The management of heterogeneous resources in Belle II

Malachi Schram, Vikas Bansal, Antonio Ledesma

Pacific Northwest National Laboratory 902 Battelle Boulevard, 99352 - Richland, WA, USA

E-mail: Malachi.Schram@pnnl.gov

Abstract. The Belle II experiment at the SuperKEKB collider in Tsukuba, Japan, will start taking physics data in early 2018. The proposed plan is to accumulate 50 ab^{-1} which is approximately 50 times more data than the Belle experiment. The collaboration expects to manage and process approximately 190 PB of data samples. Computing at this scale requires efficient and coordinated use of the compute resources in North America, Asia and Europe. We present the general Belle II computing model, the distributed data management system, and the results of recent network data transfer tests. Additionally, we present how U.S. Belle II is using virtualization techniques to augment computing resources by leveraging Leadership Class Facilities (LCFs).

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 fundemental forces of nature. The Belle II experiment is planned to collect 50 ab^{-1} of data. With this data physicists will be able to performed precision measurements that will provide stringent tests of the SM and discover or contrain new physics.

Precision flavor physics measurements to be performed by Belle II are complementary to the direct search for new particles at the LHC. If new physics is found at the LHC, flavor physics measurements are essential to identify the kind of new physics.

2. Belle II Computing Model

The Belle II collaboration was officially founded in December 2008. Today, it has more than 725 members from over 104 institutes in 24 different countries. With collaborators located in North America, Asia, Europe, and Australia it is distributed around the world.

Beam collisions are expected to start in Spring of 2018. A data sample of about 50 times the size collected by the Belle experiment is expected to be recorded by Belle II by 2025. Its data rate is predicted to be on equivalent to the LHC [3] Run I. Belle II will record events at a

	2017	2018	2019	2020	2021
CPU [kHepSPEC06]	20.11	27.56	58.90	69.71	82.97
Storage [PB]	0.31	0.81	5.04	6.50	9.28
Networking In/Out [Gbps]	0.30/0.30	0.49/0.36	1.06/0.26	1.56/0.31	1.89/0.83
CPU @ PNNL [kHEPSpecs]	0.41	6.35	40.52	29.01	41.58
CPU @ LCF [kHEPSpecs]	19.70	21.21	18.38	40.70	41.39

Table 1. U.S. Belle II computing resources requirements

rate of 6 kHz, corresponding to 11 PB of data per year from 2022 onwards. It will collect 100 PB raw data volume by 2024 with total data volume approaching 190 PB.

Figure 1 illustrates the Belle II computing model for the first three years of operations. KEK and PNNL will host a complete replica of the raw data and will process and distribute the produced data to the regional data centers. An estimation of the U.S. Belle II computing requirements

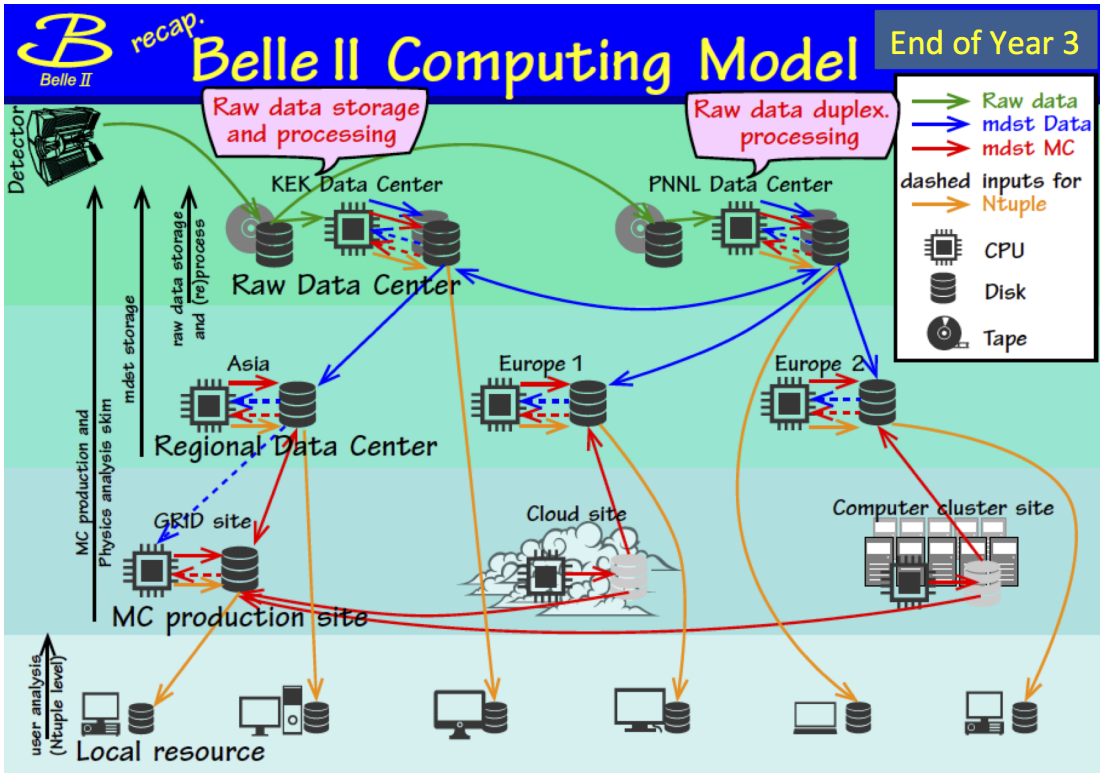

Figure 1. Belle II Computing Model for the first three years of operations.

through 2021 is provided in Table 1. U.S. Belle II also takes advantage of Leadership Class Facilities (LCFs) for MC samples production as show in Figure 2. To ensure that appropriate network bandwidths are available on network routes between major sites in North America, Asia, Europe, and Australia, they are routinely tested for network traffic throughput rates as part of data challenges. Virtual LAN (VLAN) setups were established between sites to perform network data challenges both trans-Atlantic and trans-Pacific using FTS3 [4] service. A site-to-site matrix was deployed to monitor and capture network information for latency and network

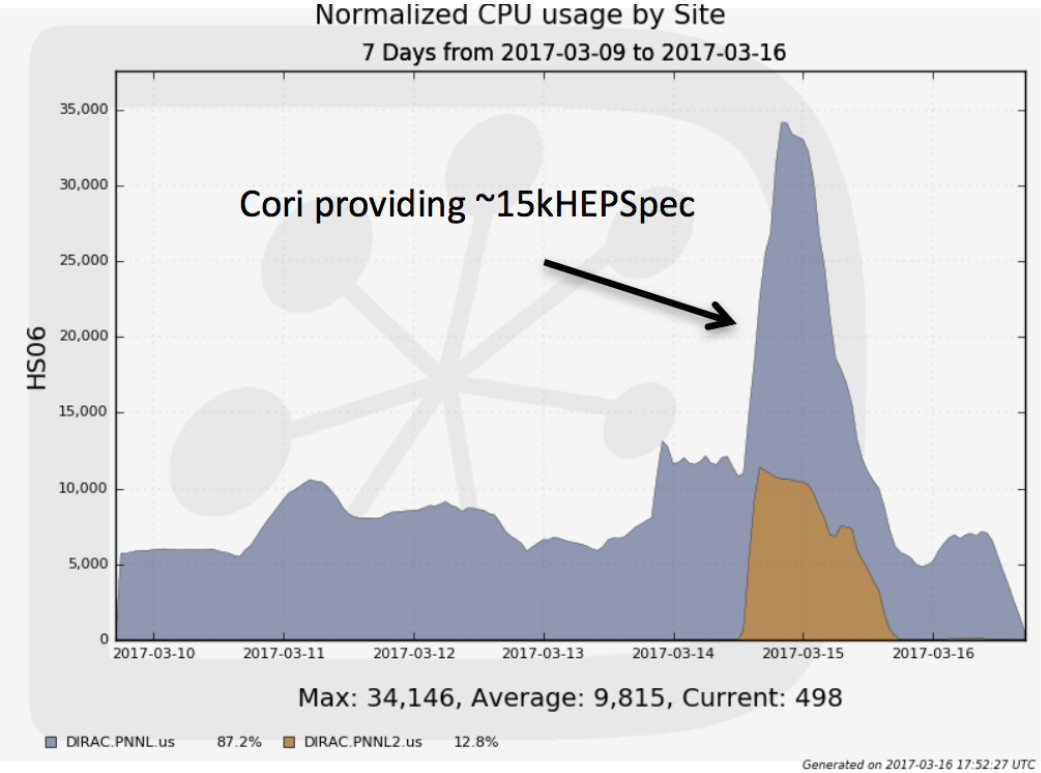

Figure 2. NERSC LCF - Cori providing 15 kHEPSpec to Belle II MC production

packet drop rates using MaDDash [5] and perfSONAR [6] services. Network information between sites will be inserted in a distributed database that will guide optimal data routing path between any two endpoints.

Belle II was recently included in the LHCONE [7] Acceptable Use Policy (AUP). This will enable Belle II member sites to interconnect via the well managed infrastructure of LHC network and hence provide maximal throughput possible. A key component of the LHCONE is that all the sites are "trusted" which alleviates the need for firewalls (as it slows down data transfer rates). Also, bandwidths via public Internet is limited and not feasible to support high data transfer rates. Matrix in Fig. 3 shows recent network data challenge results in terms of achieved bandwidth between some of the major site pairs. The numbers satisfy the bandwidth requirements assuming a 50% safety margin and continuous operations over 10 months period in a year.

Belle II has chosen DIRAC [8] to provide key functionality for their distributed computing model. DIRAC will orchestrate all its sub-components that can as well be distributed at many sites to achieve Belle II computing operations. It will also manage computing resources across all sites.

Belle II collaboration is developing various components and extensions to DIRAC, including a fabrication system that manages compute tasks over all sites, a distributed data management system, web portals for monitoring purposes, and dedicated wrappers around the Belle II reconstruction framework. Figure 4 illustrates an overview of DIRAC grid middle-ware with its major components integrating resources and other grid services to clients.

Distributed Data Management System (DDMS) is the DIRAC component responsible for managing all of Belle II data over grid. It is implemented using RPC service/client based model that is powered by MySQL database server. Many remote agents are designed and deployed

Source→ Destination↓	KEK (Gbps)	PNNL (Gbps)	DESY (Gbps)	KIT (Gbps)	CNAF (Gbps)	NAPOLI (Gbps)	SiGNET (Gbps)
KEK		6.2	11.0	5.0	9.2	15.0	3.0
PNNL	16.0		10.0	6.0	14.0	10.0	-
DESY	6.0	6.6		8.0	8.0	8.0	3.0
KIT	5.6	4.8	8.0		8.0	6.0	3.0
CNAF	18.0	14.0	10.0	6.0		8.0	3.0
NAPOLI	16.0	6.0	3.0	3.0	3.0		3.0
SiGNET	1.6	0.6	5.0	5.0	5.0	2.0	

Figure 3. Measured bandwidth in Gbps between some of the Belle II grid sites

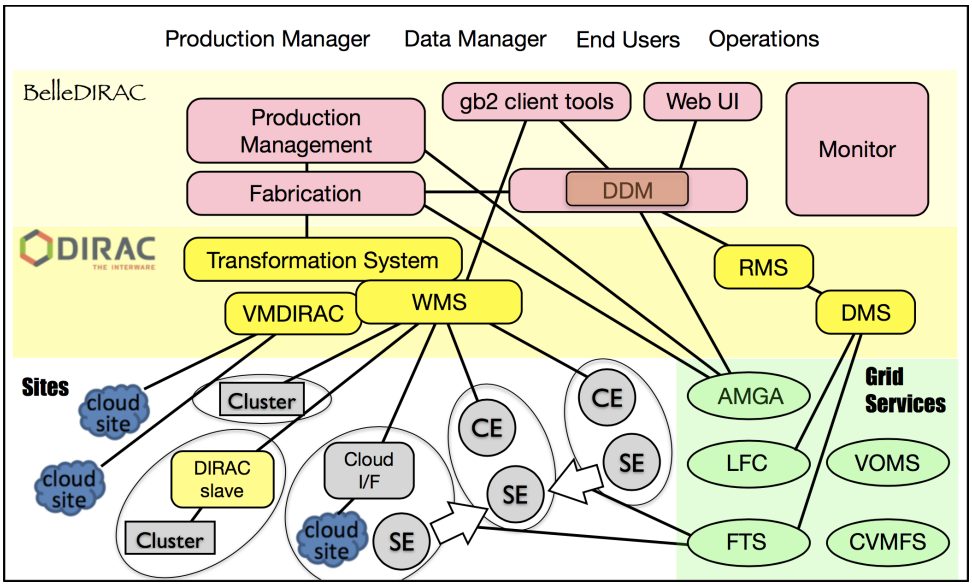

Figure 4. DIRAC, the "interware", between clients at top and grid and computing resources at bottom.

to perform the automation of choosing data destination, performing data transfer and data verification, assessing data popularity and finally cleaning of the data. One of the challenge is to resolve any race condition for a given file with multiple operations such as replicate and delete. DDMS prioritize operations as per Belle II policy and take respective actions. DDMS in on going Belle II MC production activity has shown to achieve 20k/hour file replication rates and 10k/hour file deletion rates. One of the reason deletion rates are lower is because of sanity checks we perform on each file before deleting it. We also plan to boast the rate by using a more popular 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 production system for seven MC production campaigns between 2013 and 2016. About 27 billion events have been simulated so far with concurrent jobs running on heterogeneous grid environment reaching 25K (> 200 kHEPSpec). All produced data is again distributed to storage resources at various sites.

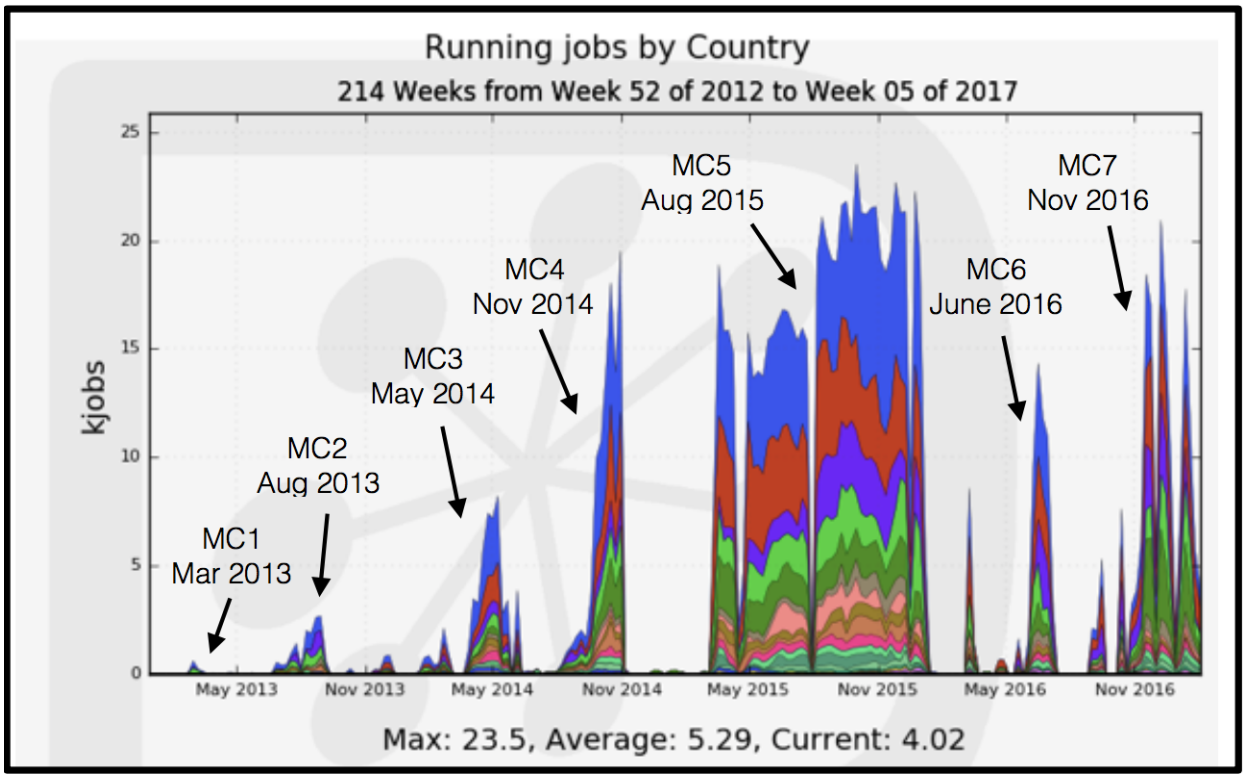

Figure 5. MC Jobs at Belle II MC production campaigns between 2013 and 2016. Y-axis shows number of production jobs and X-axis spans time.

3. Conclusions

The large increase in sensitivity of Belle II over Belle comes with a significantly larger data sample. The SuperKEKB accelerator is designed to provide 50 times more data until the year 2025. This data volume is on par with the LHC Run I and is a challenge for the new computing system. The strategy involves PNNL to host a complete replica of raw data for the first three years in operation and distribute processed data to Europe. Belle II grid brings together heterogeneous computing resources including LCFs. To ensure optimal network bandwidths, we routinely perform data challenges over the Belle II network which is now part of LHCONE AUP. Recent network data challenges achieved Belle II networking requirements. Our computing model is facilitated by DIRAC and we have successfully tested the prototype over MC production campaigns. DDMS, responsible for Belle II data management, is performing as expected.

More details about the computing at Belle II can be found in Chapter 14 of the Belle II Technical Design Report [9].

References

[1] A. Abashian *et al.*, Nucl. Instrum. Meth. A **479**, (2002) 117.
[2] M. Kobayashi and T. Maskawa, Prog. Theor. Phys., **49**, (1973) 652.

[3] http://jinst.sissa.it/LHC/
[4] http://fts3-service.web.cern.ch/
[5] http://software.es.net/maddash
[6] http://www.perfsonar.net/
[7] https://twiki.cern.ch/twiki/bin/view/LHCONE/LhcOneAup
[8] http://diracgrid.org/
[9] T. Abe *et al.*, arXiv:1011.0352 [physics.ins-det].