EPJ Web of Conferences **214**, 04031 (2019) https://doi.org/10.1051/epjconf/201921404031
CHEP 2018

The data management of heterogeneous resources in Belle II

Malachi Schram[1,*]

[1]Pacific Northwest National Laboratory - Richland, WA, USA
On behalf of the Belle II Distributed Computing group.

Abstract. The Belle II experiment at the SuperKEKB collider in Tsukuba, Japan, has started taking physics data in early 2018 and plans to accumulate 50 ab^{-1}, which is approximately 50 times more data than the Belle experiment. The collaboration expects it will require managing and processing approximately 200 PB of data.

Computing at this scale requires efficient and coordinated use of the geographically distributed compute resources in North America, Asia and Europe and will take advantage of high-speed global networks. We present the general Belle II the distributed data management system and computing results from the first phase of data taking.

1 Introduction

The Belle II experiment is the successor of the Belle experiment [1] at the KEK laboratory in Tsukuba, Japan. The Belle experiment measured charge-parity (CP) violation in the B^0 system predicted by the theory of Kobayashi and Maskawa [2]. The successful confirmation of the prediction led to the Nobel Prize to both theorists.

The Standard Model of particle physics is an incomplete description of the fundamental forces of nature. The Belle II experiment is planned to collect 50 ab^{-1} of data which requires approximately 200 Petabytes of storage. With this data, physicists will be able to performed precision measurements that will provide stringent tests of the SM and discover or constrain new physics. Precision flavor physics measurements to be performed by Belle II are complementary to the direct search for new particles at the LHC [3]. If new physics is found at the LHC, flavor physics measurements are essential to identify the kind of new physics.

Significant effort is required to develop and operate a computing infrastructure that can store and process this amount of data. The Belle II collaboration has heavily leveraged existing software and computing infrastructure from the LHC experiments. Most notably, the Belle II collaboration has built their distributed computing model around the DIRAC interware [4]. The DIRAC interware provides a collection of systems that orchestrates the use of the computing resources through a common interface. DIRAC is extremely flexible

*e-mail: malachi.schram@pnnl.gov

and expandable which allows for a collaboration, such as Belle II, to implement their unique workflows and policies.

2 Belle II Computing Model

The Belle II collaboration was officially founded in December 2008 and has grown to include over 800 members at 108 institutions in 25 countries. With collaborators geographically located across North America, Asia, Europe, and Australia a distributed computing is required in order to fulfill the processing and storage demands of the collaboration.

Figure 1 illustrates the Belle II computing model for the first three years of operations. KEK and BNL will host a complete replica of the raw data and will process and distribute the data products to the regional data centers. Belle II has chosen DIRAC to

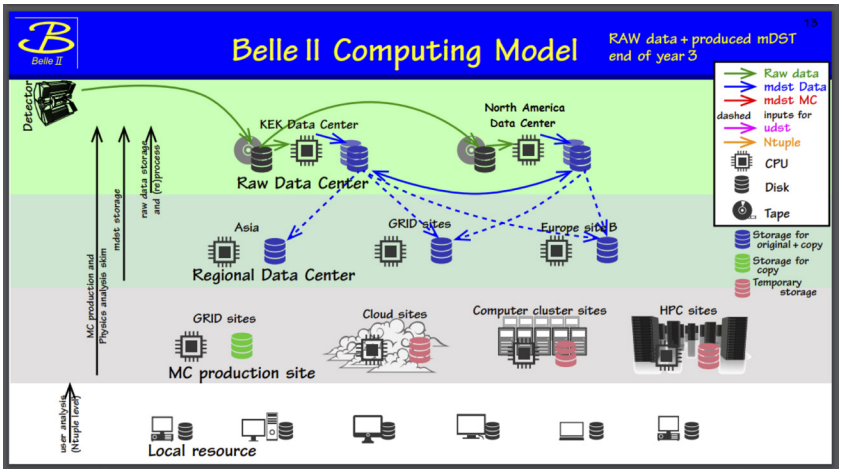

Figure 1: Belle II Computing Model for the first three years of operations.

provide key functionality for their distributed computing model [5]. DIRAC builds a layer between the users and the resources with systems that orchestrates the computing tasks that require specific types of resources. Heterogeneous computing resource providers can be seamlessly integrated into DIRAC using common interfaces. The Belle II computing infrastructure has leveraged and expanded this feature to access various resources as shown in Figure 2. Belle II computing uses dedicated and opportunistic computing resources from a variety of scheduling and computing technologies, such as:

- Schedulers: SLURM, HTCondor, DynamicTorque, CloudScheduler, etc.
- Cloud technologies: Amazon Web Services, OpenStack , etc.
- Container technologies: Docker, Singularity, Shifter, etc.
- High performance computing clusters: Cori and Edison (NERSC), Cascade (PNNL), etc.

In addition, Belle II leverages the existing WLCG storage software infrastructure, such as:

- File Transfer Service 3: Provides mass data transfer scheduling
- Storage Element technologies: BeStMan2, StoRM , dCache, DPM etc.
- File Transfer protocols: gridftp, http, xrootd, etc.

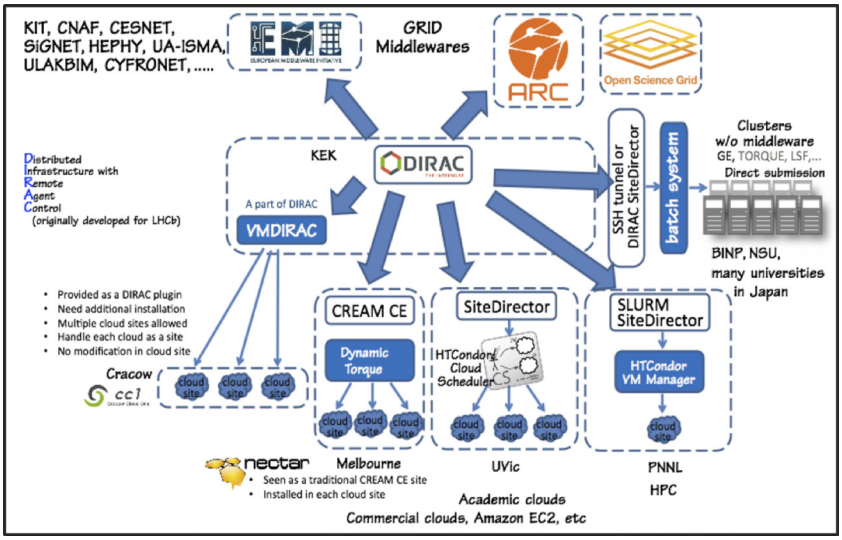

Figure 2: Schematic diagram of the Belle II computing interoperability using DIRAC.

The Belle II collaboration has developed a dedicated extension on top of the core DIRAC framework. This extension provides dedicated tools for users, modified job wrapper code to work with the Belle II software framework, and numerous systems with components that satisfy specific computing and data workflows required by the Belle II collaboration. The Belle II systems include:

- ProductionManagementSystem: provides a coordinate infrastructure to submit large scale computing tasks, such as processing the raw data and generating Monte Carlo samples
- FabricationDystem: manages blocks of compute tasks provided by the production management system
- DistributedDataManagementSystem: centrally manages the data operations from the production management system
- MonitoringSystem: provides a cohesive web application to monitor the Belle II computing systems and resources

Figure 3 illustrates the Belle II distributed computing software and middleware layers including its grid services, resources, and DIRAC components. Distributed Data Management System (DDMS) is a BelleDIRAC extension responsible for managing the Belle II data workflows over the grid. The DDMS was designed to provide a single point of access for all data operations such as replication, relocation, and deletion. This single point of access allows the production group to define the data operation priority and mitigate race conditions. For example, this system prevents a file from being deleted before it was safely replicated to another storage element. In addition to managing the Belle II data workflows, the DDMS provides auxiliary services that actively monitor the health of the storage elements and network. These auxiliary services are critical to the proper operation of the DDMS since they are integrated into the logic for data replication.

As shown in Figure 4, DDMS is implemented on top of the DIRAC core data management system with additional service, database, and agent layers. The DDMS service

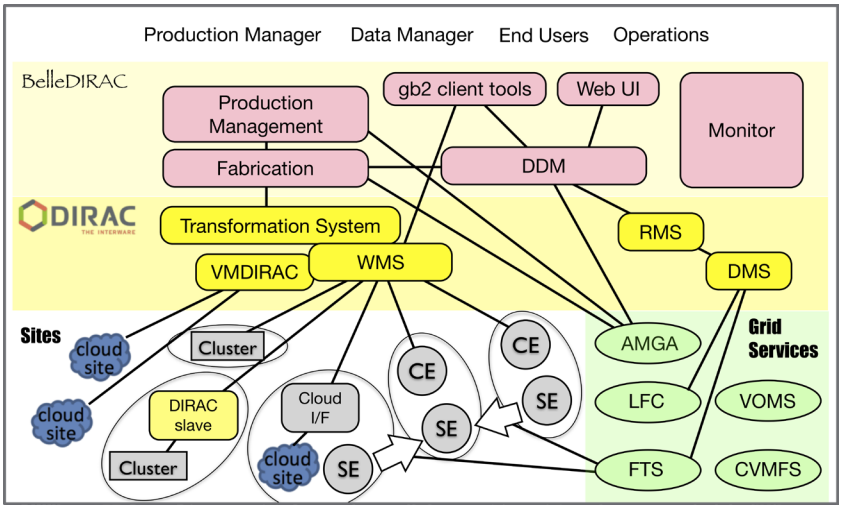

Figure 3: Schematic diagram of the Belle II distributed computing software and middleware layers.

layer provides access to the system and is implemented using RPC services with user and group authentication. There are three primary services use in the DDMS:

1. DataOperationRequest: Allow users to make data operation requests and retrieve their status
2. StorageElementStatus: Allow users to retrieve the state of a storage element and the status of each test performed
3. ReplicaPolicy: Allow users to create a data policy and retrieve the status of existing policies

The service layer is tightly integrated into the databases as most requests are persistified in MySQL database tables. There are three databases used in the DDMS to capture the state and provenance information of each operation:

1. DataOperationDB: Captures the overall state of each request and the state of each individual file throughout the data operation workflow
2. StorageElementStatusDB: Contains the state and the status of each individual test (copy, remove, etc.) performed on each storage element
3. ReplicaPolicyDB: Contains the policy state and provenance information, such as when the policy was created, the data type and level, number of replicas, who created it, etc.

The workhorses of the system are the agents that process and execute the requests. The storage and replica policy agents are autonomous within the DDMS and do not present any potential race conditions in and of themselves.

1. StorageElementStatusAgent: Periodically conducts a series of functional tests on each storage element that is defined in the DIRAC resources and update the database
2. ReplicaPolicyAgent: Submits new data operations based on the active policies

There are several data operations agents that are designed to manage race conditions and potential scaling issues. One of the challenges is to resolve any race condition for a given file with multiple operations such as replicate and delete. DDMS prioritizes operations as per Belle II policy and takes appropriate actions.

1. DataOperationLFNFanoutAgent: Identifies and populates the logical file names for a given data operation request
2. DataOperationTaskFanoutAgent: Creates individual data operation task(s) based on the data operation type, storage element, and data operation priority. This agent ensures that potential race conditions between data operation types are mitigated.
3. DataOperationRequestExecutingAgent: Submits replica requests to the Request Management System (RMS) in DIRAC
4. DataOperationDeletionExecutingAgent: Executes deletion operations within the DDMS
5. DataOperationTaskStatusUpdateAgent: Updates the status of each individual task by querying the DIRAC RMS
6. DataOperationStatusUpdateAgent: Updates the status of the requests

In a nine month period during data taking and intense Monte Carlo production, the DDMS was able to keep-up with the data operations requests of:

- 19.8M operations
- Over 50k/hr replication operations
- Over 12k/hr deletion operations

One reason why the deletion rate is lower than the replication is because of sanity checks performed on each file before deleting it. The sanity check includes making sure that there is another healthy copy of the file on the grid. We expect improved deletion rates by using HTTP over SRM protocol. The simulation samples for the Belle II experiment have been produced in a globally distributed manner, in accordance with the distributed computing model. Figure 5 illustrates the output of the DIRAC-based Belle II production system from January to July of 2018 and is summarized below:

- Average 4.16k concurrent jobs per hour
- Average 18.0k successful transfers per hour
- Current max 24.9k/hr concurrent job
- Current max 31.6k/hr successful transfers (including non-DDM transfers)

3 Conclusions

The increased sensitivity of Belle II over Belle comes with a significantly larger data volume. The SuperKEKB accelerator is designed to provide 50 times more data then its predecessor by the year 2025. The large data volume, computing requirements, and geographically distributed computing resources has led Belle II to adopt a similar computing model as the LHC experiments. As such, Belle II collaboration has developed it's own distributed computing system based on DIRAC interware. Although the development of the Belle II production system is still undergoing, it has satisfied the initial data collection and Monte Carlo requirements. However, there are many necessary improvements and developments to prepare for the full production data taking phase.

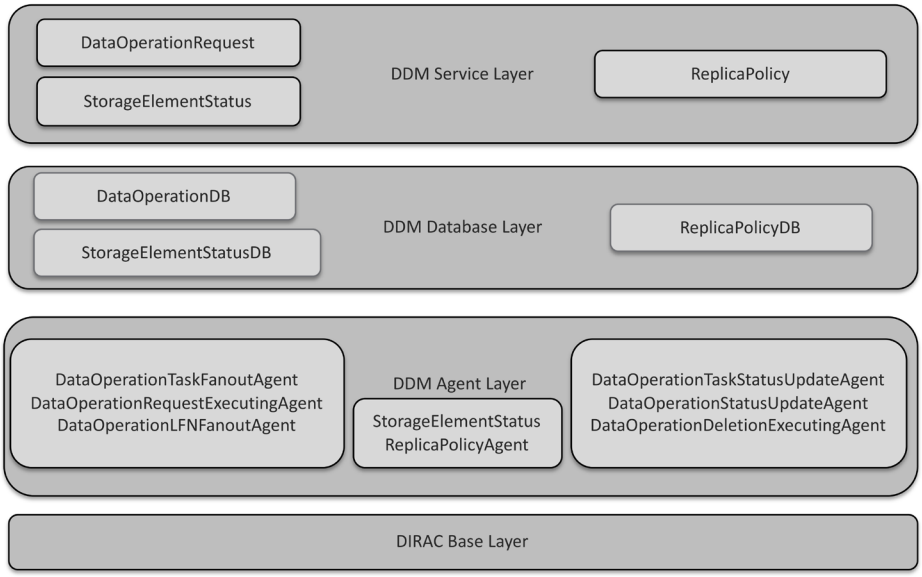

Figure 4: Schematic diagram of the Belle II DDMS components.

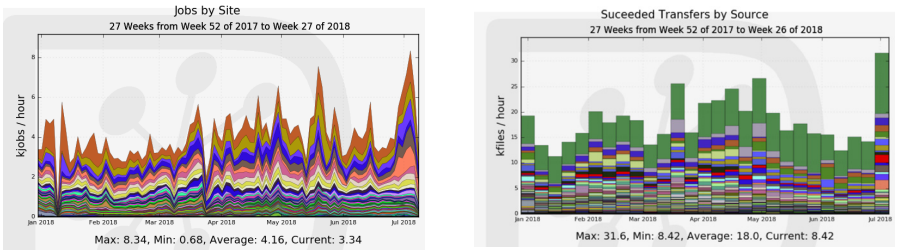

Figure 5: Belle II computing results: (a) concurrent compute jobs and (b) concurrent transfers (including non-DDMS).

References

[1] A. Abashian *et al.*, "The Belle Detector," Nucl. Instrum. Meth. A **479**, 117 (2002). doi:10.1016/S0168-9002(01)02013-7

[2] M. Kobayashi and T. Maskawa, "CP Violation in the Renormalizable Theory of Weak Interaction," Prog. Theor. Phys. **49**, 652 (1973). doi:10.1143/PTP.49.652

[3] L. Evans and P. Bryant, "LHC Machine," JINST **3**, S08001 (2008). doi:10.1088/1748-0221/3/08/S08001

[4] A. Tsaregorodtsev *et al.*, "DIRAC: A community grid solution," J. Phys. Conf. Ser. **119**, 062048 (2008). doi:10.1088/1742-6596/119/6/062048

[5] Takanori HARA and Belle II computing group, "Computing at the Belle II experiment," J. Phys. Conf. Ser. bf 664, 012002, 2015