Intelligent experiments through real-time AI: Fast Data Processing and Autonomous Detector Control for sPHENIX and future EIC detectors

Hannah Bossi (MIT) for the sPHENIX Collaboration

Laboratory for Nuclear Science

HP2024 NAGASAKI

Abstract

A collaboration of scientists from LANL, MIT, FNAL, NJIT, ORNL, and GIT, supported by the DOE Office of Science Nuclear Physics AI Machine Learning initiative, is exploring advanced AI technologies to tackle data processing challenges at RHIC and the future EIC [1]. The main objective is to develop a demonstrator for real-time processing of high-rate data streams from sPHENIX experiment tracking detectors to identify rare heavy-flavor events in proton-proton (p+p) collisions. Our innovative approach integrates streaming readout with an intelligent control system, utilizing FPGA hardware to accelerate AI inference. This improves the efficiency of collecting rare heavy-flavor events in high-rate p+p collisions (~1 MHz), optimizing the use of limited DAQ bandwidth (~15 kHz). We employ Graph Neural Network-trigger algorithms, trained on sPHENIX p+p collision simulation data, and use the hls4ml package to convert AI models into firmware. These real-time AI technologies are deployed on FLX712 boards equipped with Xilinx Kintex Ultrascale FPGAs. Our approach is also adaptable to other fields requiring high-throughput data streams and real-time detector control, including future EIC experiments. This talk will highlight AI-driven heavy-flavor triggering for sPHENIX and the development of DIS electron tagger algorithms for the EIC, showcasing the transformative potential of AI and FPGA technologies in real-time data processing for high-energy nuclear and particle experiments.

The sPHENIX Detector

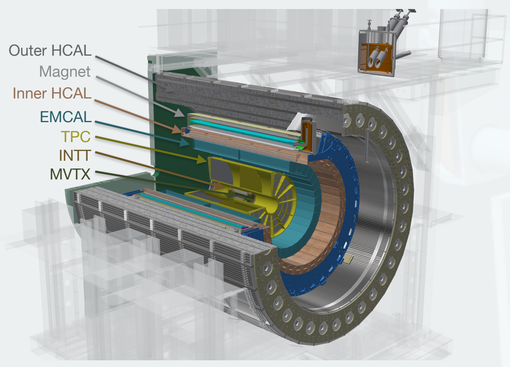


Figure 1: Schematic view of the sPHENIX detector.

- Located at the Relativistic Heavy Ion complex at Brookhaven National Laboratory.
- High resolution vertexing with inner-trackers combined with high resolution and large acceptance calorimeters make sPHENIX a state-of-the-art detector for jet and heavy-flavor physics!
- The tracking system of sPHENIX (INTT, MVTX, TPC) is capable of streaming readout (SRO).
- TPC dominates streaming rate, cannot save all streamed data.
- **Goal of this project:** Will stream INTT and MVTX data to Field Programmable Gating Arrays (FPGAs) where machine learning (ML) algorithms are embedded in order to tag heavy-flavor (HF topologies).

Firmware Implementation & Demonstrator

- For the streaming of INTT and MVTX data, the FPGA implementation of the data decoding and the hit clustering will be performed using conventional logic and the tracking and HF signal tagging will be performed using machine learning, which is shown in Figure 2.

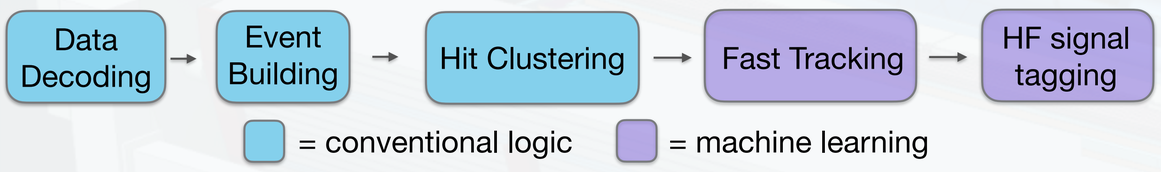
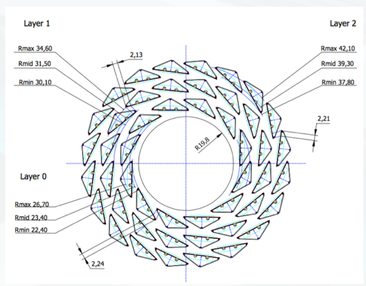


Figure 2: Schematic view of the algorithms implemented on FPGAs. Parts of the procedure that will be implemented using conventional logic are shown in blue and parts that will be implemented using machine learning are shown in purple.

Decoder

- MVTX consists of 48 staves with 9 chips per stave with > 500k pixels per chip.
- Each chip's information is sent to its own decoder.
- In p+p collisions, low occupancy (~20 hits per chip per collision).
- Decoding works sequentially where first the layer/stave value, bunch crossing ID (time), chip value and row and column of each active pixel hit is decoded, respectively.



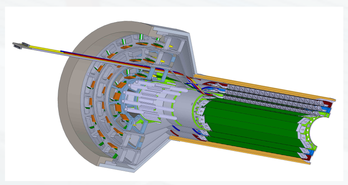
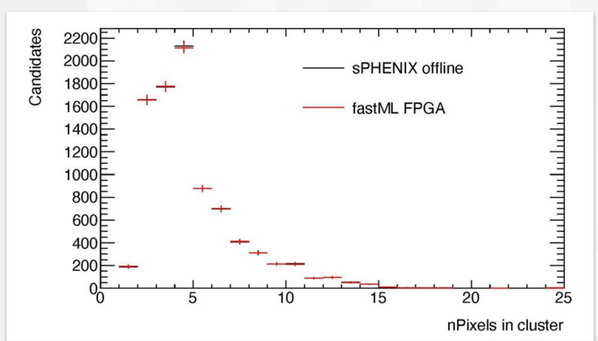


Figure 3: (Left) Stave layout of the MVTX in the beam view. (Right) Detector rendering of the MVTX.

Clusterizer



- ALPIDE chips reads out data in double columns from 0 to 1023.
- Clusters are assembled as they arrive.
- Can achieve 13.5 μm cluster resolution.
- Code written in C++ validated by comparing to the sPHENIX offline algorithm, demonstrating good agreement as shown in Figure 3.
- This code was the translated to VHDL using vitis_hls [2].

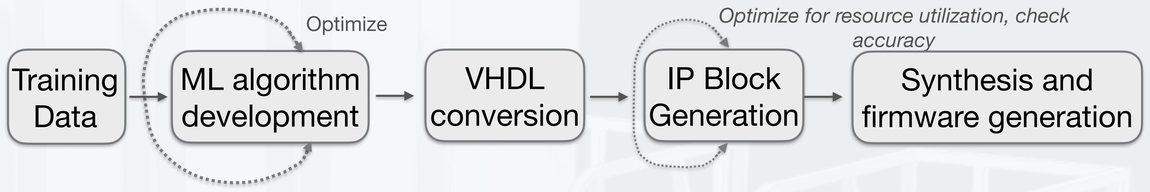
Figure 4: Comparison of the number of pixels in a cluster between sPHENIX offline clustering algorithm and the FPGA result.

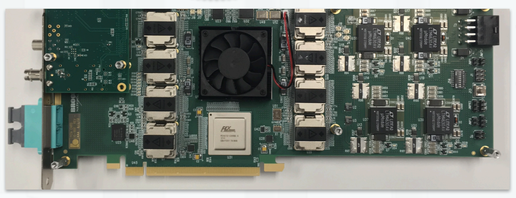
References

[1] NP 2022 Accelerator R&D | U.S. DOE Office of Science (SC). https://science.osti.gov/np/Research/DOE-N P-Accelerator-R-D-PI-Meeting/DOE-NP-Accelerator-R-D-PI-Meeting-Presentations-2022.
[2] AMD Vitis HLS, https://www.amd.com/en/products/software/adaptive-socs-and-fpgas/vitis/vitis-hls.html
[3] hls4ml, https://fastmachinelearning.org/hls4ml/
[4] "FlowGNN: A Dataflow Architecture for Real-Time Workload Agnostic Graph Neural Network Inference", arXiv:2204.13103
[5] "Semi-Supervised Classification with Graph Convolutional Networks", T. Kipf and M. Welling, arXiv:1609.02907

ML algorithms on FPGAs

- Our ML algorithms must have low latency and resource utilization. Not trivial since the algorithms themselves are complex.
- To help with this, use hls4ml [3] to both translate algorithms into high level synthesis and also generate Intellectual Property (IP) core. In some cases FlowGNN [4] also used to generate VHDL code.





- Currently using a BNL FLX712 board which contains Xilinx Kintex Ultrascale FPGA for decision hardware.
- Advantageous due to large firmware and software support and large amount of optical IO. Also is the same as what is currently in use at sPHENIX DAQ.

Figure 5: (Left) Photograph of BNL FLX712 board.

Track construction

- **Step 1:** Perform an edge candidate generation by connecting all clusters (nodes) together via edges and also applying some geometric constraint.
- **Step 2:** Perform a candidate classification of the edges using a graph convolutional network [5] that predicts the true edge candidates.
- **Step 3:** Then construct the track from the edge candidates.
- **Step 4:** Use a least squares method to determine the track momentum via its curvature.

Edge Prediction Step	
LUT	194 k (14.9%)
FF	214 k (8.2%)
BRAM	406 k (20.2%)
DSP	488 k (5.4%)

Table 1: Resource utilization for edge prediction using an Alveo U280 accelerator card and translation via FlowGNN.

Aggregation Step	
LUT	22.8k (3.44%)
FF	15.6 k (1.18%)
BRAM	0 (0%)
DSP	76 (1.38%)

Edge Prediction Step	
LUT	19 k (2.8%)
FF	27.5k (2.1%)
BRAM	498 (9.02%)
DSP	311 (1.4%)

Table 2: Post logic synthesis resource utilization for the aggregation step (latency 140 ns) and the edge prediction step (latency 365 ns) using hls4ml.

- A few different iterations where small changes lead to big improvements!
- **Using least squares method lead to a 14% improvement in the accuracy.**
- Reducing number of edges by 50% saved on resources, but only cost 0.4% of the accuracy

HF signal tagging

- For the tagging of the HF signal topology, a Bipartite Graph Network with a Set Transformer (BGN-ST).
- This is an attention-based algorithm that allows modeling of effects such as two tracks sharing a common vertex, determination of the collision vertex, and whether or not the track origin vertex is centered around the collision vertex.
- Track node input vectors contain a total of 37 features including…
  - 5 hits (INTT + MVTX)
  - Length of each edge
  - Angle between edges
  - Total length of the edges
  - Track radius (proportional to track p_T)
- Aggregators (primary and secondary vertices)
- **Performance on $D^0 \rightarrow K\pi$**
  - Current tracklet algorithm has excellent accuracy of > 87%
  - AUC > 93%
- **Performance on Beauty Decays (no pileup, 0.05% of events)**
  - Current tracklet algorithm has excellent accuracy of > 91%
  - AUC > 97%
- Significant improvement over tagging method using clusters (hit-based) instead of tracks.

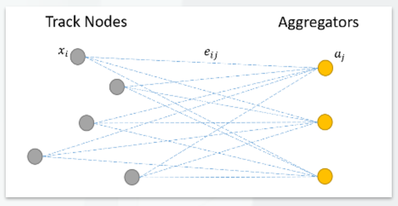


Figure 6: Track nodes and aggregators where $e_{ij} = s_{ij}x_i$ and s_{ij} is the weight.

Project Timeline

- **2021** Project started. Initial simulations constructed. First data for training.
- **2022** SRO development for INTT/MVTX. Fast tracking and trigger algorithm development. Initial FPGA bitstream synthesis. GPU feedback machine R&D.
- **2023** Refine interface between algorithms and detector readout. Update to match latest data stream and commissioning info
- **2024** Deploy device at sPHENIX
- **2030s** Deploy device at the EIC

Summary and conclusions

- Artificial intelligence and machine learning have the potential to revolutionize our approach collecting, reconstructing and understanding data, and thereby maximizing the discovery potential in the new era of nuclear physics experiments.
- In this project we use ML algorithms that are embedded onto FPGAs in order to to tag heavy-flavor event topologies using streamed data from the inner trackers (INTT + MVTX) of sPHENIX.
- This is beneficial as it promises a dramatic increase for the amount of available data for heavy-flavor analyses, crucial to the physics program of sPHENIX.
- All components in the FPGA pipeline are developed, putting all components together on a single FPGA is in progress!

U.S. DEPARTMENT OF ENERGY | Office of Science | OAK RIDGE National Laboratory | Fermilab | Los Alamos NATIONAL LABORATORY | UNT | Brookhaven National Laboratory