Contribution ID: **70**   Type: **Oral**

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

Monday 2 October 2023 15:25 (20 minutes)

The RHIC interaction rate at sPHENIX reaches around 3 MHz in pp collisions and requires the detector readout to reject events by a factor of over 200. Critical measurements 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 propose a new approach based on real-time AI technology, adopt an FPGA-based implementation using a FELIX-712 board with the Xilinx Kintex Ultrascale FPGA, and deploy the system in the detector readout electronics loop for real-time trigger decision.

Summary (500 words)

The interaction rate in sPHENIX at RHIC will be increased to 3 MHz for pp collisions. The sPHENIX experiment produces streamed readouts for the MAPS based Vertex Detector (MVTX), Intermediate Silicon Tracker (INTT), and Time Projection Chamber (TPC) tracking detectors and triggered readouts for the calorimeters. Archiving the full detector data streams exceeds the current DAQ bandwidth limits. The proposed triggering system consists of a comprehensive pipeline based on Graph Neural Networks (GNN) to perform online displaced vertex tracklet-based analysis to trigger the TPC streams, enabling efficient data acquisition.

The MVTX and INTT send raw data readout streams (300 kHz and 3 MHz, respectively) into their respective Front-End Link eXchange (FELIX)-712 boards, with six boards for MVTX and eight boards for INTT. The FELIX consists of a Kintex Ultrascale XCKU115FLVF1924-2E FPGA, a 16-lane Gen-3 PCIe card, and 48 transmitter and receiver optical links. Two FELIX-712 will be used to implement the GNN (AI Engine), with each FELIX-712 processing half hemisphere of the MVTX and INTT event. The aim is to reuse the PCIe Interface of the FELIX boards, which allows taking advantage of the FELIX Software Infrastructure. A total of 144 incoming data fibers with 3.2 Gbps bandwidth from MVTX alone must be zero-suppressed, multiplexed, and transferred to the AI engine using SFP+ optical links with 8b10b encoding. The optical links have been tested for transferring 14 Gbps data rate reliably with a low bit error rate. Consequently, we reduce the number of fibers from 144 to 48. There will be 24 links available for MVTX and 24 links for INTT for each hemisphere.

The AI Engine performs a fast inference for displaced tracks from heavy quark decays. The implementation is based on GNN using PyTorch ecosystem with FlowGNN models. The hls4ml package generates an intellectual property (IP) core for FPGA implementation. The entire latency for the trigger signal must be within the order of 10μs, which leaves the GNN model to perform inferences with less than 5μs latency per event.

Moreover, the reconstruction of heavy flavor particles requires continuous monitoring and adjustment of the beam trajectory, detector alignment, conditions, anomalies, and GNN latent variables. This implies the need for autonomous monitoring and feedback to control the overall real-time workflow. The final system will be based on a CPU/GPU platform connected to the AI engine over the PCIe, creating an embedded system.

The first version of AI Engine is based on the rm-4.11 version of the FELIX-711 Firmware and interfaces the tdaq-09-04-00 TDAQ Release (Software 4.2.4 and Driver 4.9.1). It utilized 51.8% LUT, 38.86% FF, and 59.68% BRAM.

The current effort aims to develop a demonstrator that uses VC-709/FELIX-711 as a data streamer and FELIX-712 as an AI engine to verify the system's feasibility in bandwidth, latency, and FPGA utilization. This is the first trigger R&D at RHIC using a unique hybrid mode of continuous and triggered readouts based on

FPGA accelerated AI/ML modules. A deployment is planned for sPHENIX 2024 run and possibly the future Electron-Ion Collider.

Primary author: KVAPIL, Jakub (Los Alamos National Laboratory (US))

Co-authors: Mr OLVERA, Alejandro (University of North Texas); HAO, Callie; DEAN, Cameron (Massachusetts Inst. of Technology (US)); DA SILVA, Cesar Luiz (Los Alamos National Laboratory (US)); Prof. YU, Dantong (New Jersey Institute of Technology); ROLAND, Gunther (Massachusetts Inst. of Technology (US)); JHENG, Hao-Ren (Massachusetts Inst. of Technology (US)); PEREIRA DA COSTA, Hugo; SCHAMBACH, Johann Joachim (Oak Ridge National Laboratory - (US)); DURHAM, John Matthew (Los Alamos National Laboratory); CHEN, Kai (Central China Normal University); RIGATTI, Micol (Fermi National Accelerator Lab. (US)); LIU, Ming Xiong (Los Alamos National Laboratory); TRAN, Nhan (Fermi National Accelerator Lab. (US)); WUERFEL, Noah; HEN, Or; LI, Pan; HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US)); Prof. FU, Song (University of North Texas); LI, Xuan (Los Alamos National Laboratory); Dr CORRALES MORALES, Yasser (Los Alamos National Laboratory (US)); LEE, Yen-Jie (Massachusetts Inst. of Technology (US)); Dr LIN, Youzuo (Los Alamos National Laboratory); Dr SHI, Zhaozhong (Los Alamos National Laboratory)

Presenter: KVAPIL, Jakub (Los Alamos National Laboratory (US))

Session Classification: Trigger and Timing Distribution

Track Classification: Trigger and Timing Distribution