Prepared for submission to JINST

A demonstrator for a real-time AI-FPGA-based triggering system for sPHENIX at RHIC

J. Kvapil,[a,1] **G. Borca-Tasciuc,**[b] **H. Bossi,**[c] **K. Chen,**[d] **Y. Chen,**[d] **Y. Corrales Morales,**[c] **H. Da Costa,**[a] **C. Da Silva,**[a] **C. Dean,**[c] **J. Durham,**[a] **S. Fu,**[e] **C. Hao,**[f] **P. Harris,**[c] **O. Hen,**[c] **H. Jheng,**[c] **Y. Lee,**[c] **P. Li,**[f] **X. Li,**[a] **Y. Lin,**[a] **M. X. Liu,**[a] **A. Olvera,**[e] **M. Purschke,**[g] **M. Rigatti,**[h] **G. Roland,**[c] **J. Schambach,**[i] **Z. Shi,**[a] **N. Tran,**[h] **N. Wuerfel,**[j] **B. Xu,**[f] **D. Yu,**[k] **H. Zhang**[f]

[a] *Los Alamos National Laboratory,*
Bikini Atoll Rd, Los Alamos, NM 87545, United States
[b] *Rensselaer Polytechnic Institute,*
110 8th St, Troy, NY 12180, United States
[c] *Massachusetts Institute of Technology,*
77 Massachusetts Ave, Cambridge, MA 02139, United States
[d] *Central China Normal University,*
No.152, Luoyu Rd, Wuhan 430079, China
[e] *University of North Texas,*
1155 Union Cir, Denton, TX 76205, United States
[f] *Georgia Institute of Technology,*
225 North Ave, Atlanta, GA 30332, United States
[g] *Brookhaven National Laboratory,*
PO Box 5000 Upton, NY 11973, United States
[h] *Fermilab,*
PO Box 500. Batavia IL 60510, United States
[i] *Oak Ridge National Laboratory,*
P.O. Box 2008. Oak Ridge, TN 37831, United States
[j] *University of Michigan,*
500 S State St, Ann Arbor, MI 481091, United States
[k] *New Jersey Institute of Technology,*
323 Dr Martin Luther King Jr Blvd, Newark, NJ 07102, United States

E-mail: jakub.kvapil@lanl.gov; jakub.kvapil@cern.ch

[1] Corresponding author.

Abstract: The RHIC interaction rate at sPHENIX will reach around 3 MHz in pp collisions and requires the detector readout to reject events by a factor of over 200 to fit the DAQ bandwidth of 15kHz. Some critical measurements, such as heavy flavor production in pp collisions, often require the analysis of particles produced at low momentum. This prohibits adopting the traditional approach, where data rates are reduced through triggering on rare high momentum probes. We explore a new approach based on real-time AI technology, adopt an FPGA-based implementation using a custom designed FELIX-712 board with the Xilinx Kintex Ultrascale FPGA, and deploy the system in the detector readout electronics loop for real-time trigger decision.

Keywords: Detector control systems (detector and experiment monitoring and slow-control systems, architecture, hardware, algorithms, databases); Trigger algorithms; Trigger concepts and systems (hardware and software)

ArXiv ePrint: not yet

1 Motivation

Realizing the science potential of modern nuclear physics (NP) experiments at colliders relies on the collection and processing of very large datasets, with sustained data rates from future detectors exceeding Terabits per second. Critical measurements in NP data often require the analysis of complex final states of particles produced at low momentum. This prohibits adopting the traditional approach used in high energy physics, where data rates are reduced through online event selection ("triggering") on rare high momentum probes that can be readily distinguished from the background. On the other hand, archiving the full detector data stream exceeds current DAQ bandwidth limits and would lead to cost-prohibitive offline storage and analysis computing requirements. We propose to develop real-time AI technologies, implemented in the detector readout electronics loop, that address these challenges for the next generation of NP experiments at RHIC and EIC. First, we will deploy a demonstrator that is being developed under the current Fast-ML project for the pp running in the sPHENIX experiment in 2024, and then generalize our approach for applications in experiments at the EIC, using future generations of EIC detector technologies.

The triggered readout rate of sPHENIX is limited to 15 kHz due to the design of the calorimeter readout system, which places limitations on the overall data volume and the expected collision rate. In pp collisions, RHIC delivers collision rates of 3 MHz, limiting sPHENIX to collect less than 1% of heavy-flavour (HF) events of the total pp (and p+Au) rate when using triggered readout. The extended Streaming readout (SRO) of the tracking detectors can further improve the statistics up to 10% of the total luminosity. The goal of this project is to sample the remaining luminosity further enhancing the collected data samples. The aim is to deploy a future system on the Electron-Ion Collider (EIC) to identify the (non)interesting Deep-Inelastic-Scattering processed in the e+p/A collisions.

2 sPHENIX detector

The sPHENIX detector [1] is located at the RHIC accelerator complex in BNL. The solenoid magnet provides a magnetic field of 1.4 T and the detectors have a pseudorapidity coverage of $|\eta| < 1.1$. The sPHENIX running period is 2023 - 2025, where 2023 was dedicated to commissioning, 2024 to pp collisions, and 2025 to Au+Au collisions. The main central barrel tracking detectors are the Microvertex Detector (MVTX), Intermediate Tracker (INTT), Time Projection Chamber (TPC), and TPC Outer tracker (TPOT), and the central barrel calorimetry system - Electromagnetic Calorimeter (EMCAL) and Hadronic Calorimeter (HCAL). The sPHENIX detector schematic is shown in figure 1 (left). The tracking detectors are capable of SRO. Even-though they are able to record all data, the data volume of the TPC detector exceeds the capability of the computing center and therefore a down-selection of what to save must be done. In the absence of online reconstruction to identify interesting HF events, a random selection of events based on the minimum bias trigger detector is done. The aim of this project is to reconstruct tracklets from the silicon detectors in order to search for signatures of HF decays based on unique topology, and provide an additional trigger to sPHENIX.

Silicon detectors The three innermost tracking layers (the MVTX detector) are based on Monolithic Active Pixel Sensors (MAPS), the ALPIDE, originally developed for the ALICE experiment.

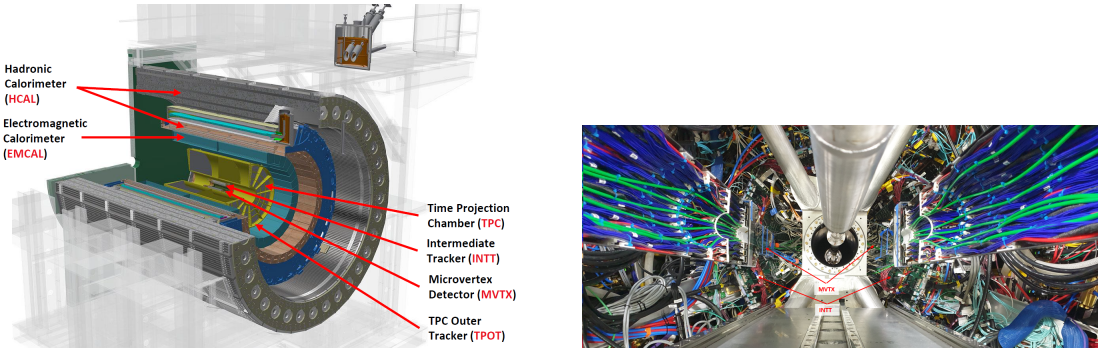

Figure 1. Left: The sPHENIX detector with highlighted central barrel detectors. Right: Installation of MVTX and INTT detectors.

The ALPIDE offers very fine pitch of 27 μm x 29 μm, collision event time resolution of 5 μs; the MVTX contains a total of 270M channels. The next two layers (the INTT detector) contain silicon strip sensors (manufactured by Hamamatsu) with a pitch of 78 μm x 16 (or 20) mm with 360k channels in total. Both detectors are shown in figure 1 (right).

sPHENIX readout, trigger and timing distribution The schematic of the sPHENIX readout chain is shown in figure 2. The tracking detectors' Front-End Electronics (FEE) sends data to the Event Buffer and Data Compressor (EBDC) through the FELIX (FLX-712) [2] interface card. The calorimetry detectors' Front-End Modules (FEM) send data to the SubEvent Buffer (SEB) through the Data Collection Module (DCM2). The trigger and timing information is distributed via the Granule Timing Module (GTM). The Global Level 1 Trigger (GL1) and machine clock are transmitted to the FELIX cards for the tracking detectors and to the FEM for the calorimeters. There can be up to 4 LEMO and 4 Fiber connections to the GTM with 64 trigger inputs total. The most used trigger inputs are from the Hadronic Calorimeters for the cosmic and high energy jet triggers (normally not pre-scaled) and Minimum-Bias Detector for the beam collision trigger (heavily pre-scaled). The goal is to provide an additional high efficiency trigger input for HF events in pp collisions.

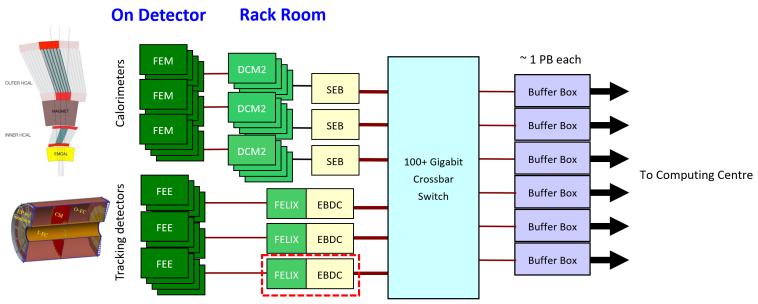

Figure 2. A schematic of sPHENIX readout chain. The calorimetry data are sent from Front-End Modules (FEM) to the SubEvent Buffer (SEB) through the Data Collection Module (DCM2), and the tracking data from Front-End Electronics (FEE) are sent to the Event Buffer and Data Compressor (EBDC) through a FELIX interface card.

2.1 The DAQ-AI Data Flow

Data from the MVTX and INTT are transmitted from the FELIX-DAQ board to the AI-Engine. It was decided that the AI-Engine will be hosted on a widely HEP community supported FELIX board for the following 3 reasons. First, FELIX offers 48 high-speed optical links to receive data; second, to reuse its Wupper module for the PCIe communication and associated software tools; and third, to use the sPHENIX infrastructure for tracking detectors that was built around the FELIX cards. The AI-Engine houses the raw data decoder, event builder, clusterizer, and GNN models to provide fast tracking and the trigger decision. The trigger decision is sent via a LEMO cable to the GTM. Since the decision is based on a event topology, a reference point (beam spot), which changes in time, must be precisely known and monitored. A GPU based feed-back system will be in place to process the data from the buffer boxes, reconstruct the beam spot position and update the position in the AI-Engine. The schematic of the data flow is shown in figure 3.

There are 144 optical links running at 3.2 Gbps per fiber for the MVTX alone, thus two engines will be used, one for each MVTX/INTT hemisphere. This will allow to have 24 links for MVTX and 24 links for INTT. Since the DAQ, AI-Engine, and GTM sit in the Counting house, away from the radiation environment, it is not necessary to use radiation-hardened protocols. The FELIX optical links have been tested up to 14 Gbps with BER $< 10^{-16}$ with an external loop-back measurement. INTT offers excellent time resolution to tag 100 ns RHIC bunch-crossing time to assign a unique timing to each collision event.

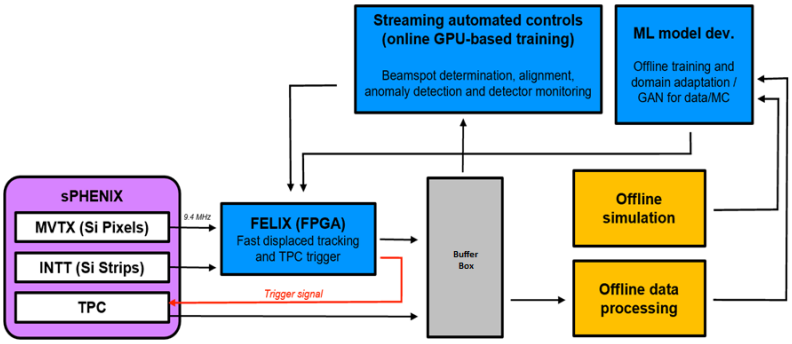

Figure 3. The schematic of the DAQ-AI data flow.

2.2 Latency breakdown

The TPC buffers can hold up to 30 μs of data. As we need to capture both sides around the triggered event, the aim is to deliver the trigger decision with 10 μs. The MVTX contributes up to 5 μs latency, the cables between the interaction region and counting house contribute around 0.3 μs (81m fibres). Forwarding data to the AI-Engine and decoding them takes up to 0.6 μs, which results from the maximum depth of FIFO holding the decoded hits (128). This leaves around 4 μs for the AI engine to perform tracking and trigger decision. The real latency of the decoding and data transmission depends on the occupancy, which is estimated to be around 50 physics hits per chip per event. The AI-Engine needs to ensure the latency is fixed, either by delaying the trigger decision, or vetoing the processing by decoding the bunch-crossing numbers from MVTX and INTT.

Table 1. Efficiency and Background Rejection with 1% and 0.1% signal/noise Ratios.

1% signal/background ratio			0.1% signal/background ratio		
Bg. rejection	Efficiency	Purity	Bg. rejection	Efficiency	Purity
90%	72.5%	7.25%	90%	78%	0.78%
99%	15.0%	15.0%	99%	17%	1.7%

3 Model description

The model is based on Graph Neural Network (GNN) using PyTorch and PyTorch geometric. The aim is to reconstruct the decay topology of the tracklets and search for secondary (displaced) vertices, which is one of the prominent features of HF decays. The displaced vertex will be $\sim$ O(100 μm) away from the primary vertex which has resolution $\sim$ O(10 μm). We propose a novel method for treating events as graphs consisting of tracks as nodes and interconnection between tracks when they belong to the same particle decay rather than hit graphs [3]. Further improvement can be achieved by estimating the transverse momentum, p_{T}, based on the tracks. A 15% improvement is observed in the trigger decision performance, as expected. There are three stages in event processing: hit clustering, track reconstruction, and trigger decision. We use the GNN models in the second stage to reconstruct tracks, remove outliers (TrackGNN), and regress the track momentum onto the learned tracks. In the third stage, we design a bipartite GNN to reconstruct displaced vertices based on the corresponding tracks generated from decays and the estimated track momentum and estimate the probability of event being a trigger. The flowchart is shown in figure 4 and the efficiency and background rejection with 1% and 0.1% signal/noise ratios are summarized in Tab. 1.

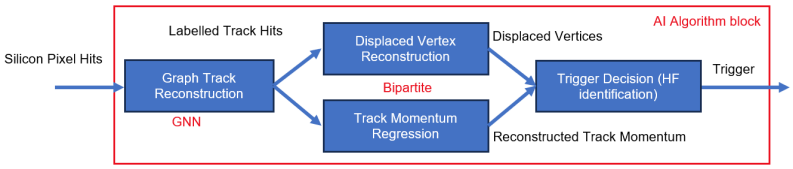

Figure 4. Flowchart of the GNN network.

4 Generation of the GNN IP core

We propose two parallel efforts to implement the software sPHENIX model onto FPGA, both use High-Level Synthesis (HLS) tools to generate synthesizable Register Transfer Level (RTL) code, i.e., Verilog. The first effort, manually translates the sPHENIX model into synthesizable C code and feeds it into the HLS tool, Vitis HLS [5], and then performs targeted optimization of the model following the FlowGNN architecture [6], which is the state-of-the-art GNN architecture on FPGA. The second effort is based on the hsl4ml framework [4], which is a generalized package to translate neural networks, such as deep neural networks (DNNs) and GNN, into an IP core. The target is to have a model implemented on FPGA that can process 100-200 nodes (hits) and 200-500 edges (hit connections) within $\sim$ O(10 μs).

FlowGNN Architecture with on-FPGA Implementation The current TrackGNN model in sPHENIX we are using has one GNN layer, which includes 4 multi-layer perceptron (MLP) layers for both node and edge embedding with a dimension of 8. The proposed architecture follows the message-passing framework in FlowGNN [6]: the node embeddings are processed first, followed by an adapter to orchestrate the node information to the correct edge processing units for edge embedding computation and message aggregation. We also use quantization to reduce the data precision and to reduce the memory and computation requirements. We use `ap_fixed<18, 6>` for node embeddings, edge embeddings, and model weights and bias, and use `ap_fixed<21, 9>` for messages and input node features. The target FPGA board is the Alveo U280, approximately twice as big as FELIX FLX-712. The resource utilization is as follows: 194K (14.9%) LUT, 214K (8.2%) FF, 406 (20.2%) BRAM, and 488 (5.4%) DSP. The processing latency is measured on-board in an end-to-end fashion, including graph and weight loading, model computation, and results readback. The latency for an average-sized input (92 nodes and 142 edges) is 8.82 μs at a 285 MHz clock. The time-size scatter plot for nodes and edges are shown in figure 5, where the x-axis is the time spent processing one graph, and the y-axis is the graph size in terms of the number of nodes and edges. Compared to CPU calculations, the FPGA was 99.86% accurate.

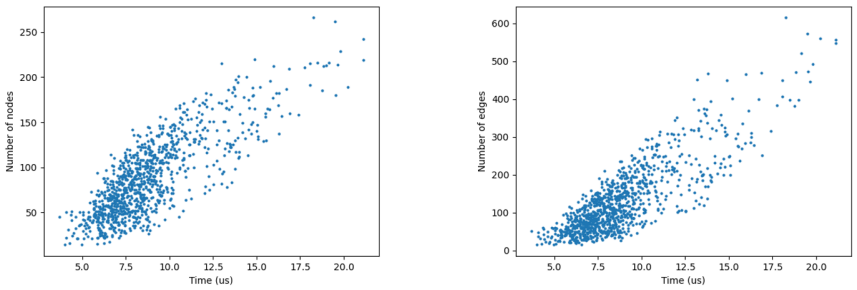

Figure 5. The time-size scatter plot of the TrackGNN model measured at Alveo U250. Left: for the nodes size. Right: for the edges size.

hls4ml A parallel effort to translate the model using hls4ml just started. The framework has been extensively used at CMS and we expect to have a first version of the inference at the end of October.

5 Summary and Outlook

Different modules of the AI-Engine has been tested independently, the final development is focused towards combining them into a single FPGA board. A further iterations with the model developers is done to account for the FPGA utilization. The beam test is expected in 2024 during RHIC pp run.

Acknowledgments

This project is funded by the United States Department of Energy, funding calls FOA-0002490, FOA-0002875 and Los Alamos National Laboratory LDRD program.

References

[1] sPHENIX Collaboration *sPHENIX Beam Use Proposal*, **sPH-TRG-2020-001** (2020).

[2] K. Chen, H. Chen, J. Huang, F. Lanni, S. Tang and W. Wu, *A Generic High Bandwidth Data Acquisition Card for Physics Experiments*, **IEEE Trans. Instrum. Meas. 69** (2020) no.7, 4569-4577,

[3] Xuan, T., Borca-Tasciuc, G., Zhu, Y., Sun, Y., Dean, C., Shi, Z. & Yu, D. *Trigger Detection for the sPHENIX Experiment via Bipartite Graph Networks with Set Transformer.*, **Machine Learning And Knowledge Discovery In Databases** (2023) pg. 51-67.

[4] F. Fahim, B. Hawks, C. Herwig, J. Hirschauer, S. Jindariani, N. Tran, L. P. Carloni, G. Di Guglielmo, P. Harris and J. Krupa, *et al. hls4ml: An Open-Source Codesign Workflow to Empower Scientific Low-Power Machine Learning Devices*, arXiv:2103.05579.

[5] Vitis High-Level Synthesis, Xilinx. https://www.xilinx.com/products/design-tools/vivado/high-level-design.html

[6] R. Sarkar, S. Abi-Karam, Y. He, L. Sathidevi and C. Hao, *FlowGNN: A Dataflow Architecture for Real-Time Workload-Agnostic Graph Neural Network Inference*, **2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)**, Montreal, QC, Canada, 2023, pp. 1099-1112, doi: 10.1109/HPCA56546.2023.10071015.