The Neural Network First-Level Hardware Track Trigger of the Belle II Experiment

Christian Kiesling[*1]

*Max-Planck-Institute for Physics (MPP), Garching, Germany

E-mail: cmk@mpp.mpg.de

Abstract. We describe the principles and performance of the first-level (L1) hardware track trigger of Belle II, which uses the information of Belle II's Central Drift Chamber (CDC) and provides three-dimensional track candidates based on neural networks. The inputs to the networks are 2D track candidates in the plane transverse to the electron-positron beams, obtained via Hough transforms, and selected information from the stereo layers of the CDC. The networks then provide estimates for the origin of the track candidates in direction of the colliding beams (z-vertex), as well as their polar emission angles θ. Using a suitable cut d on the z-vertices of the "neural" tracks allows us to identify events coming from the collision region ($z \approx 0$), and to suppress the overwhelming background from outside. The networks also enable a minimum bias trigger, requiring a single 2D track candidate validated by a neural track with a momentum larger than 0.7 GeV in addition to the $|z|$ condition. We also sketch our concepts for upgrading the neural trigger in view of rising instantaneous luminosities, accompanied by increasing backgrounds.

1 Introduction

SuperKEKB[1], an upgrade of the B-factory KEKB[2], is operating since 2019 and continues to produce increasing world-record luminosities, targeting 6×10^{35} cm^{-2}s^{-1}, at the center-of-mass energy of 10.58 GeV. This collision energy corresponds to the mass of the $\Upsilon(4S)$ resonance, which decays predominantly into pairs of B mesons, observed with the Belle II detector[3]. In addition, a continuum of the lighter mesons is produced in the annihilation of electrons and positrons, and the important two-body leptonic final states $\mu^+\mu^-$ and $\tau^+\tau^-$.

Apart from the decay products of some long-lived particles, the collision products have their origin (vertex) within the small collision volume of the electron-positron beams (interaction point IP). There is, however, a sizeable background caused by interactions of the beam particles along the beam (z-direction) with the residual gas in the beam pipe, or with elements of the magnetic beam guiding and focusing system. When these interactions occur not too far from the IP ($z = 0$), they may emit particles into the Belle II detector and create tracks similar to the desired annihilation events. The size of this effect is hsown in fig. 1. The events were triggered by the conventional 2D trigger of Belle II, which is derived from the axial wires (parallel to the z-axis) of the central drift hamber (CDC).

We report here on the new global L1 track trigger for Belle II, using neural networks, which estimates the z-component of the tracks in beam direction (z-Trigger). It uses inputs from the 2D tracks of the conventional 2D trigger, and in addition a subset of the stereo wires of the CDC, to estimate the z-impacts and polar emission angles of the tracks.

---

[1]for the German AI Trigger Group, KIT, Karlsruhe: Jürgen Becker, Torben Ferber, Greta Heine, Marc Neu, Lea Reuter, Slavomira Stefkova, Kai Unger, MPP: Simon Hiesl, Timo Forsthofer, Christian Kiesling, TUM: Alois Knoll, Felix Meggendorfer
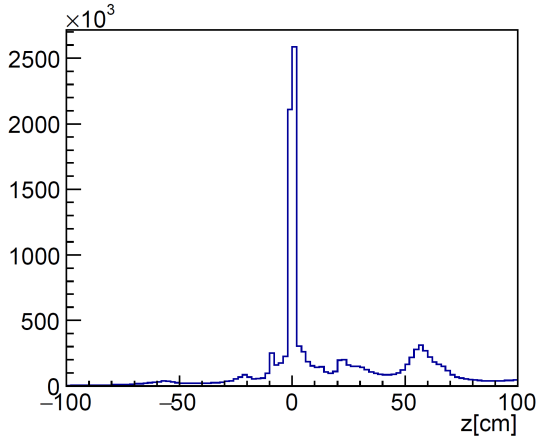
Figure 1: Distribution of the z-positions in centimeters for offline reconstructed tracks over the entire z range. The data is from the running in the year 2020, before the launch of the z-Trigger. Only about 20% of the tracks come from the interaction region.

2 Principle of the Neural z-Trigger

The CDC, with an inner (outer) radius of $r = 16$ (113) cm, is equipped with 56 cylindrical layers of axial and stereo wires. It contains 14336 drift cells with a size of $2\,\mathrm{cm}$. Six adjacent wire layers are combined to form nine so-called super-layers (SL). The innermost SL ha s eight layers with smaller (half-size) drift cells to cope with the increasing background towards smaller radii.

The track finding for the trigger is based on a reduced set of sense wires in each of the nine SLs (5 axial and 4 stereo SLs). So-called track segments (TS) combine wires in five adjacent layers within a SL to form patterns similar to hour-glass shapes (see fig. 2, right side, for the TS in SL0 the wire set is shown on the left side). For a TS to fire, its set of wires must satisfy plausible patterns originating from traversing tracks. The central wire (priority wire) defines the spatial position of the TS. In case the central priority wire is not hit, two so-called 2nd priority wires are defined, which take the role of the priority wire. Apart from the geometrical position of the priority wires, their drift time with a resolution of 2 ns is provided by the CDC frontend electronics.

Introducing the TSs provides a powerful suppression of random noise in the CDC. The set of possible wire patterns within the TSs has been encoded in look-up tables (LUT), which are synthesized statically during design-time. In the conventional track trigger the priority wires in the TSs from the five axial SLs are combined to find 2D-tracks in the plane transverse to the beam direction ($r\phi$ plane). Using the positions of the priority wires of the axial TSs, improved by the drift time, the track finding is done using Hough transform techniques [4, 5]. The Hough method assumes the track origins at $(x = 0, y = 0)$ in the $r\phi$ plane and returns a set of 2D track candidates, which are defined by their azimuthal emission angles ϕ at the origin and their track curvatures $\omega = 1/R$, where R is the radius of the track orbit in the $r\phi$ plane. At least 4 of the 5 axial TSs are required in the Hough transform to establish a 2D track candidate.
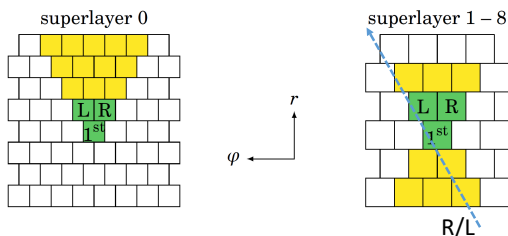

Figure 2: Left: TS of the innermost (axial) SL of the CDC consisting of 15 drift cells, arranged in triangular shape. Right: TS in the other eight SLs consisting of 11 drift cells in hour-glass shape. The drift cells used for the TS are marked in yellow. For each TS the wire marked as 1st is called the priority wire. The two wires labeled R and L are used in case the priority wire did not fire (see text).

For any given 2D track candidate found by the Hough transform, a set of possible stereo TSs associated with the 2D track is selected and the stereo TS with the shortest drift time is chosen for each of the four stereo SLs. The neural nets are then used to yield accurate estimates for the z-impact and the polar emission angle of the track. For each of the up to 9 priority wires three variables are calculated (preprocessing, see fig. 3 and their definition in the next section) which are fed into a single hidden layer

feed-forward neural network. The two outputs of the network are the z position and polar angle θ of a neural track.

The global L1 track trigger at Belle II is operated as follows: Whenever two or more 2D track candidates have been found in an event, at least one neural track with typically $|z| < 15$ cm is required for a valid L1 decision. In addition, using the estimate of the polar angle θ, the momentum of each neural track is calculated. The momentum is used to enable a trigger requiring only one 2D track candidate together with a neural track. This Single Track Trigger (STT) also requires $|z| < 15$ cm, and in addition a momentum $p > 0.7$ GeV. Since only a single neural track is required in an event, the STT opens up the full phase space for the second and further tracks in the event, which either have very low transverse momentum (< 250 MeV), or are emitted with very shallow polar angles.

3 Neural Architecture, Preprocessing, Training

The neural algorithms, yielding for each track in an event an estimate for the z-coordinate and the polar emission angle θ, are executed on Field Programmable Gate Arrays (FPGA) in the CDC track trigger electronic boards. Due to latency limitations at L1, only about 300 ns are available for the execution of the neural algorithm, including input variable preprocessing and delivery of the output to the global decision logic (GDL). Therefore, only single-hidden-layer feed-forward network architectures were considered.

The inputs to the neural networks consist of triplets of variables derived for each priority wire in the associated TS in each of the nine SLs (shown on the left side in fig. 3). The triplet of variables consists of the crossing angle α of the track through the TS, the signed drift time t_{drift} and the relative azimuth angle φ_{rel} of the priority wire. The crossing angle α is the inclination (or zenith) angle of the track passing the TS (see also fig. 2 for illustration). The angle φ_{rel} is given by the difference of the priority wire position in azimuth, ϕ_{pw}, and the value ϕ_{extr} extrapolated from the Hough parameters ϕ and ω of the 2D track, measured at the end plate of the CDC. For the axial TSs this angle is usually close to zero. The right-left (R/L) ambiguities for the drift times are lifted by pre-determined LUTs, analyzing the hit patterns in the TSs from a set of simulated tracks spanning a wide range of momenta and emission angles. Whenever a clear decision can be made for the track passing right (left) from the priority wire, the drift time is assigned to a negative (positive) value (see fig. 2). In the case no decision can be made, the drift time is set to zero (track passing close to the wire), irrespective of the actual value of the drift time.
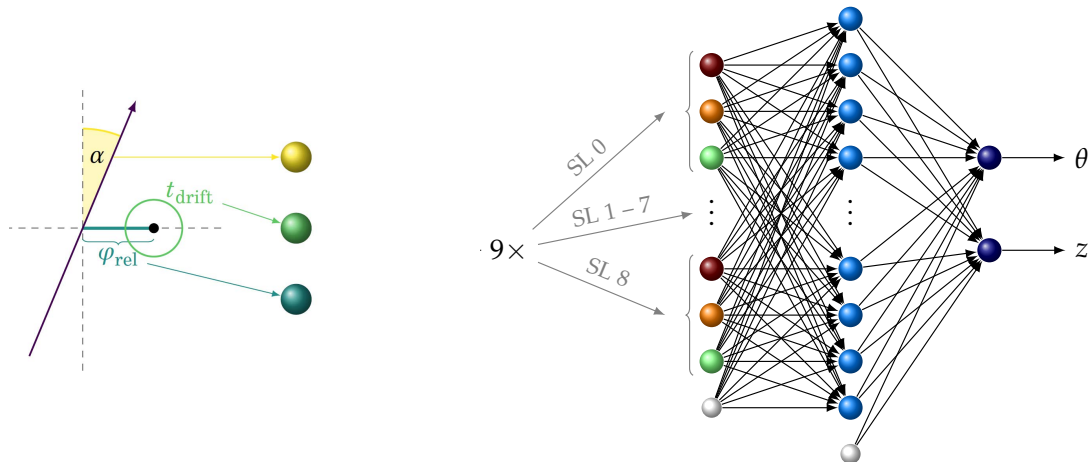

Figure 3: Input variables to the neural network and the neural architecture. A triplet of variables is prepared for the TS in each of the 9 SLs, α, φ_{rel}, and the signed drift time t_{drift} (see text). The network architecture is a fully connected feed-forward neural network with 81 hidden nodes. Two output nodes provide the estimates for the z origin of the track and its polar emission angle θ.

Given a 2D track candidate, the possible stereo SLs are selected using look-up tables (LUT), pre-determined from fully reconstructed tracks prior to the network training. This set of possible stereo TS candidates is defined within a range $\Delta\phi$ around the mean value ϕ_{mean} of the two axial TS sandwiching the stereo SL. From this set the stereo TS with the shortest drift time is chosen, yielding the value $\varphi_{\mathrm{rel}} = \phi_{\mathrm{pw}} - \phi_{\mathrm{mean}}$ for the priority wire in the TS, with ϕ_{pw} again measured at the end plate of the CDC.

The training of the networks is done with the target values z_{reco} and θ_{reco} taken from the fully reconstructed tracks (reco tracks) in event samples recorded during the data taking. For the training of the neural trigger the PyTorch library [6] is used. Here, we also include a second norm weight regularization into the loss function to punish excessively large weights and thus avoid numerical instabliites when going to the integrized version on the FPGA hardware. The loss function is minimized by Adam [7], a first order optimizer, using mini-batches. The gradients are calculated with the backpropagation algorithm implemented in PyTorch. Convergence is controlled by an independent validation data set. The sample sizes for training and convergence tests are generally of order 300 k tracks. The training itself is done in floating point arithmetic. All the inputs, outputs and network parameters are then transformed to integers for the computations on the FPGAs.

4 Performance of the Neural z-Trigger

In late 2020 the z-Trigger was commissioned, being monitored but not yet activated at that time. Due to increased background, coming along with rising instantaneous luminosity, the neural trigger was activated from 2021 onward and operated as a global track trigger at L1. To assert a track trigger, at least one neural track is required in the event obeying the condition $|z| < d$. Typical values of d are 20 cm or less. For the performance studies we compare the neural tracks (neuro track) with the corresponding reco tracks. Examples for the achieved resolutions for the two parameteres z-impact and θ are given in fig. 4, with data from the high luminosity / high background running before the long shutdown in July 2022. The plots show the differences $\Delta z = z_{\text{neuro}} - z_{\text{reco}}$ (z-resolution, left side) and $\Delta p = p_{\text{neuro}} - p_{\text{reco}}$ (momentum resolution, right side). The distributions are well described by double-Gaussians, where the wide Gaussian represents the neuro tracks with one missing stereo TS. The core Gaussian for the neural z estimate yields a resolution of about 3 cm, the width of 60 MeV for the core Gaussian of the momentum distribution mainly describes the resolution below $p \sim 1$ GeV.
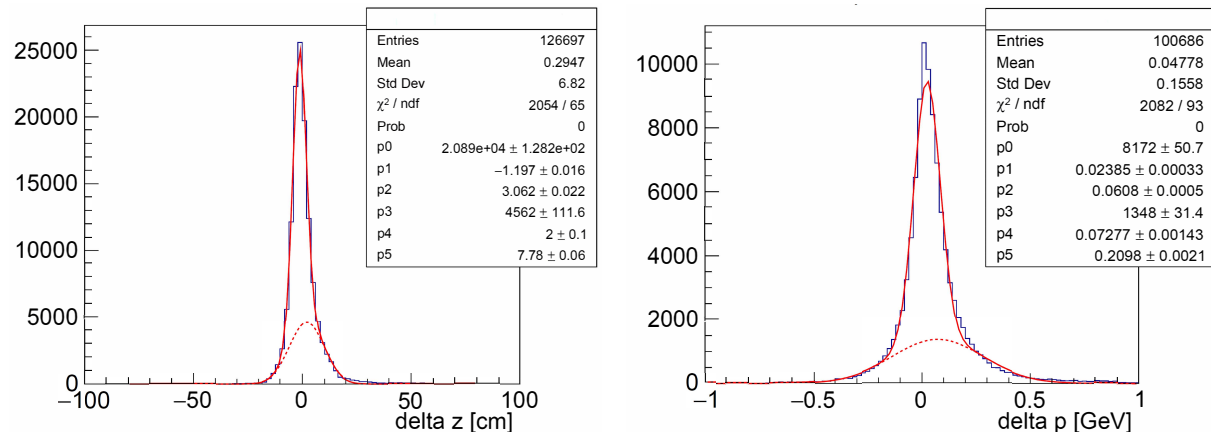

Figure 4: Distributions of $\Delta z = z_{\text{neuro}} - z_{\text{reco}}$ (left side) and $\Delta p = p_{\text{neuro}} - p_{\text{reco}}$ (right side). Fits to the distributons with double-Gaussians are also shown.

Single Track Trigger (STT): In addition to the requirement of a neuro track for all standard $\geq$ 2-track triggers, a minimum- bias single track trigger was launched, requiring $|z| < 15$ cm, and supplemented by an additional minimal requirement for the track momentum of $p > 0.7$ GeV. The momentum cut largely reduces the huge QED background $e^+ e^- \rightarrow e^+ e^- e^+ e^-$, but also strongly reduces the off-IP background generated by spallation protons, with typical momenta around 500 MeV.

At the trigger level, the momentum of the neuro track is calculated from the curvature ω of the input 2D track, the known solenoid field B of Belle II, and the polar scattering angle θ, estimated by the network (second output of the networks, see fig. 3):

$$p[\text{GeV}] = \frac{1}{\omega[1/\text{m}] \ \sin(\theta) \ 0.3 \ B[\text{T}]}.$$

Based on this expression the momentum of a neuro track is determined in the GDL using a look-up table for the trigonometric function $\sin(\theta)$.

The STT was launched as an unprescaled physics trigger for low multiplicity events during the year 2021. Its contribution to the total trigger rate was observed at the level of roughly 20%, which is very reasonable for a minimum bias trigger. Since only a single track is required, further tracks have no limitation set by the acceptance of the CDC. Therefore also 2-track events are triggered where the second track is only going through a few planes of the CDC and the offline reconstruction is mainly done by the vertex detectors. A comparison of the STT relative to the two main multi-track triggers (still requiring a neural track, see above) is shown in fig. 5 for the reaction $e^+e^- \rightarrow \mu^+\mu^-(\gamma)$. The plot shows the event efficiency as a function of the track with the smaller transverse momentum. The observed improvement of the efficiency for momenta larger than 0.7 GeV is evident.
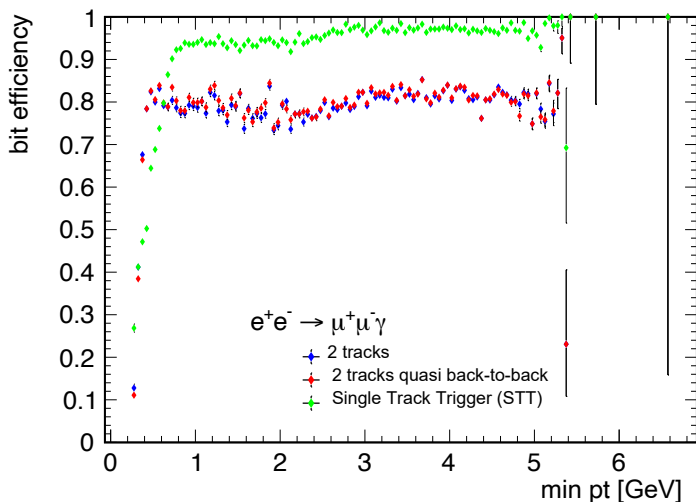

Figure 5: Efficiency of the STT in comparison to the main two-track triggers (requiring at least one neural track) for the reaction $e^+e^- \rightarrow \mu^+\mu^-(\gamma)$, as function of the transverse momentum of the track with the smaller p_T.

Towards the end of the data taking period, before the Long Shutdown LS1 in June 2022, the machine conditions were characterized by extremely large backgrounds in the CDC, leading to increased rates for the z-Trigger, dominated by the STT: Instead of the usual 20% share in the total trigger budget, this share has increased to about 50%. On the other hand, the efficiencies for vertex tracks did not change within their uncertainties. The main reason is an increase in the number of fake neuro tracks with z values inside the acceptance band, which do, however, not correspond to any real reco track. Fake tracks are typically generated by an increasing rate of 2D input track candidates, which are formed by or contaminated with random background hits.

5   Ongoing Developments

We envisage several ways to stabilize the STT and the multi-track z-Trigger for future running. Since new and more powerful custom-made trigger boards (UT4, equipped with Virtex UltraScale XCVU080/160 FPGAs) have become available recently for the z-Trigger, more resources are now available to overcome the limitations of the present single hidden-layer networks. This means that the neural network architecture of the z-Trigger can now be extended to deep-learning networks, with typically three to four hidden layers with $\mathcal{O}(100)$ nodes each.

Most importantly, the input track candidates for the networks need to be made robust against background. Presently, the track finding is carried out in 2D only. To reduce the chance of track candidates formed by background TSs, we have developed a scheme [8], where the track candidates are searched in a three-dimensional Hough space, adding as a third dimension the polar track angle θ. The enlarged Hough space has the advantage that now all 9 SLs (instead of only the 5 axial SLs at present) can be used for track finding, making random noise much less likely. As a further advantage of the 3D Hough space the track candidates are forced to come from IP, thus naturally suppressing tracks from outside.

There are also studies going on to implement a trigger to identify track pairs with a strongly displaced vertex (Displaced Vertex Trigger DVT). Such event signatures are expected, for example, in models that

predict inelastic dark matter production in association with a dark Higgs boson [9], but cannot be efficiently triggered by the STT. Since the origin for the displaced vertices is unknown, the method of choice is to put a grid of possible vertex locations (we call them macro cells) over the transverse plane of the CDC. A typical number of the finite-size macro cells is of order 100 and the track finding via Hough transforms has to be executed in parallel with each of these macro cells as vertex hypothesis. In the new UT4, this seems possible within the latency attributed to the L1 track trigger. More details on the neuro trigger and its upgrade program can be found elsewhere [10].

6 Summary and Outlook

For the Belle II experiment at SuperKEKB the first fully operating global L1 track trigger based on neural networks has been realized. The neural trigger uses the information from the axial and stereo wires of the Central Drift Chamber (CDC). The input to the networks are the track candidates provided by the conventional 2D track trigger using Hough transformations. Adding also the stereo wires of the CDC as input, a set of single hidden layer networks provide as outputs the vertices of the 2D track candidates along the beam (z)-direction and their polar emission angles θ. With this information the neural trigger rejects within 300 ns the events coming from outside of the electron-positron interaction point ($z = 0$): All track triggers with two or more 2D candidates require at least one neural track with the condition $|z| < 15$ cm. Using the polar angle θ, a minimum-bias single track trigger (STT), requiring a mild momentum cut of 0.7 GeV in addition to the acceptance cut of $|z| < 15$ cm, became possible. The STT is particularly effective in selecting events with low charged multiplicity.

While the physics performance of the STT was not affected by increasing backgrounds, it showed an over-proportional increase in rate, generated mainly by the background-prone 2D track finding algorithm. In order to provide stable operation for the STT in the future, a comprehensive upgrade program is in progress, replacing the traditional 2D track finding by a novel three-dimensional Hough space analysis, and extending the network structure to deep-learning architectures with additional wire inputs. The expected increase of performance of this new scheme is encouraging.

Since the neural triggers are optimized for tracks coming from the interaction point, they will miss event signatures with tracks coming from a strongly displaced vertex. Such event signatures are expected in various extensions of the standard model that predict long-lived particles. We shortly sketched our plans to add neural algorithms to the Belle II L1 track trigger system for such anomalous event types.

References

[1] S. Hashimoto et al., *LoI for KEK Super B Factory, Part I: Machine*, KEK-Report 2004-4 (2004).

[2] KEKB B-Factory Design Report, KEK Report 95-7, August 1995.

[3] T. Abe et al., *Belle II Technical Design Report*, KEK-REPORT-2010-1, arXiv:1011.0352v1 [physics.ins-det] (2010).

[4] Hough, P.V.C. *Method and means for recognizing complex patterns*, U.S. Patent 3,069,654, Dec. 18, 1962; Duda, R.O.; Hart, P. E. (January 1972). *Use of the Hough Transformation to Detect Lines and Curves in Pictures*. Comm. ACM. 15: 11–15, Jan. 1972.

[5] S. Pohl, *Track Reconstruction at the First Level Trigger of the Belle II Experiment*, PhD Thesis, Ludwig-Maximilians-Universität München, 2018.

[6] A. Paszke, S. Gross et al., *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, Advances in Neural Information Processing Systems 32 (2019).

[7] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, 3rd International Conference for Learning Representations, San Diego (2015)

[8] S. Skambraks, *Efficient Physics Signal Selectors for the First Trigger Level of the Belle II Experiment based on Machine Learning*, PhD Thesis, Ludwig-Maximilians-Universität München, 2020.

[9] See e.g.: David Smith and Neal Weiner. *Inelastic Dark Matter*. Physical Review D, 64(4), 2001; Duerr, M., Ferber, T., Garcia-Cely, C. et al. *Long-lived dark Higgs and inelastic dark matter at Belle II*. J. High Energ. Phys. 2021, 146 (2021). https://doi.org/10.1007/JHEP04(2021)146.

[10] S. Bähr et al., *The Neural Network First Level Hardware Track Trigger of the Belle II Experiment*, hep-ex arXiv:2402.14962.