PROCEEDINGS OF SCIENCE

Online Estimation of Particle Track Parameters based on Neural Networks for the Belle II Trigger System

Steffen Baehr* **and Kai Lukas Unger and Juergen Becker**
Karlsruhe Institute of Technology (KIT)
E-mail: steffen.baehr@kit.edu, kai.unger@kit.edu, juergen.becker@kit.edu

Felix Meggendorfer and Sebastian Skambraks and Christian Meggendorfer
Max-Planck-Institute for Physics, Munich, Germany

The Belle II particle accelerator experiment is experiencing substantial background from outside of the interaction point. To avoid taking data representing this background, track parameters are estimated within the pipelined and dead time-free level 1 trigger system of the experiment and used to suppress such events. The estimation of a particle track's origin with respect to the z-Axis, which is along the beamline, is performed by the neural z-Vertex trigger. This system is estimating the origin or z-Vertex using a trained multilayer perceptron, leveraging the advantages of training to current circumstances of operation. In order fulfil the requirements set by the overall trigger system it has to provide the estimation within an overall latency of 5 μs while matching a refresh rate of up to 31.75 MHz for new track estimations. The focus of this contribution is this system' current status. For this both implementation and integration into the level 1 trigger will be presented, supported by first data taken during operation as well as figures of merit such as latency and resource consumption. In addition its upgrade plan for the near future will be presented. The center of these is a Hough based track finding approach that uses Bayes theorem for training the weighting of track candidates. Characteristics of this system's current prototypical implementation on FPGAs as well as present plants towards integration for future operation will be presented.

Artificial Intelligence for Science, Industry and Society, AISIS2019
October 21-25, 2019
Universidad Nacional Autónoma de México, Mexico City, México

*Speaker.

PoS(AISIS2019)010

1. Introduction

The Belle II is an asymmetric electron-positron particle collider experiment located at the KEK facility in Tsukuba, Japan. To enable new discoveries beyond the standard model, it is aiming at achieving a 40-fold increase of its predecessor's luminosity which would break the current world record. While an increase in luminosity is advancing the chances of new discoveries, it is requiring the development of new concepts for a data readout system capable of handling the subsequent increase in data rates. A system capable of transmitting the entire amount of data to be produced, is meanwhile very expensive to be realised. As such, they are accompanied by online data processing that reduces the data rate to be supported and thus the cost for implementing the data readout. The most efficient reduction is hereby achieved by performing this reduction as early as possible and thus as close as possible to the sensors. This typically requires dedicated processing platforms that are capable of being interfaced with the respective sensors. One popular platforms used across several experiments are FPGAs due to their high and diverse IO capabilities. The mechanisms tasked with the reduction can be meanwhile separated into online data reduction [2] and trigger systems. Both approaches are based on the fact that the origin of most observed particle tracks is not related to the collisions but are rather a result of background effects. Efficient and lossless reduction is then achieved by distinguishing data belonging to background from physics events. A trigger system is then using this classification in order to decide online during the runtime of the experiment when to send the detector data. As such, they are managing the readout frequency, that is typically constrained by the data acquisition system of the experiment.

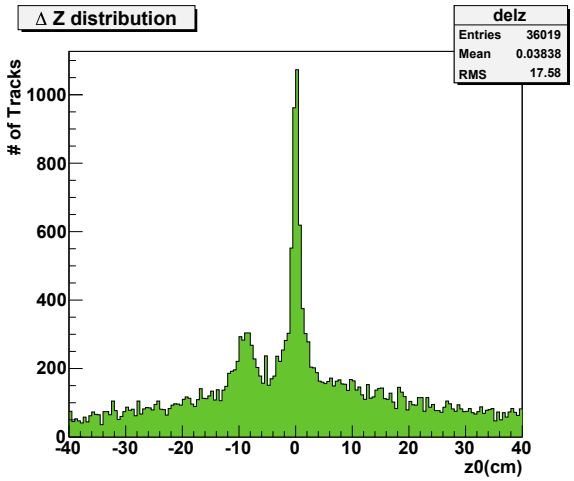

Figure 1: z-Distribution from the Belle experiment [1].

One of the possibilities for identifying background events is to a particle track's point of origin. Only those tracks that are originating from the interaction point are part of the physics experiment. Thus, by accurately estimating the point of origin, background can be identified and suppressed. The effectiveness hereby depends on the amount of tracks originating from outside the interaction point. As most of the fundamental detector design is based on the predecessor Belle, it is a good starting point for investigation. It was observed that a high amount of data that was read out was due to particles outside of z = 0, which is the interaction point along the z-Axis. The distribution of

events and their reconstructed origin is shown in figure1. A significant part of the read out particle tracks did as a result not contribute to new physics. This situation is expected to worsen for Belle II, as a result additional effort is put into a trigger based on determining the z-Vertex. One of the mechanisms to be used in the experiment is a neural z-Vertex trigger that is capable estimating the parameter by using neural networks. It is an FPGA-based trigger system that uses a multilayer perceptron that was trained to estimate the z-Vertex. To stay within the latency budget of the entire trigger system and match the throughput requirements, the estimation has to be generated within the overall latency of 5 μ,s while achieving a throughput frequency of 31.75 MHz [3].

This paper is organized in the following way. Section 2 provides the fundamentals for L1 trigger system based on the central drift chamber of the experiment, which forms the basis of operation of the neural trigger. In Section 3 related track trigger concepts and design flows are discussed. The neural trigger is presented in section 4 together with first results from the experiment. Section 5 concludes the paper with a discussion and outlook

2. The Belle II CDC Trigger at the First Level

The essential part of the observation of collisions is an exact reconstruction of the subsequent decays. The experiment uses three detectors, which are primarily used for reconstruction of particles tracks. These are the pixeldetector (PXD), silicon vertex detector (SVD) and central drift chamber (CDC). The PXD and the SVD are completely new developments for the Belle II experiment, while the CDC is mostly reused from Belle. The PXD forms the innermost detector of the experiment and is placed directly around the IP. It provides a much higher precision in spatial measurements, but has a relatively long integration time due to the used DEPFET sensor technology. The SVD is enclosing the PXD and is in contrast it is a strip detector with less accuracy in spatial resolution but a shorter readout time and higher coverage of space. Both are enclosed by the CDC that is mainly used to determine a track's momentum and charge by estimating the track parameters. These three detectors are used to determine decay nodes and find tracks with low momentum. The tracking detectors are enclosed by particle identification detectors. The time of propagation detector is located directly after the CDC at the barrel while the particle identification detector is installed at the endcaps. The CDC[4] is hereby the most important for this paper. It is constructed according to a layered model. Each layer consists of a number of wires stretched parallel to the z-Axis and surrounded by a gas mixture. This gas mixture consists of a so-called low-Z gas. Interactions of the passing particles with the gas mixture is producing charge carriers at the wires. These are then used to detect the passing of a particle. The wires are grouped in layers, in total 56 layers are forming the entire detector. Successive layers are arranged with an increasing distance to the IP. Additionally, the number of wires varies across layers. Layers that are placed further away contain more wires, to cover the correspondingly larger space to be covered for detection of particles. The layers are grouped together into nine SuperLayer (SL) that are enumerated from zero to eight. All wires have properties specific to their SL. They have an orientation, in which there are either aligned parallel to the z-axis, called axial. Or they have a stereo orientation in which they are aligned with an angle between 45.4 to 74.0 mrad to the z axis depending on the SL. This angle is introduced to enable three-dimensional reconstruction of tracks.

The trigger system of the CDC is the most important sub-trigger of Belle II. It has the longest processing pipeline and therefore requires the highest overall latency to generate the necessary signals. For the entire processing from frontend to global decision logic (GDL) including all communication 5μs is fixed, while it has to generate trigger signals with a frequency of 30 kHz. The NNT developed in this paper is one of its sub-systems. To understand the background and resulting requirements, the system architecture is outlined.

Overall a multi-level and pipelined processing architecture is used. Data of individual SLs is continuously read out by the FrontEnd. It is then transferred to merger units, which concentrate data of the entire CDC. The concentrated data is then sent to the first stage of the trigger logic, the track segment finder (TSF) The task of the TSF is to combine the individual wires into so called track segments (TS). These are generated for each SL separately in parallel by dedicated FPGAs. TS are additionally distinguished into axials and stereo according to the orientation of the wires within the processed SL. Axials are first processed by the 2D track finder, which tries to find a suitable 2D-tracks. At the same time using the CDC an event time by an event time finder. The outputs from these two modules together with the stereo TS are then sent to the 3D-Track Finder and the NNT. Both of these have the responsibility to send an estimation of the z-Vertex to the GDL.

3. Comparable Trigger Approaches

With the usage of neural networks, the relationship based on data taken from the experiment or simulation can be learned without the necessity to create a precise algorithm. Another popular approach is the usage of a Hough transformation or linear regression for track finding. These approaches extrapolate tracks by using single detector hits and fitting a track around them. They are viable options for readout triggering, achieving accurate results in the past. However, since the background is at the moment not completely understood, simulations showed that neural networks are projected to be more robust for future deployment. Additionally, adjusting these algorithms to the changing behaviour of the experiment is not an easy task, since they have to be redesigned. Neural networks on the other hand can be simply retrained using newly acquired data. Since neural networks offer the described advantages compared to traditional approaches, they were already used in past experiments for example the H1-Level 2 Trigger of the HERA experiment [5] However, that trigger was implemented on dedicated processors and had a rather large time budget of 200μs compared to the requirements of the Belle II L1 trigger at 5μs. Employing neural networks on FPGAs gained significant traction over the last years. FINN [6] showed how an abstract network description can be transferred efficiently onto FPGAs. It is nowadays used as reference for further improvements. However, these frameworks are not meant to be used for high energy physics use cases as they dont provide the necessary interfaces are not targeted at their specific low latency requirements. They rather focus on predefined architecture that cover a broad range of applications. The lack of a suitable design flow for this domain was recognized with HLS4ML which is a framework that can generate a neural network architecture [7] based on a physics use case. However, it is more focused on the investigation of feasibility of neural networks for trigger systems in general, while the presented work represents a complete system that is currently in use and integrated into the data flow of the experiment.

4. Realisation and Evaluation of the Neural z-Vertex Trigger

Algorithmically the employed approach for estimating the z-Vertex consists of two main stages [8, 9]. At first a preprocessing stage transforms the detectors data into a more viable representation. Using this the neural network is estimating of the z-Vertex. Often the preprocessing stage can be omitted and compensated by using a more complex network. Since both the latency and resource budget is tightly constrained, a preprocessing stage that makes use of the geometrical information of the CDC is used reducing the number of necessary networks and neurons. Reasonable accuracy for the estimation can already be achieved using just one network. However, the final implementation uses a total of five networks that can compensate missing wires in the stereo SLs. In case one of the four stereo layers does not have a matching TS, a specialized network is loaded during runtime

The preprocessing calculates three separate input variables for each of the SLs of the CDC. This results in 27 inputs in total for the network. The three variables are the crossing angle α, which is between the estimated 2D particle track and an active reference wire of the CDC, the wires position relative to the estimated 2D track ϕrel and the drift time that describes the time between the particle passing through the space and a hit being registered at a wire. The geometrical representation of the inputs is shown in figure 4. For the network itself an MLP with one input, hidden and output layer is used. Each node in the hidden and output layer represents a neuron that computes a weighted sum over its input values. From a processing point of view, the MLP mainly consists of a set of multiply and accumulate operations (MAC). For each neuron's output a predefined activation function is applied, for which we use the hyperbolic tangent. In our setup all input and output values of the MLP are scaled to values within the range $[-1,1]$. Training of the MLP showed that using 81 neurons for the hidden layer and two neurons for the output layer is sufficient to achieve the required results for estimating the z-vertex. Our network training is performed using the iRPROP algorithm [10] which is an improved variant of the RPROP backpropagation algorithm. This algorithm has been demonstrated to be more effective than the classical backpropagation, because the magnitude of the weight update is independent of the magnitude of the derivative of the cost functions and only dependent on the dynamics of the past weight updates. The effect is a faster convergence to a minimum of the cost function with the same minima found compared to the classical backpropagation.

Results from trigger operation taken during early runs of the experiment are shown in figure 2. The first plot is representing the difference in the z-Vertex estimated by the hardware and reconstructed by later precise analysis. A single peak at 0 would be the optimal results, however we are seeing that the hardware estimation is not already exact as the distribution has some broadness. However when applying a z-Cut at $-40:+40cm$ still a significant reduction of background events can be achieved. Accompanying this, the second plot is showing a scatter plot comparing the results of the hardware and a software emulation that was performed with the same input data that was used by the hardware. An linear correlation an be seen here across the entire z-Axis, which shows the general correctness of the hardware trigger. However, there are still some outliers which are mostly due to mismatches in the offline analysis. These outliers will be addressed by improvement of the analysis.

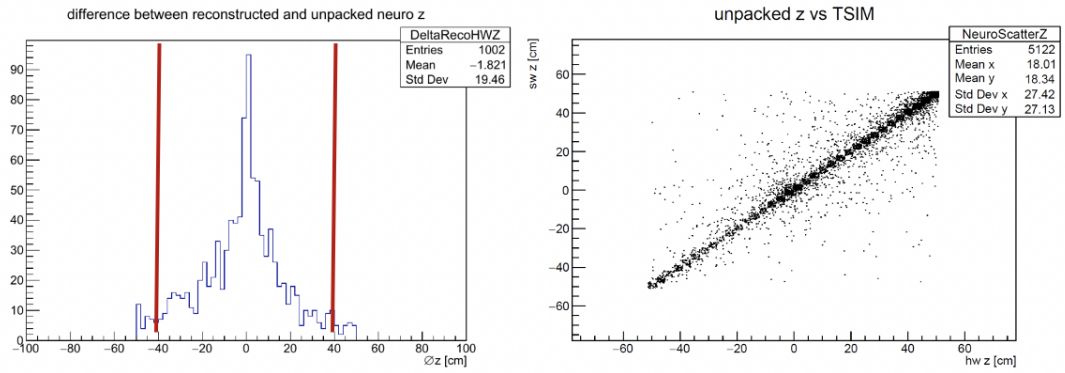

Figure 2: Estimated z-Vertex and scatter plot between HW results and SW emulation of the algorithm.

5. Conclusion

This paper provides an overview over the neural z-Vertex trigger that is used in the Belle II experiment. It is used to suppress background event identifiable by particle tracks with an z-Vertex outside of the interaction point. As this reduction has to be performed as early as possible to achieve efficient suppression of the overall data rate, it is located at the L1 trigger system for the central drift chamber. To fulfil its hard real-time requirements the trigger system is implemented on FPGAs with deterministic design of the processing architecture. The system is meanwhile already operational and early results show promising performance with a reasonable resolution of about 40 cm, which can be used for suppression at this stage of the experiment.

References

[1] T. Abe et al., "Belle II Technical Design Report" , ArXiv Nov. 2010

[2] S. Baehr et al., "Online-analysis of hits in the Belle-II pixeldetector for separation of slow pions from background", Journal of Physics: Conference Series, 2015

[3] Y. Iwasaki et al., "Level 1 trigger system for the Belle II experiment", IEEE Trans. Nucl. Sci. 58 (2011) 1807

[4] N, Taniguchi et al., "Central Drift Chamber for Belle-II". Journal of Instrumentation, 12(06):C06014, 2017.

[5] A. Gruber et al, "A neural network architecture for the second level trigger in the H1-experiment at the electron proton collider HERA", arXiv:1406.3319 [physics.ins-det]

[6] Y. Umuroglu et al. "FINN: A Framework for fast scalable binarized Neural Network inference", arXiv:1612.07119

[7] J. Duarte et al., Fast inference of deep neural networks in FPGAs for particle physics, arXiv:1804.06913

[8] S. Skambraks et al., A z -Vertex Trigger for Belle II, [arXiv:1406.3319]

[9] S. Pohl, "Track vertex reconstruction with neural networks at the first level trigger of Belle II", Ph.D. thesis Ludwig-Maximilians-University Munich

[10] C. Igel et al., "Improving the Rprop Learning Algorithm", Proceedings of the Second Int. Symposium on Neural Computation, NC 2000