AI4EIC

P. Harris J. Huang

Summary

In this session we had 5 talks :

- Overview of AI in HEP readout : Dylan Rankin (MIT)
- EIC readout overview : Fernando Barbosa (JLab)
- Real-time AI tracking and tagging : Dantong Yu (NJIT)
- Real-time data compression with Bicephalous Convolutional Auto-Encoder : Yi Huang (BNL/CSI)
- Event tagging and triggering on FPGA : Sergey Furletov (JLab)

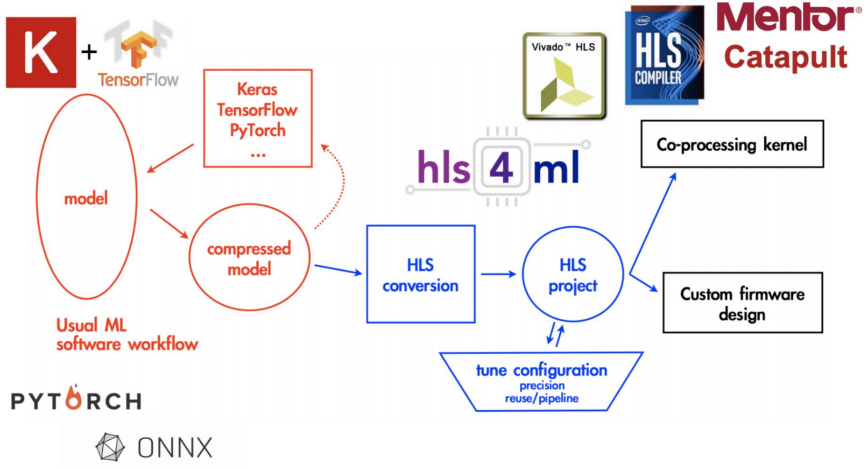
Focus of this session was on how do we extract data from the detector efficiently

Dylan Rankin: Overview of AI in HEP Readout

Review of the many new technologies developed at the LHC

LHC is working on integrating AI into all levels of data taking L1+HLT Trigger

hls4ml Workflow

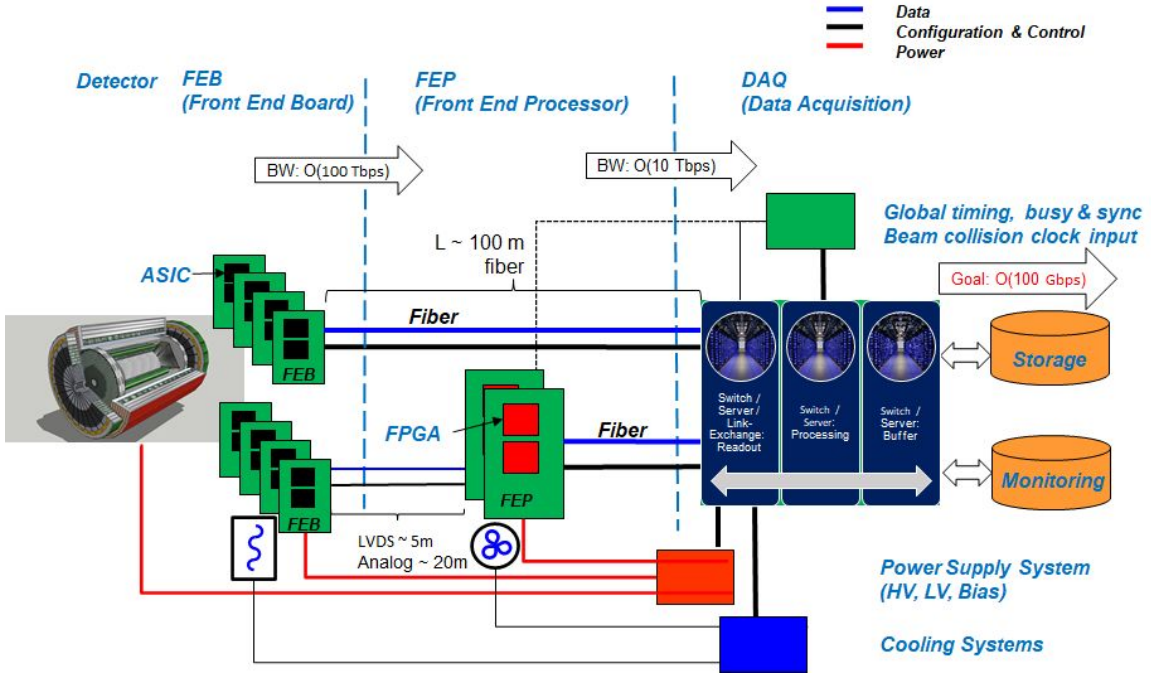


HLS4ML a multipurpose tool capable of converting NNs to FPGA firmware to run at very low latency

SONIC (inference as a service) enables the possibility of integrating GPUs/FPGAs/… to HLT

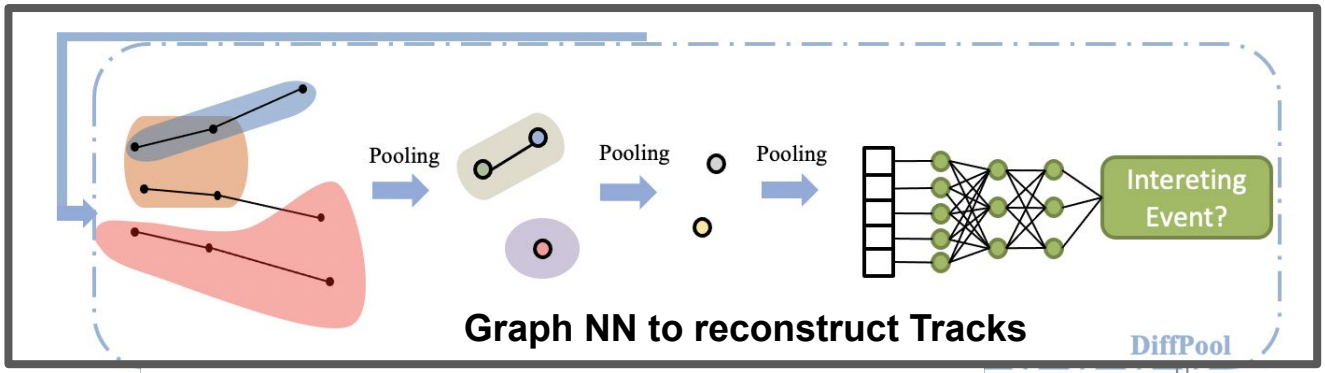
Fernando Barbosa: EIC readout overview

- Detailed discussion on the project of EIC readout system
- Layout landscape of AI application in EIC Readout
- Aim to record all O(100)Gbps EIC collision signal
- But AI readout pipeline design need to accommodate the uncertain noise and background level in experiment

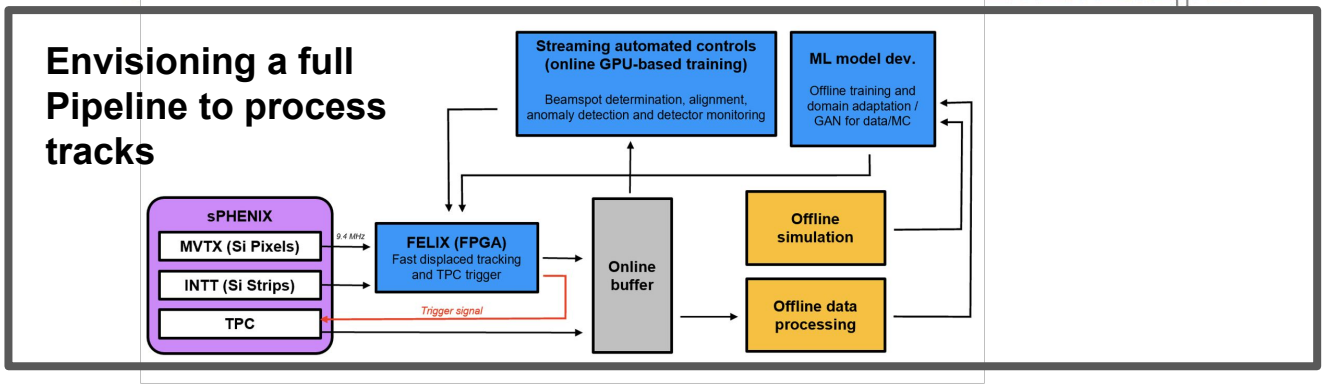


Dantong Yu: Real-time AI tracking and tagging

A trigger software pipeline for sPHENIX aimed at Real-time Tracking



A first version of a graph NN is envisioned which can perform tracking

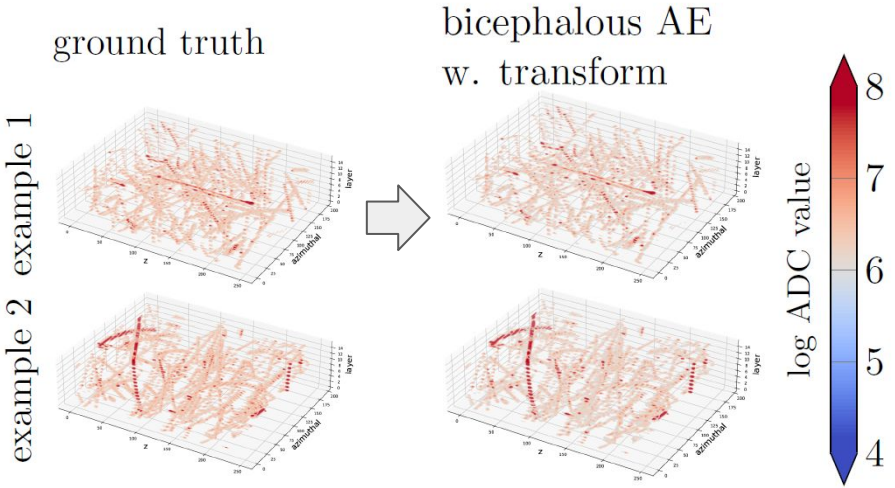


With this first version we can start to envision a full pipeline of operation

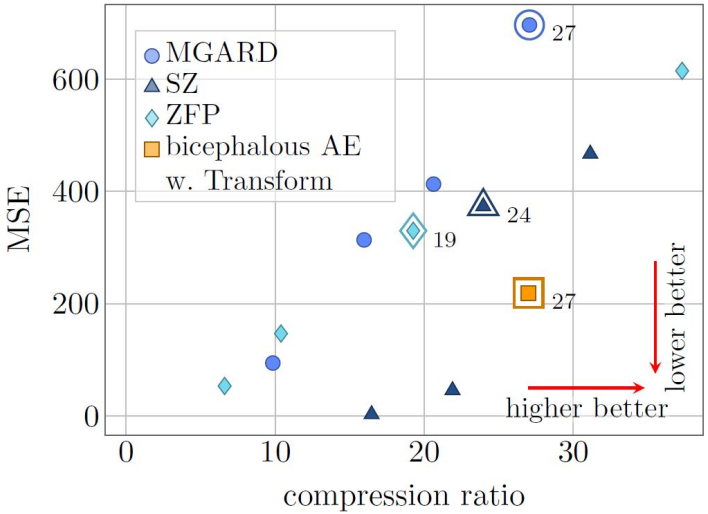
Yi Huang: Real-time data compression with Bicephalous Convolutional Auto-Encoder

- Custom designed CNN auto-encoder to compress tracking data with zero-suppression

Before compression → after decompression

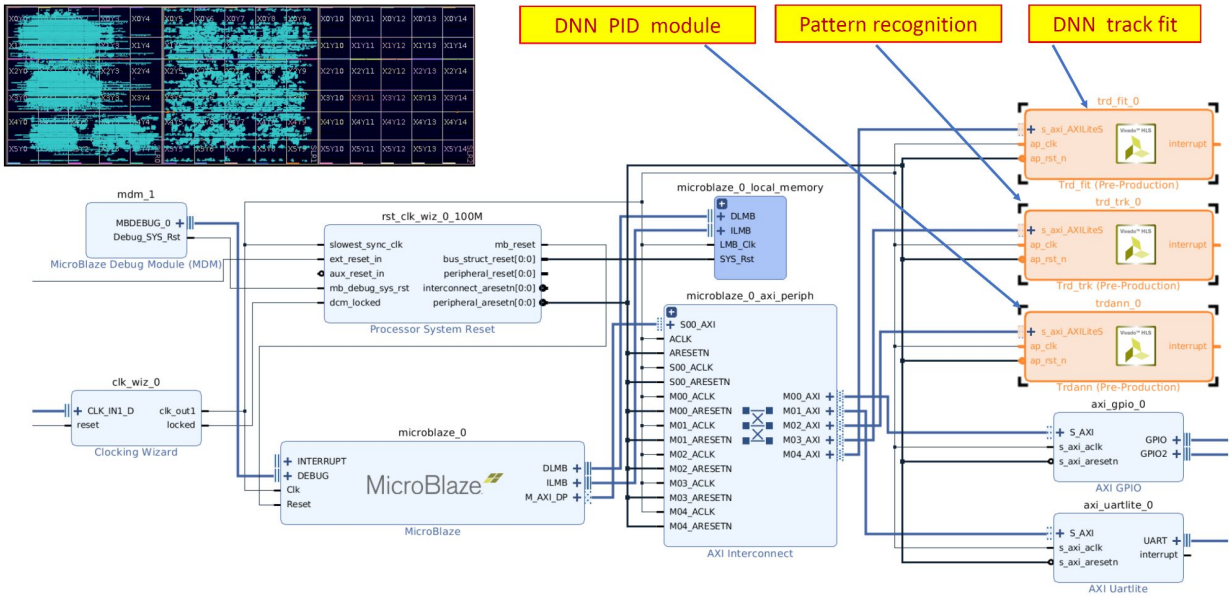


Best performance among lossy compressions



Sergey Furletov: Event tagging and triggering on FPGA

- A series of deployment of MLP/DNN on FPGA for PID and tracking of the GEMTRD detector and calorimeter. Promising for EIC physics-based data reduction in DAQ



Discussion Session

There is a trade off between our ability to buffer and our ability to compute:

- Given the large uncertainty of EIC noise/bgd, how flexible can we be?
  - At 100 Tb/s a 1ms buffer is 100 Gb, which is certainly possible
  - Allows for a long possible reconstruction latency, throughput is the same

What is the complexity of the models that we can build on FPGAs?

- HLS4ML/others can run very large (> 5M weights) models on FPGAs
- These models can take a long time to process
- Understanding the balance between model size and resources is critical