborators to contribute. The distributed computing system is based on existing technologies and extended by Belle II specific features. An essential concept is the workflow abstraction from individual files and jobs to datasets and projects.

This concept and underlying infrastructure were successfully tested in three MC production campaigns. This success was only possible by the CPU, storage, and network resources kindly provided to us by sites and NRENs. The campaigns showed the issues and bottlenecks of our system and we managed to address several of them already during the production campaign so that a final job failure rate of about 1% was reached.

The fourth MC production campaign is started on September 2014. We include more sites and cloud resources. Another important use case of the system is the user analysis on the grid which may show completely different issues than a centrally managed MC production.

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 Republich, DESY, GridKa, LRZ/RZG in Germany, INFN/CNAF 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. We acknowledge support from MinES of RF (grant 14.610.21.0002, identification number RFMEFI61014X0002).

References

[1] Kobayashi M and Maskawa T 1973 Prog. Theor. Phys. **49** 652
[2] Abashian A *et al.* 2002 Nucl. Instrum. Meth. A **479** 117
[3] Aubert B *et al.* 2002 Nucl. Instrum. Meth. A **479** 1
[4] Abe T *et al.* arXiv:1011.0352 [physics.ins-det]
[5] Casajus A *et al.* [LHCb DIRAC Collaboration] 2010 J. Phys. Conf. Ser. **21**9 062049; Tsaregorodtsev A *et al.* 2010 J. Phys. Conf. Ser. **219** 062029
[6] Ahn S *et. al.* 2010 Journal of the Korean Physical Society **5**7 issue 4 715
[7] Blomer J, Buncic P, Charalampidis I, Harutyunyan A, Larsen D and Meusel R 2012 J. Phys. Conf. Ser. **396** 052013
[8] Moll A 2011 J. Phys. Conf. Ser. **331** 032024
[9] Lange D J 2001 Nucl. Instrum. Meth. A **462** 152
[10] Agostinelli S et al. [GEANT4 Collaboration] 2003 Nucl. Instrum. Meth. A **506** 250