B-factory Programme Advisory Committee

Evaluation of JFY2021 - 2024 Offline Computing Resource Requirement

Sub-Committee for Computing Resource Review
G. Carlino (Naples), P. Mato (CERN), P. McBride (FNAL),
W. Hulsbergen (Nikhef) and chaired by T. Nakada (EPFL)

17 July 2020

1 Overview

The computing and storage resources needed for the years from 2021 to 2024 had firstly been presented by the Belle II collaboration during the annual B-factory Programme Advisory Committee (BPAC) meeting in February 2020 and was elaborated in a dedicated meeting on June 5th 2020. A final document with the latest luminosity projections has been presented in the BPAC meeting of June 30th 2020. The estimation is based on the foreseen integrated luminosity, the trigger conditions and the computing model parameters.

The SuperKEKB luminosity profile, adjusted to the latest projections, is summarised in Table 1. The main computing model parameters are the event sizes and the processing times. These parameters have been measured with the latest software release 04-00-01 for different classes of events and different background levels. Foreseen software improvements have been taken into account as well.

Year	Apr 2020 Mar 2021	Apr 2021 Mar 2022	Apr 2022 Mar 2023	Apr 2023 Mar 2024	Apr 2024 Mar 2025
$\int \mathcal{L} dt$ (ab^{-1}/year)	0.23	0.92	1.5	3.1	4.1
Cumulative $\int \mathcal{L} dt$ (ab^{-1})	0.24	1.2	2.7	5.7	9.9

Table 1: Expected SuperKEKB integrated luminosity profile for the period 2020 to 2025.

Computing and storage needs in the years 2021 to 2024 have been separately estimated for different activities, in particular prompt processing and calibration of real data, reprocessing of the real data, production of simulated events, skimming of the real and simulated data, and physics analysis.

The CPU resource requirements are mainly driven by the event simulation. The collaboration has defined the types of simulated data samples to be produced and their statistics. The largest sample is composed of generic e^+e^- annihilation events to be generated (event generation, detector simulation and event reconstruction), skimmed and analysed every year with statistics depending on the collected integrated luminosity for the real data. The ratio between the number of events to be generated and the number of real data events is an important parameter in the resource estimates. The collaboration has defined the minimum ratio that is required for physics analysis. The

ratio is largest in the first two years of data taking and will be gradually reduced in the following years with the increase of the cumulative integrated luminosity.

The computing resource requirements for the period from Japanese Fiscal Year (JFY) 2021 to 2024 are summarised in Table 2. It is assumed that the computing activities are evenly distributed over the year, with an efficiency of about 80%. The losses account for inefficiencies of the sites, experimental software and middleware for the distributed computing. This number has been estimated from experience gained in recent years, and is in agreement with the efficiency at the LHC during the first years of data taken.

Year	Apr 2021 Mar 2022	Apr 2022 Mar 2023	Apr 2023 Mar 2024	Apr 2024 Mar 2025
Tape (PB)	3.2	7.2	16	27
Disk (PB)	8.9	19	23	39
CPU (kHS06)	498	511	642	883

Table 2: Estimated requirements on computing resources for years 2021 to 2024.

Given the pledges by the national facilities for JFY2020 amount to 1.4 PB of tape space, 11 PB of disk space and 207 kHEPSpec of CPU, the request for JFY2021 represents no increase in disk space, an increase of about a factor 2.3 in tape space and a factor 2.4 in CPU.

2 Comments from the committee

The reviewers are impressed by the progress achieved by the Belle II collaboration in the last year. Thanks to the success of the experiment, an appreciable amount of real data has become available. This has lead to a reduction of the uncertainty in computing and storage resource requirements for the coming years. For instance, now that the impact of machine background and detector noise on the event size is known, the required disk space is better known as well. Furthermore, implementation of previously planned improvements in the simulation and reconstruction software have reduced uncertainties on CPU estimates.

The main uncertainty that impacts the resource requirements is the evolution of the SuperKEKB performance. The machine group has recently adjusted the luminosity projections for the years 2021 to 2024, leading to a reduction in the luminosity projection per year of up to 30%. This in turn affects the required resources both for real data processing and the production of simulated events. As a minimum number of simulated events is needed for preparations of physics analyses, which dominates the needs, the variations in the computing resource requirements are smaller, but uncertainties remain large nonetheless.

With the recently acquired data as input, the collaboration is actively working on improving the simulation of the detector response. Events simulated with older versions of the software are discarded and replaced by new simulation. As the resources dedicated to the generation of simulated data are costly, the committee encourages the collaboration to review the actual utilisation of large simulated samples in physics analysis and is looking forward to hearing the outcome.

The committee noticed that the collaboration does not plan to reprocess existing simulation samples with new software versions. Saving the output of the event simulation step such that the digitisation and reconstruction can be redone on existing samples would save CPU time at the expense of extra storage space. The committee advises

that the collaboration investigates this strategy in order to prolong the useful lifetime of generated samples and reduce CPU expenses in the future.

In the past, the model used for resource estimates was largely based on processing simulated events. As the experiment and accelerator are now operating successfully, a larger fraction of the resources is spent for processing real data. The committee recommends that the computing model be revisited in view of lessons learned from the first large-scale processing of the real data.

By the end of 2021 the expected data sample of the Belle II experiment will be comparable to that of Belle. It is of prime importance to the experiment that these data can be efficiently analysed such that competitive results on the benchmark channels will be produced soon after data taking. The committee recommends that the collaboration review the physics analysis chain and identify bottle necks that could affect a prompt exploitation of the large data samples.

3 Conclusions

The committee takes note for the computing resource requirements estimated by the Belle II collaboration for the coming four years. The CPU and storage resources requested for the activities foreseen in JFY2021 are well substantiated and the committee recommends the funding agencies to grant them.

Although the model used for the estimation has significantly improved over the last years, the projections for the years beyond JFY2021 still have significant uncertainties