Run Dependent Monte Carlo at Belle II

Alberto Martini

Deutsches Elektronen–Synchrotron, 22607 Hamburg, Germany

E-mail: `alberto.martini@desy.de`

Abstract. The Belle II is an experiment taking data from 2019 at the asymmetric e^+e^- SuperKEKB collider, a second-generation B-factory, in Tsukuba, Japan. Its goal is to perform high-precision measurements of flavour physics observables. One of the many challenges of the experiment is to have a Monte Carlo simulation with very accurate modelling of the detector, including any variation occurring during data taking. To this goal, a dedicated "run dependent" Monte Carlo has been developed, using the detector conditions during data taking, as well as beam-induced background collected with random triggers. In this article, the procedure for the setup and processing of run-dependent Monte Carlo at Belle II is described.

1. Introduction

The Belle II experiment is located at KEK, Tsukuba, Japan and operates at SuperKEKB, a second-generation B-factory, providing collisions of e^+e^- beams at a specific energy value of 10.58 GeV [1]. A sketch of the accelerator machine and the detector system is reported in Fig. 1. The goal is to explore the high-intensity frontier, performing high precision measurements and new physics searches in many fields (flavour, CP-Violation, rare decays, charm, taus physics) with a very large dataset, thanks to the integrated luminosity of 50 ab^{-1} data, roughly fifty times that of the Belle [2] experiment. Performing world-leading analysis with the Belle II data

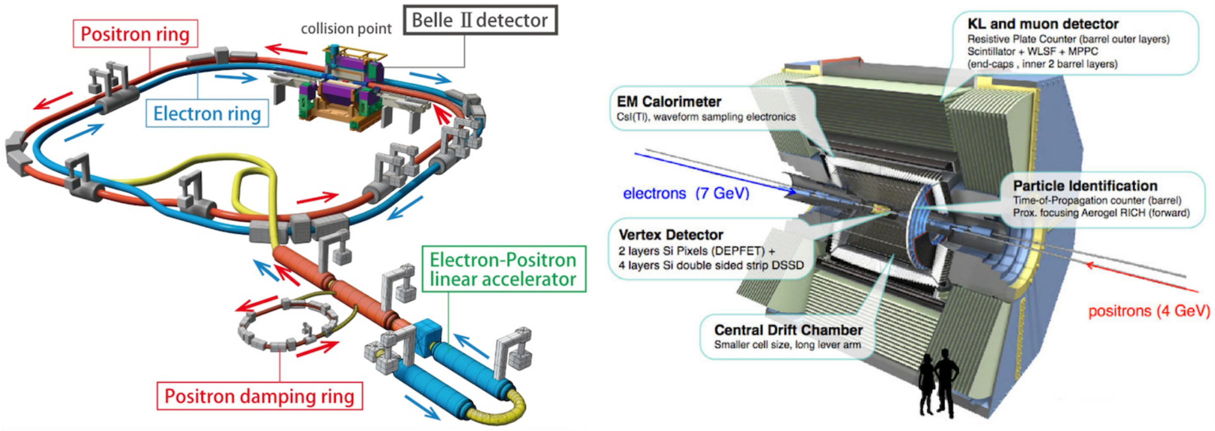
Positron ring
collision point
Belle II detector
Electron ring
Electron-Positron linear accelerator
Positron damping ring
KL and muon detector
Resistive Plate Counter (barrel outer layers)
Scintillator + WLSF + MPPC
(end-caps , inner 2 barrel layers)
EM Calorimeter
CsI(Tl), waveform sampling electronics
electrons (7 GeV)
Particle Identification
Time-of-Propagation counter (barrel)
Prox. focusing Aerogel RICH (forward)
Vertex Detector
2 layers Si Pixels (DEPFET) +
4 layers Si double sided strip DSSD
positrons (4 GeV)
Central Drift Chamber
Smaller cell size, long lever arm

Figure 1. Schematic view of the superKEKB accelerator (left) and the Belle II detector (right).

requires excellent modelling of the experiment, meaning that Monte Carlo (MC) simulations are crucial.

2. Data-production workflow

The raw data collected by the experiment is transferred from the online system to the server placed at KEK laboratories, KEKCC, and then data is registered to the GRID and replicated. A skimming[1] based on High-Level Trigger (HLT) [3] is performed to extract a few streams of raw data, which are used for calibration purposes. These skimmed raw data are prepared at KEKCC, registered to the GRID and transferred to BNL, where the "prompt" calibration is performed. Among them the "delayed bhabha" samples are produced, in raw data format, and used as starting point for the MC run-dependent (MCrd) production. The usage and production of such samples are better described in Sec. 4.2

The calibration step starts with a fast calibration process, called "prompt" calibration, requiring inputs from experts on the detector configuration for the specific data-taking period. Such information is collected as payloads which are stored in the condition database. The data processing of the HLT "hadron" sample will begin immediately after the "prompt" calibration is finished. Data processing of the full data will start thereafter. As soon as the data is ready, analysis skims on data will begin with the highest priority, together with the run-independent and MCrd preparation.

3. MC run-dependent preparation

At Belle II we prepared one major MC production per year, following the software schedules happening at the same rate. As of today, we are producing 2 different types of simulations: MC run-independent (MCri) and MCrd. The former is considering averaged configurations of the detector, as we expect them from pure simulation studies. In addition, the background levels are entirely simulated, based on the best understanding of the background sub-group. Therefore, possible changes in the running conditions of data-taking are not taken into account properly and can lead to data-MC differences in physics analysis. The latter relies as much as possible on data-driven quantities, coming from each sub-detector component. This means having background levels coming directly from data and detector configurations that follow the data-taking periods with high-level precision. The usage of such simulation becomes crucial for high-precision measurements that will require not only a large amount of statistics but also a solid knowledge of the background contributions, required to accurately distinguish them from the signal one.

The MCrd dataset is around four times larger than the measured data collected so far, corresponding to an integrated luminosity of $\sim$1.5 ab^{-1}, and it is being processed for the first time using the GRID [4]. Instead, the run-independent simulation samples are being produced independently of the data sample size, reaching an integrated luminosity of 0.1 to 3 /ab, depending on the process. This was the only simulation that was used by the experiment before the data-taking period started. Producing MCrd is particularly complex for the GRID production system, making the effort very challenging. A general schema of the GRID system is shown in Fig. 2.

In Sec. 4, a description of each step needed to produce such MC simulations is given.

4. MC run-dependent production

Data-taking periods at Belle II are identified via experiment and run numbers. Each experiment corresponds to a specific configuration of the detector or accelerator, and the variation of

[1] Data and MC are processed according to specific analysis requirements in order to reduce the number of events and the size of the samples.

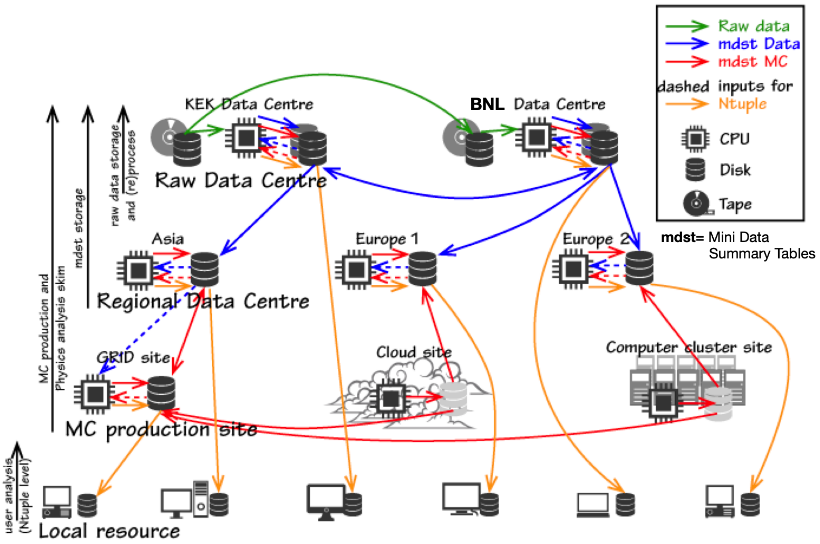

Figure 2. General view of the GRID system and connections with different resources around the world [5].

experimental conditions implies a change in experiment number. Moreover, experiments consist of a set of runs with different sizes and time lengths. The MCrd follows this run structure, therefore it is more complex to prepare and process than the MCri. Each step is described in the following: background preparation, detector configuration and submission on GRID.

4.1. Samples differentiation

The produced samples can be divided into two main sets of MC: generic and signal. The generic ones include the $e^+e^- \to q\bar{q}$/taupair/$B\bar{B}$/low-multiplicity (intended as events with a total number of tracks less than 5) samples, and the corresponding integrated luminosities being produced at Belle II so far are reported in Tab. 1.

Table 1. Summary of the generic samples being produced at Belle II, with corresponding cross sections (σ) and integrated luminosities (L^{int}). The gg corresponds to diphoton production, while $llXX$ and hhISR (Initial State Radiation) includes respectively the channels $ee\tau\tau$, $ee\pi\pi$, $eeKK$, $eepp$, $\mu\mu\tau\tau$, $\mu\mu\mu\mu$ and K^+K^-ISR, $K^0\bar{K}^0$ISR, $\pi^+\pi^-$ISR, $\pi^+\pi^-\pi^0$ISR.

Channel	σ [nb]	L^{int} [/fb]
$q\bar{q}$	3.68	1697
taupair	0.919	1697
$B\bar{B}$	1.05	1697
$\mu^+\mu^-$	1.148	1697
e^+e^-	295.8	42
$eeee$	39.55	424
$ee\mu\mu$	18.83	424
gg	5.1	849
$llXX$	2.01	424
$hhISR$	0.216	424

The production of signal samples is also being addressed, and it is very demanding in terms of processing on GRID. Many signals $O(1000)$ are produced, a number significantly larger than the generic ones $O(10)$, but typically each production has fewer events $O(10^6)$ than the generic ones $O(10^9)$. On the other hand, the events are spread over all runs/experiments, so a large number of short processes are performed, stressing the production system. So far, most of the signals were processed as MCri, but about 10% have been produced as MCrd: we expect this number to increase in the future.

4.2. Background preparation

The background contributions to MCrd are extracted from data, using a "delayed bhabha" trigger to select the background events. These correspond to events acquired after a time interval from a bhabha trigger, and they are stored at a 5 Hz frequency. The rate is limited by the throughput since the full waveform from the electromagnetic calorimeter, which is a heavy and time-consuming process, is saved. Thereafter, data are being "unpacked" to get the digit information associated with the tracks and clusters in the event. Finally, the digit information from the background is overlayed to the MC events to be produced, in order to have the simulation with data-driven background levels.

This production is done per experiment number and it is not very demanding in terms of processing time and CPUs.

4.3. Detector configuration

The most time-consuming part of the process concerns gathering together the calibration constants dedicated specifically to MCrd productions. In fact, each sub-detector is asked to provide data-driven payloads such as alignment, dead channel mappings and many more. This is currently one of the main bottlenecks, and future improvements will definitively reduce the time needed for production.

4.4. Submission on GRID

Once all the payloads are collected, the MCrd production scripts configure the production. They include information about the total luminosity to be produced, the length of each run to split properly the events, as well as the amount of background events available for each run to avoid too large reuse rate. The GRID needs to process a very large amount of jobs since they correspond to the number of runs.

The final submission to the GRID happens after a test is performed on a dedicated server, and the produced simulations are correctly uploaded. More details on the production system can be found in [5].

5. Conclusions

The Belle II experiment is collecting a huge amount of data and the proper production of MC samples is a crucial part of the analysis work being carried out by the collaboration. For this reason, a MCrd production system is built and, although being complex and computationally demanding, providing us with the best simulations. In fact, the Belle II MCrd samples replicate the run-by-run evolving detector conditions, therefore reducing the data-MC discrepancies significantly. With ∼400 fb^{-1} of data collected by Belle II as of today, we already produced MCrd samples corresponding to ∼1700 fb^{-1}, for the high multiplicity samples.

References

[1] Abe T *et al.* (Belle-II) 2010 *eprint* (*Preprint* 1011.0352)

[2] Altmannshofer W *et al.* (Belle-II) 2019 *PTEP* **2019** 123C01 [Erratum: PTEP 2020, 029201 (2020)] (*Preprint* `1808.10567`)

[3] Itoh R *et al.* 2020 *EPJ Web Conf.* **245** 01040

[4] Tsaregorodtsev a *et al.* 2008 *Journal of Physics: Conference Series* **119** 062048

[5] Miyake H, Grzymkowski R, Ludacka R and Schram Malachi 2015 *Journal of Physics: Conference Series* **664** 052028