PROCEEDINGS OF SCIENCE

Design and Performance of the Belle II High Level Trigger

M. T. Prim,[a,*] **N. Braun,**[b] **Y. Guan,**[c] **O. Hartbrich,**[d] **R. Itoh,**[e] **T. Konno,**[f] **K. Lautenbach,**[g] **C. Li,**[h] **Z-A. Liu,**[i] **M. Nakao,**[e] **SH. Park,**[j] **S. Reiter,**[g] **B. Sprück,**[k] **S. Y. Suzuki,**[e] **S. Yamada,**[e] **J. Zhao**[i] **and Q. Zhou**[e]

[a] *Universität Bonn - Physikalisches Institut, Bonn, Germany*
[b] *Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany*
[c] *University of Cincinnati, Cincinnati, U.S.A*
[d] *University of Hawaii, Honolulu, U.S.A*
[e] *High Energy Accelerator Research Organization (KEK), Tsukuba, Japan*
[f] *Kitasato University, Sagamihara, Japan*
[g] *Univeristy of Giessen, Giessen, Germany*
[h] *LiaoNing Normal University (LNNU), Dalian, China*
[i] *Insutitute of High Energy Physics (IHEP), Beijing, China*
[j] *Yonsei University, Seoul, Korea*
[k] *University of Mainz, Mainz, Germany*

E-mail: markus.prim@belle2.org

The data acquisition system of the Belle II detector is designed for a sustained first-level trigger rate of up to 30 kHz at the design luminosity of the SuperKEKB collider. The raw data read out from the subdetector frontends is delivered in realtime into an online High Level Trigger (HLT) farm, consisting of up to 20 computing nodes housing around 5000 processing cores, where it is reconstructed by the full Belle II reconstruction software in realtime. Based on this online reconstruction, events are filtered before being stored to disk for later offline processing and analysis. In this manuscript we will present the newly developed data flow throughout the HLT system utilizing the open-source library ZeroMQ, which was first used in the beam run period in fall 2019. Further, we show the performance of the HLT reconstruction and how the current setup scales with the anticipated future data rates when SuperKEKB increases its luminosity to the design value.

40th International Conference on High Energy physics - ICHEP2020
July 28 - August 6, 2020
Prague, Czech Republic (virtual meeting)

*Speaker

PoS(ICHEP2020)769

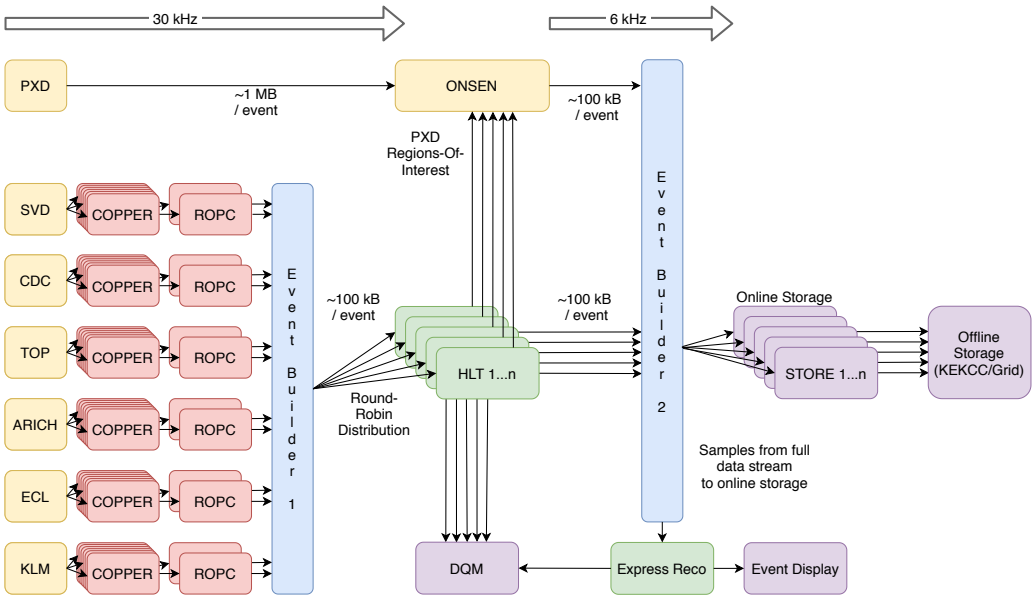

Figure 1: Schematic overview of the data flow in the Belle II data acquisition. The data of the individual detectors, the Silicon Vertex Detector (SVD), Central Drift Chamber (CDC), Time-of-Propagation Detector (TOP), Aerogel Ring Imaging Cherenkov counter (ARICH), Electromagnetic Calorimeter (ECL), and K_L^0 and Muon Chamber (KLM) are sent through the DAQ readout system to the first event builder. The Pixel Detector (PXD) readout is treated in a separate way as indicated in the schematic. For a discussion of the data flow starting from the first event builder see the text.

1. Introduction

The High Level Trigger (HLT) consists of several independent units to leverage the inherent parallelism by the independence of the collision events. The design setup consists of roughly 5000 CPUs distributed up to 20 HLT units, which include 10-20 worker nodes each. Each worker performs event reconstruction and filtering on each of their CPUs. Additionally, each unit has a dedicated input, output, and storage node. With this setup, the HLT is designed to cope with an input rate of up to 30 kHz where each event is of the size of 100 kB excluding the pixel detector (PXD) data (the PXD data flow is treated separately). The HLT input nodes are connected via a fiber network to the detector readout electronics. The event builder processes running on these machines aggregate the raw data without PXD data for the same collision event from different sub-systems, which is streamed round-robin to the individual HLT nodes via TCP. Load-balancing is performed within specific HLT units. The processed data is streamed in form of ROOT objects to a second event builder and in form of binary data to the pixel readout system. The readout of the pixel readout system is based on the binary data and streamed to the second event builder, where the raw data and PXD data are combined and streamed to the storage system. The integration of the HLT into the overall data acquisition data flow at Belle II is schematically shown in Figure 1.

The requirements which are imposed with or into the above described data flow are summarized in the following. The HLT has to reconstruct the full event excluding the hits from the PXD and perform a trigger decision to either keep or discard the event. The event reconstruction is done using the same reconstruction software which is also used in the offline data processing. This means

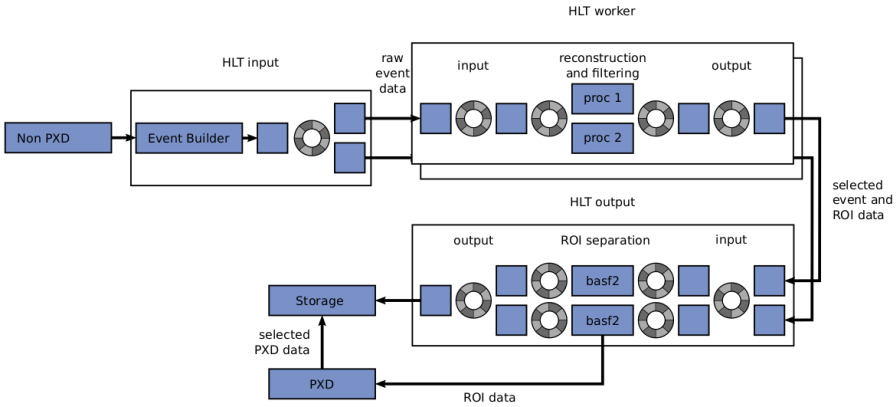

Figure 2: The data flow through the different processes (blue boxes) and nodes (black frames) utilizing the custom shared memory ring-buffers (grey circles). This sketch does only show a single worker with two processes and omits other parts such as the data quality monitoring for better visibility. The abbreviation ROI stands for regions-of-interest. The ROIs are required for the PXD readout. Taken from [2].

that there is no dedicated "fast" reconstruction running on HLT, to avoid additional systematic uncertainties by a different reconstruction for the HLT trigger decision. The average time per event reconstruction is $0.3-0.5$ s. Additionally to the trigger decision, so-called regions-of-interest (ROI) for the PXD readout have to be calculated and send to the PXD readout system. This ROI has to be provided to the PXD readout buffer within 5 s. In case of discarded events the metadata has still to be kept in the data-flow to be able to clear the buffers in the PXD readout system. The HLT farm, at its design hardware specification, has to sustain an input rate of 20 kHz, with peaks up to 30 kHz. To ensure data quality during beam time, the HLT also performs live data quality monitoring (DQM) in form of histograms of pre-defined variables of interest and an event display.

In Section 2 we briefly describe the old and new data flow implementations and follow with an overview of the software operated on the HLT in Section 3. In Section 4 we summarize the current status of the HLT. A comprehensive overview of the Belle II online software can be found in [2], from which this manuscript borrows heavily.

2. HLT Data Flow

The core component of the initial HLT implementation is a custom shared memory ring-buffer, which is read/write accessible by multiple processes. The consistency of the ring-buffer is handled by System V semaphores, which serialize parallel access. Ring-buffers are used between two subsequent stages in the data flow to interchange data on a single node. Communication between different nodes is handled via a custom TCP protocol. Figure 2 sketches the data flow through the HLT utilizing the ring-buffers.

The system based on the ring-buffer setup was successfully deployed during the commissioning of Belle II until 2019. However, some technical limitations which could not easily be solved in this setup became apparent: First, the semaphores require a persistent state. This is problematic when the data processing terminates in an unforeseen way and a residual, possibly invalid, state may be

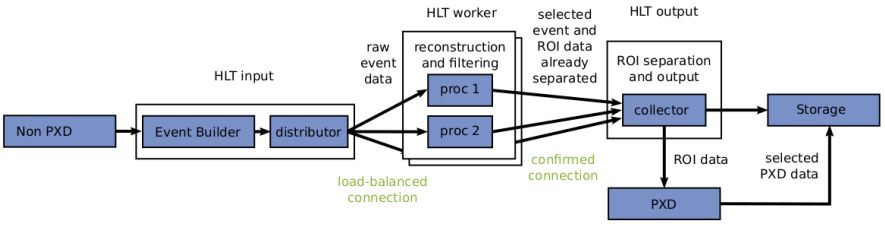

Figure 3: The data flow through the different processes (blue boxes) and nodes (black frames) utilizing ZeroMQ. This sketch does only show a single worker with two processes and omits other parts such as the data quality monitoring for better visibility. The abbreviation ROI stands for regions-of-interest. The ROIs are required for the pixel readout. Taken from [2].

left-over. Cleaning up these residual states is a complex and error-prone procedure and depends on human intervention. Second, non-data communication, which would allow for control commands is not possible. This would allow initializing the processing pipeline before the first data is incoming. Among other considerations, this lead to the modernization of the software for the HLT.

ZeroMQ [1] is a well maintained, open-source library, which has established itself as an industry-standard for high-performance broker-less asynchronous messaging in distributed applications. The inter- and intra-node communication is now based on ZeroMQ TCP connections and the incoming and outgoing messages are buffered via a TCP message queue. This approach removes the necessity of ring buffers and thus removes the aforementioned problem of the residual states. ZeroMQ allows for a variety of message formats, e.g. data- or flow-controlling messages, and the implementation of advanced features such as back-pressure or fair-queuing. The initialization and cleanup of the inter-process and inter-node connections are done automatically by ZeroMQ. Figure 3 sketches the data flow through the HLT utilizing the ZeroMQ library.

3. HLT Processing

The HLT runs the Belle II software framework `basf2` [3, 4] to perform a full reconstruction of the recorded events. Utilizing the offline reconstruction software, additional systematics which could be introduced by a simplified reconstruction are avoided. Based on this reconstruction the event is classified within all available trigger lines. If the event is *not* tagged by any trigger line, the data is discarded and only the event meta data is streamed out. If the event is tagged by any trigger line, and is not discarded by potential prescales, the full event is streamed out with the calculated ROIs, and assembled in the second event builder (see Figure 1). The HLT aggregates variables of interest for events which pass any of the trigger lines and provides this information centralized in the DQM. As PXD information is only available at the second event builder, additional DQM information including the PXD data is calculated on the *Express Reco*, where the data stream through the second event builder is sampled.

In addition to the above described *filter* operation mode, the HLT can also operate in a *monitoring mode*. In this mode the filter decision is still calculated and events are tagged, but all events are kept. This allows for direct measurements of the HLT trigger efficiencies. At the current input rate of $4 - 5$ kHz the HLT is able to run in this mode.

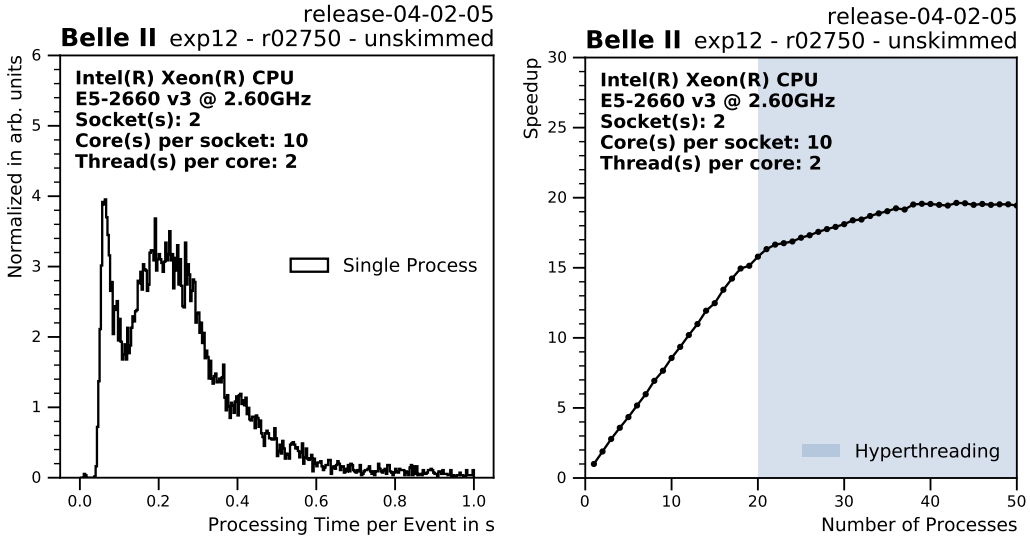

Figure 4: (left) Processing time per event for data recorded with the Belle II detector. The distribution represents a mixture of different event types, e.g. Bhabha and other low multiplicity events. (right) Speedup of the `basf2` reconstruction software operated on the HLT. It shows good scaling behavior with the available physics cores and hyper-threads.

The current setup can cope with the currently delivered luminosity of the SuperKEKB accelerator and Level 1 trigger settings. The average processing time for the current mixture of events defined by the Level 1 trigger is shown in Figure 4 (left). Due to the highly parallel environment of the HLT, `basf2` has to demonstrate that it scales well with the hardware. This is illustrated in Figure 4 (right.)

Considering the anticipated sustained input rate of 20 kHz and scaling with the scheduled hardware additions to the HLT farm (Figure 5), the system is not able to cope with that rate yet. However, promising developments are ongoing in terms of general optimizations of the software, and in particular optimizations in the track reconstruction, which requires the most time in the full reconstruction chain. Additionally, there are trivial optimizations possible which were not considered yet, e.g. removing reconstruction steps which are not necessary for the filter decision. The trivial optimizations can already provide a performance increase by up to 10%.

4. Summary

The new data transportation scheme based on ZeroMQ was successfully deployed in the fall run 2019 and tests have shown that it can handle rates up to 10 kHz per HLT unit, which is beyond any of the necessary requirements. Its new monitoring features have already been proven useful for debugging. Further, switching from raw TCP sockets and custom ring buffers to ZeroMQ has allowed for rapid development and bug fixes.

The `basf2` software running on the HLT shows good scaling behavior, and although the requirements are not yet fulfilled, the estimates on future optimizations are promising so that we can keep the online reconstruction identical to the offline reconstruction to avoid additional HLT trigger systematics.

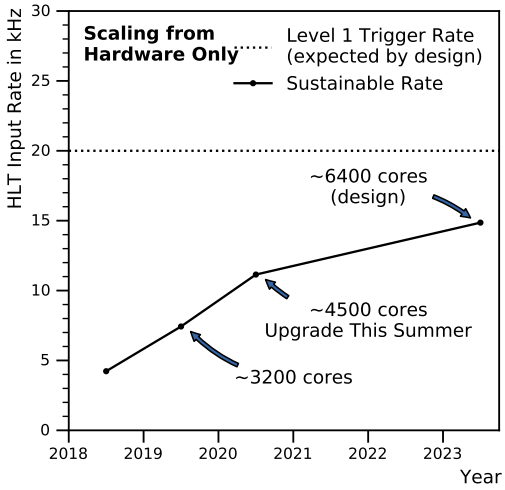

Figure 5: Scaling of the sustainable input rate with current Level 1 trigger prescale factors, reconstruction software, and background levels in the coming years.

References

[1] https://zeromq.org/

[2] Braun, N., Kuhr, T., *Software for Online Reconstruction and Filtering at the Belle II Experiment*, arXiv:2003.02552.

[3] Kuhr, T., Pulvermacher, C., Ritter, M., Hauth, T., Braun, N., *The Belle II Core Software*, arXiv:1809.04299, 10.1007/s41781-018-0017-9, Comput.Softw.Big Sci. 3 (2019) 1, 1.

[4] Bertacchi, V. et. al., *Track Finding at Belle II*, arXiv:2003.12466, 10.1016/j.cpc.2020.107610, Comp.Phys.Comm 259 (2021)107610.