B-factory Programme Advisory Committee Comments on 2019 - 2022 Offline Computing Resource Requirement

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

16 March 2018

1 Overview

The computing and storage resources needed for the years from 2019 to 2022 have been presented by the Belle II collaboration in a dedicated meeting on 29th of November 2017 and during the annual B-factory Programme Advisory Committee meeting on 12th to 14th of February 2018. The estimation is adjusted to the latest projection of the SuperKEKB luminosity profile summarised in Table 1 assuming nine months of data taking per calendar year. For the event sizes and processing times, measurements with the latest software release 01-00-00 for different classes of events and different background levels are used. Foreseen software improvements are taken into account as well.

Year	2018 Jan-Dec	Jan 2019 Mar 2020	Apr 2020 Mar 2021	Apr 2021 Mar 2022	Apr 2022 Mar 2023
Luminosity (ab^{-1}/year)	0.05	2.87	6.07	9.43	12.08
Total Luminosiry (ab^{-1})	0.05	2.92	9.00	18.43	30.51

Table 1: The SuperKEKB Luminosity profile

Needs in the years 2019-2022 can be divided in the following activities: processing of real data, production of Monte Carlo samples, skimming of the real and simulated data and physics analysis.

The collaboration has defined the types of Monte Carlo data samples to be produced and their statistics. The largest sample is composed of generic Monte Carlo to be generated, skimmed and analysed every year with statistics depending on the collected integrated luminosity. The ratios of the number of events to be generated for the generic Monte Carlo 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 total integrated luminosity.

The total resource requirements for the years from 2019 to 2022 are summarised in Table 2. It is assumed that the computing activities are equally distributed over the year, with an efficiency of about 80%. These losses account for inefficiencies of the sites, experimental software and middleware for the distributed computing.

Year	Jan 2019 Mar 2020	Apr 2020 Mar 2021	Apr 2021 Mar 2022	Apr 2022 Mar 2023
Tape (PB)	10	24	45	71
Disk (PB)	11	21	24	42
CPU (kHS06)	299	432	508	663

Table 2: Resource requirement

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. Therefore, the committee believes that the resource estimation has reached a good level of reliability.

Nevertheless, some uncertainties are still present in the estimation. At this point in time, the expected machine background and the effect of foreseen improvements in the software are still the main sources of those uncertainties. The analysis of Phase 2 data will give soon a better understanding of the background.

Some activities requiring a non negligible amount of resources seem to have still a large margin of improvement:

- Data skimming has much improved with respect to last year with the implementation of a realistic model. However, an optimised model, where a large fraction of skims (at least those belonging to the same analysis working group) are run at the same time producing inclusive datasets, should be tested to see whether the needs of CPU and storage could be reduced.

- 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 committee expects that the estimation will be revised by the collaboration in the next years based on the experience.

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 analysis of the resource breakdown shows that the computing resources are mainly determined by the required number of Monte Carlo events. The Collaboration gives a detailed physics motivation. The ratio of Monte Carlo samples to real data after the first years of data taking has been reduced with respect to previous years requests. Still the increase of the SuperKEKB integrated luminosity will cause a large growth of the Monte Carlo statistics and of the resulting needs of computing resources. The committee accepts the current proposal in the near term but emphasises that the needs must be reviewed in the future based on the experience.

The committee has some minor comments about the processing power needed for the reconstruction of real data:

- It is estimated that the calibration steps takes about 10% of the event reconstruction time. The calibration procedure is actually not part of the reconstruction process and should be treated in an independent way. A realistic estimate of the resources for calibration is required.

- A scale factor (0.83) for the foreseen software upgrade is considered. Evolution of the improvement should be carefully monitored.

1.2 Conclusions

The committee finds that the computing and storage resources requested by the Belle II collaboration for the activities foreseen in 2019 are reasonable and recommends that they will be granted by the Funding Agencies.

The committee finds also that uncertainties are still present in the model and the experience that will be gained in the first year of data taking and the evolution of the computing model might lead to changes in the resource estimations for the years 2020-2022. Therefore, the values shown in Table 2 will be carefully reviewed in the forthcoming reviews.