EPJ Web of Conferences **245**, 04005 (2020) https://doi.org/10.1051/epjconf/202024504005
*CHEP 2019*

# The Belle II Raw Data Management System

*Michel* Hernández Villanueva[1,*] and *Ikuo* Ueda[2]
on Behalf of the Belle II computing group

[1]University of Mississippi
[2]KEK IPNS

**Abstract.** The Belle II experiment, a major upgrade of the previous $e^+e^-$ asymmetric collider experiment Belle, is expected to produce tens of petabytes of data per year due to the luminosity increase from the upgraded SuperKEKB accelerator. The distributed computing system of the Belle II experiment plays a key role, storing and distributing data in a reliable way to be easily accessed and analyzed by more than 1000 collaborators. In particular, the Belle II Raw Data Management system has been developed with an aim to upload output files onto grid storage, register them into the file and metadata catalogs, and make two replicas of the full raw data set using the Belle II Distributed Data Management system. It has been implemented as an extension of DIRAC (Distributed Infrastructure with Remote Agent Control) and consists of a database, services, client and monitoring tools, and several agents that treat the data automatically. The first year of data taken with the Belle II full detector has been managed by the Belle II Raw Data Management system successfully. The design, current status, and performance are presented. Prospects for improvements towards the full luminosity data taking are also reviewed.

## 1 Introduction

The SuperKEKB accelerator and the Belle II detector are major upgrades of the KEKB accelerator and Belle detector, located at the KEK laboratory in Tsukuba, Japan [1]. Electrons and positrons collide with a high rate, with an expectation to accumulate a data sample of 50 $ab^{-1}$, approximately 50 times more than its predecessor. With this projected luminosity, Belle II is expected to produce tens of petabytes of real and simulated data per year. The computing system plays a key role in the success of the experiment, as it must manage the large amount of data and MC such that it can be easily accessed and analyzed.

Adopting a distributed computing model, Belle II uses DIRAC [2] as the framework for its distributed computing to interact with distributed resources [3, 4]. The grid system of the Belle II experiment consists of 58 computing sites and 21 storage elements around the world, centrally managed with DIRAC. An extension of the DIRAC framework, named BelleDIRAC [5], handles automated productions and data placement.

Raw data files recorded by the experiment are stored and processed on the grid. We keep two full sets of raw data on the grid for safety and to make it possible to reprocess them even during data taking. To achieve this, the raw data files are uploaded from the offline servers to

*e-mail: mhernan7@olemiss.edu

the grid storage at the host laboratory, KEK, registered in the replica and metadata catalogs, and replicated to a raw data center geographically remote from the host laboratory. For this purpose, the distributed computing group of Belle II has designed a customized system to perform the registration and replication of raw data files efficiently.

## 2 Raw Data Registration and Replication

The Belle II Raw Data Management system is the main component of BelleRawDIRAC, another extension of DIRAC on top of BelleDIRAC, dedicated to registration and replication of raw data files. BelleRawDIRAC consists of a database, services interacting with clients, and several agents working in parallel for each of the steps in uploading, registration and replication. An instance of BelleRawDIRAC runs on a dedicated server, separated from the rest of production activities, ensuring a stable registration of raw data files into the grid. Another dedicated server is prepared to have an instance on stand-by for redundancy.

Figure 1 shows the workflow of the Raw Data Management system and its interaction with other systems inside the Belle II computing model. The upload and registration of raw data start with the submission of a list of files located on the offline servers. Initially submitted by a raw data manager, the registration list is now automatically provided by the Belle II Online-Offline Data Operations system [11]. An agent confirms the accessibility of the file, extracts the metadata embedded on the files, and groups them into data sets based on the corresponding experiment number, run type, and run number. The file is uploaded into a temporal storage element at the host laboratory (KEK) and its information is registered into the replica catalog LFC [6] and the metadata catalog AMGA [7–9]. Once the process is completed, replication to permanent raw data centers is triggered using the Distributed Data Management system of Belle II [10]. One raw data center is located at the host laboratory in Japan and, in the early stage of Belle II, the second one is located at the Brookhaven National Laboratory (BNL) in the USA. Finally, data is reprocessed for user analysis using the Belle II distributed computing system. Figure 2 illustrates the data flow during the registration and replication performed by BelleRawDIRAC.

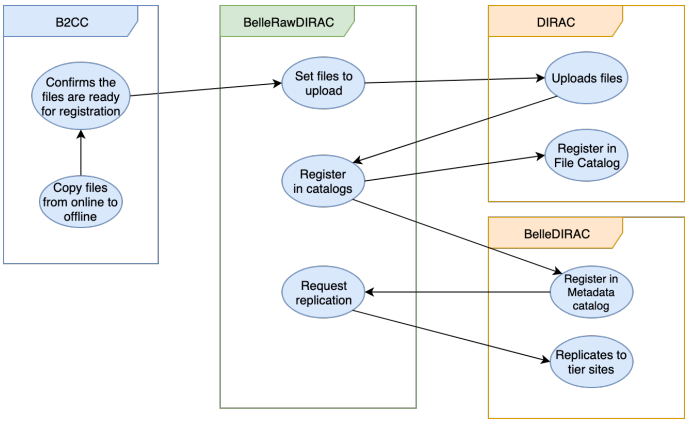


**Figure 1.** Workflow diagram of the Raw Data Management system. The Online-Offline data operations system (B2CC) submits a list of files to be registered and replicated. Then the Raw Data Management system, the main component of BelleRawDIRAC, processes them in communication with systems implemented in DIRAC and BelleDIRAC.

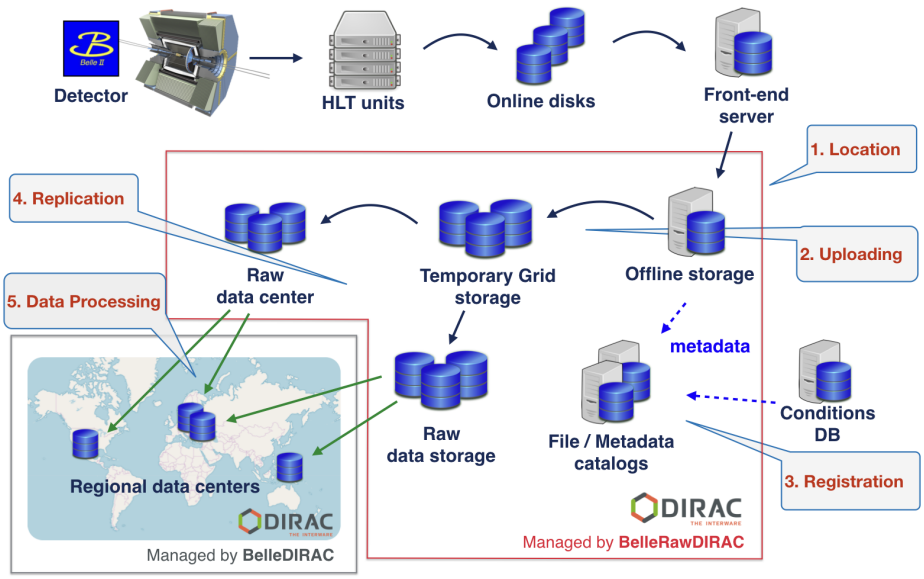


**Figure 2.** Belle II raw data flow through online - offline - grid. Raw data files recorded by the detector are transferred from the online system to offline servers. BelleRawDIRAC takes the raw data files from an offline server and performs the registration and replication. With the files located on the two raw data centers, reprocessing of the raw data files is performed on the grid.

APIs allow the communication between BelleRawDIRAC and the Belle II Online-Offline Data Operations system. Command-line tools have been developed for the interaction between the Raw Data Management system and the raw data manager (human). BelleRawDIRAC agents automatically perform all of the required tasks and reliability checks. The status of files and datablocks[1] are stored in the DB of the system, giving information to the agents regarding which action is required for each file. It also gives information if some error occurs in the process. Figure 3 shows the status diagram used for files during the processing.

## 3 Monitoring

The Raw Data Management system is isolated from the rest of the data production operations by design and is only accessible by authorized data managers. However, data production shifters will monitor the process of registration and replication and, therefore, sending information into the monitoring system of Belle II is required. For this purpose, an agent of BelleRawDIRAC compiles information on the latest operations over raw data files and their status and sends them to the BelleDIRAC monitoring system. Figure 4 shows the communication between the Raw Data Management system and the monitoring system.

## 4 Raw Data Processing during the Early Operation of Belle II

Belle II started operations in 2018 [12]. During the first months of data taking, registration of files was performed using command-line tools implemented in BelleDIRAC and their replication was triggered manually with the Distributed Data Management system. Starting in

[1] A "datablock" is a unit of data management in the Belle II computing model, containing at most 1000 files.

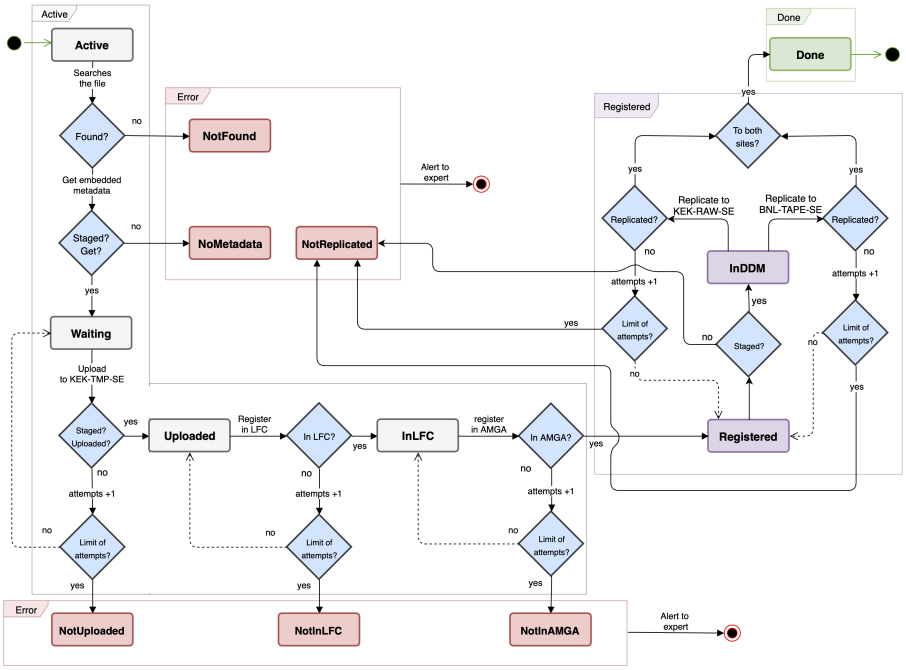


**Figure 3.** Status and minor status diagram for raw data files. Each step is treated by an agent, which changes the status at the end of its cycle.

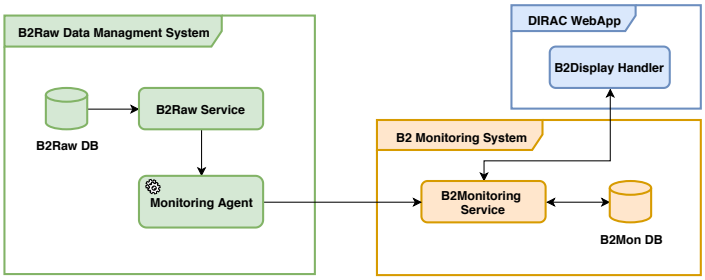


**Figure 4.** The Raw Data Management system sends information to the BelleDIRAC monitoring system, aiming for monitoring the registration and replication of raw data files.

2019, the registration and replication of raw data files are being handled by the Raw Data Management system. The left side of Figure 5 shows the cumulative transfer of raw data starting from the first files registered by BelleRawDIRAC in Feb 2019. Currently, ~400K files, corresponding to 315 TB, have been uploaded to KEK, registered, and then replicated to BNL. The right side of figure 5 shows the throughput of raw data files being uploaded in a typical period with BelleRawDIRAC in normal operation. On average, 253 files are uploaded and registered each hour. With a sustained throughput of ~600 files/hr the duty cycle is 40%.

Therefore, we would be able to handle an increase of twice the current throughput assuming that the system can continuously perform the task at the current rate.

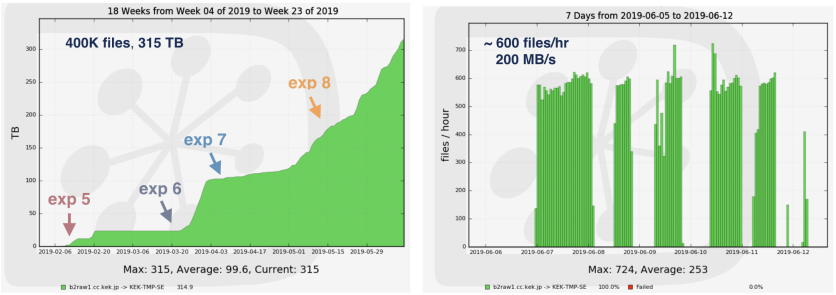


**Figure 5.** (Left) Cumulative data processed by BelleRawDIRAC. To the date, 400K raw data files have been uploaded, registered and replicated, corresponding to 315 TB. (Right) Throughput of upload and registration during a normal operation day. In average, 253 files are being registered per hour, with a sustained throughput of ~600 files per hour.

Figure 6 shows the elapsed time of files being registered. On average, the completion of file replication to the raw data centers since it is registered into the database of BelleRawDIRAC takes one hour. During the operation, bugs and performance issues have been fixed.

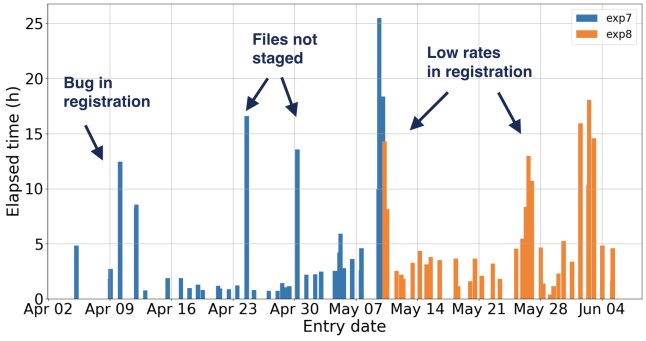


**Figure 6.** Elapsed time in hours for files being registered and replicated by the Raw Data Management system. Each bar corresponds to a list of files submitted from the offline system, and the elapsed time corresponds to the time between the files entering into BelleRawDIRAC to its completion with two replicas created. An abnormally long time indicates an issue during the process. So far, we have identified bugs in the registration, files not staged on disk, and abnormally low rates in the registration product of many files submitted at once.

## 5 Conclusions and Perspectives

The registration of raw data files on the grid and their replication to raw data centers are critical steps in the data production workflow of the Belle II experiment. The Raw Data Management system is the main component of BelleRawDIRAC, taking care of the upload

and registration of raw data files onto the grid and triggering their replication. The system has been working successfully, creating two full replicas of the early data set of Belle II and enabling data reprocessing on the grid. Currently, all of the raw data is being uploaded to the storage element of the host laboratory, KEK, and replicated to BNL.

In the coming months, the Distributed Data Management system of Belle II will be replaced by Rucio [13]. Starting in 2021, one replica will be kept at the host laboratory, KEK, and a second replica will be distributed among several countries [14]. The interaction between Rucio and BelleRawDIRAC is currently under design.

Given the luminosity profile, we may expect an increase in data throughput in the coming months. We are working to improve the performance of BelleRawDIRAC, using additional streams that will handle an increase in the luminosity of the experiment.

BelleRawDIRAC is under constant development to stabilize operation, fix bugs, and provide monitoring tools. Parallelization of the workflow has been successfully implemented, with an agent taking care of each step in the process. During the current operation, no major issues have been observed and tools for manual intervention are ready.

## References

[1] E. Kou et al., *The Belle II Physics Book*, arXiv:1808.10567 [hep-ex]

[2] A. Tsaregorodtsev et al., *DIRAC: A community grid solution*, J. Phys. Conf. Ser. **119** (2008) 062048

[3] T. Hara et al., *Computing at the Belle II experiment*, J. Phys. Conf. Ser. **664** (2015) no. 1, 012002

[4] F. Stagni et al., *DIRAC in large particle physics experiments*, J. Phys. Conf. Ser., **898** (2017), 092020

[5] H. Miyake et al., *Belle II production system*, J. Phys. Conf. Ser. **664** (2015) no. 5, 052028

[6] https://edms.cern.ch/ui/file/579088/1/LFC-Administrator-Guide-1.3.4.pdf

[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** (2010) 715-724

[8] G. Park et al., *Directory search performance optimization of AMGA for the Belle II experiment*, J. Phys. Conf. Ser. **664** (2015), 042030

[9] J. Kwak, et al., *Improvement of AMGA Python Client Library for Belle II Experiment*, Journal of Physics: Conference Series **664** (2015), no. 4, p. 042041

[10] M. Schram et al., *The data management of heterogeneous resources in Belle II*. EPJ Web Conf., **214** (2019) 04031

[11] M. Barrett et al., *The Belle II Online–Offline Data Operations System*. In preparation

[12] F. Abudinén et al., *Measurement of the integrated luminosity of the Phase 2 data of the Belle II experiment*, Chinese Physics C **44** (2020) no. 2, 021001

[13] M. Barisits et al., *Rucio: Scientific Data Management*, Computing and Software for Big Science **3**, (2019) no. 1, 11

[14] S. Pardi, *Network in Belle II*, CHEP 2019 proceedings.