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

Micol Rigatti
Fast Machine Learning for Science Workshop 2023
25 September 2023
The sPHENIX Experiment at RHIC/BNL will take $p + p$, $p + Au$ and $Au + Au$ data at $\sqrt{S_{NN}} = 200 \text{ GeV}$ with unprecedented collision rates, approaching 10 MHz for $p + p$.

The maximum sPHENIX detector raw data rates exceed the throughput limits of the current sPHENIX DAQ design (300 Gbit/s) and the capacity of tape storage system and Computing resources.

Realizing the potential of modern nuclear physics (NP) experiments at colliders rely on the collection and processing of very large datasets, with data rates exceeding Terabits per second.
We propose to develop real-time AI technologies implemented in the detector readout electronics loop that address the challenges for the next generation of NP experiments at RHIC and EIC.

REAL-TIME AI

Extraction of critical data via selective streaming

Automated control and feedback for detector operation

that in the sPHENIX experiment means

particle tracking and heavy flavor reconstruction

beam spot and alignment determination
Technical Approach

A fast search for **displaced tracks** will be performed by AI-trained FPGA (FELIX) to identify tracks from heavy quark decays that are pointing away from the nominal beam center.
Hardware implementations of selective data streaming in sPHENIX will use the FELIX board.

FELIX is a 16-lane Gen-3 PCIe card with 48 transmitters and receiver optical links. The on-board FPGA is a Kintex Ultrascale XCKU115FLVF1924-2E.
The data for the Graph Neural Network are input from the **optical interface**. The GNN will search for displaced tracks and when found, will send a trigger signal to the TPC.

The **PCIe interface** is used in the developing phase to interact with the GNN and will be maintained with the same purpose.
Challenges

Meeting the **timing constraints**

- For the experiment → develop a GNN fast enough with its own resolution and space constraints
- Of the firmware → synchronize the data flow and avoid the worst negative slack

Test the functionality of the NN and the data routing mechanism

The FELIX card is a **data router**. It needs an application to be instructed on the data movement. The **application** relies on an **OS** and a **driver** interfacing the FELIX card.

**Interfacing the board to test the NN is the other real challenge**
Graph Neural Network

ML models have shown steady increases in performance on the triggering problem

**Good progress:**
- integrating ML models in FPGA
- reduced the number of hits while preserving the trigger performance
- end-to-end pipeline performance increases from 84% to 88%
- implemented event pileup. We can handle 20 events with an end-to-end accuracy of 78.08%.

<table>
<thead>
<tr>
<th>Data</th>
<th>Year</th>
<th>Metric</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT Tracks</td>
<td>2023</td>
<td>Accuracy</td>
<td>90.22%</td>
</tr>
<tr>
<td>Predicted Tracks</td>
<td>2022</td>
<td>Accuracy</td>
<td>84.01%</td>
</tr>
</tbody>
</table>
HLS4ML Results

- Network inputs: nodes=80, edges=100
  - Input network
    - Can be parallelized to be “nodes” times faster (i.e., 15ns)
  
  \[
  \begin{array}{|c|c|c|c|c|}
  \hline
  \text{Latency} & \text{BRAMs} & \text{DSPs} & \text{FFs} & \text{LUTs} \\
  \hline
  1.2 \text{us} & 6.5\% & 0.3\% & 5\% & 7.5\% \\
  \hline
  \end{array}
  \]

- Edge network

  \[
  \begin{array}{|c|c|c|c|c|}
  \hline
  \text{Latency} & \text{BRAMs} & \text{DSPs} & \text{FFs} & \text{LUTs} \\
  \hline
  3 \text{us} & 15\% & 2\% & 20\% & 65\% \\
  \hline
  \end{array}
  \]

- Node network (results from HLS synthesis)
  - Neet to optimize the scatter_add function (expecting ~2us for the net)

  \[
  \begin{array}{|c|c|c|c|c|}
  \hline
  \text{Latency} & \text{BRAMs} & \text{DSPs} & \text{FFs} & \text{LUTs} \\
  \hline
  12 \text{us} & 42\% & 7\% & - & - \\
  \hline
  \end{array}
  \]
Thank you!
The sPHENIX application

- Starting in 2023
- Tracking system:
  - MAPS-based vertex detector (MVTX)
  - Silicon strip tracker (INNT)
  - Time projection chamber (TPC)
- Calorimeters: electromagnetic (EMCAL) and hadronic (HCAL)
- 1.4 T super conductive solenoid

sPHENIX will study the hot QCD medium, called Quark-Gluon Plasma (QGP), using scale-dependent probes to search for the existence of novel quasi-particles.