B-factory Programme Advisory Committee Comments on 2020 - 2023 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)

22 June 2019 (revised)

1 Overview

The computing and storage resources needed for the years from 2020 to 2023 have been presented by the Belle II collaboration during the annual B-factory Programme Advisory Committee meeting on February 11th to 13th and in a dedicated meeting on May 9th. Further information on the most up to date projection on the integrated luminosity provided on 19th of June 2019 was also taken into account. 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 assuming eight months of data taking per calendar year. The main computing model parameters are the event sizes and the processing times. These parameters have been measured with the latest software release 02-01-00 for different classes of events and different background levels. Foreseen software improvements are taken into account as well.

Year	Jan 2019 Mar 2020	Apr 2020 Mar 2021	Apr 2021 Mar 2022	Apr 2022 Mar 2023	Apr 2023 Mar 2024
$\int \mathcal{L}dt$ (ab^{-1}/year)	0.10	0.25	2.73	5.66	6.19
Cumulative $\int \mathcal{L}dt$ (ab^{-1})	0.10	0.35	3.08	8.74	14.93

Table 1: The SuperKEKB integrated luminosity profile

Computing and storage needs in the years 2020-2023 are calculated for the following activities: processing of real data, production of simulated events, skimming of the real and simulated data and physics analysis.

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, skimmed and analysed every year with statistics depending on the collected integrated luminosity for the real data. The ratios of the number of events to be generated for the generic simulation sample to the number of real events are larger 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 total resource requirements for the years from 2020 to 2023 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.

Year	Apr 2020 Mar 2021	Apr 2021 Mar 2022	Apr 2022 Mar 2023	Apr 2023 Mar 2024
Tape (PB)	1.4	7.4	19.8	33.3
Disk (PB)	2.2	12.6	17.7	26.3
CPU (kHS06)	207	447	571	675

Table 2: Computing resource requirements

1.1 Comments from the committee

The reviewers are impressed by the progress achieved by the Belle II collaboration in the last year. The computing model has significantly improved thanks to experience gained in the Monte Carlo campaigns and in the processing and analysis of the first real data. Therefore, the committee believes that the uncertainties in resource estimation has greatly been reduced.

Nevertheless, such an estimation suffers of the large uncertainties in the understanding of the machine background whose origin has not yet been completely identified. The analysis of Phase 2 and Phase-3 data will give soon a better understanding of the background allowing a more robust assessment of the computing needs.

In general, the committee strongly recommends to reinforce the effort in code optimisation for simulation and reconstruction with high priority in order to minimise CPU needs.

The main source of required computing resources is given by the number of the simulated generic type events to be produced. As already stated in the last year's report, the committee finds the physics motivations for the base of the estimation reasonable for the moment. Nevertheless it emphasises that the needs must be reviewed in the future.

The estimation of user analysis needs should be revised with the experience gained with the analysis of the Phase-3 data. The chaotic nature of analysis activities makes a realistic determination of the resources complicated. However, the number of concurrent analysis seems large. A mechanism of centralised production of ntuples by analysis working group is suggested.

The computing infrastructure of the Belle II collaboration is growing and is reaching a dimension comparable to the LHC experiments in the first years of data taking.

Therefore, the committee recommends to put in place monitoring systems providing information about the efficiency in the use of the computing resources, such as the CPU usage and the actual use in the analysis of the collected and produced data, including the generated MC samples.

The committee is pleased to note that the collaboration has recently implemented a CPU accounting system based on information extracted from the EGI portal or, for the sites not exporting to EGI, from the KEK Dirac portal. This accounting system provides a detailed monthly breakdown of each site?s performance. However, the choice not to group the sites using the Belle II hierarchical site classification and without any distinction between sites offering pledged or only opportunistic resources could be a source of confusion. The committee recommends to provide separate accounting information for sites offering pledged resources using the Belle II site classification and to explicitly indicate the amount of pledged resources provided by each site or federation of sites.

Furthermore, the committee believes that it will be useful to have data accounting systems based on data popularity information in order to manage efficiently the storage infrastructure and to have a clear understanding of the number of simulated events necessary to fulfil the physics program.

1.2 Conclusions

The committee finds that the computing and storage resources requested by the Belle II collaboration for the activities foreseen in 2020 are reasonable and recommends that they will be granted by the Funding Agencies.

The committee finds that uncertainties are still present in the model used for the estimation and the experience that will be gained in the first full year of data taking in 2019 and the evolution of the computing model might lead to changes in the resource estimations for the years 2021-2023. Therefore, the values shown in Table 2 should be carefully reviewed in the forthcoming reviews.