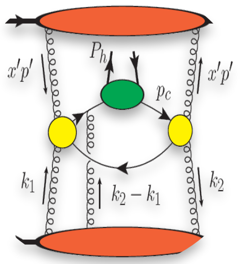
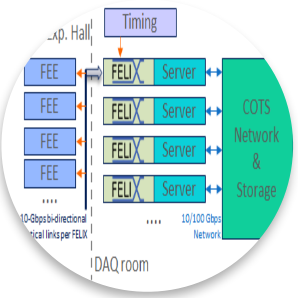
AI for Streaming readout

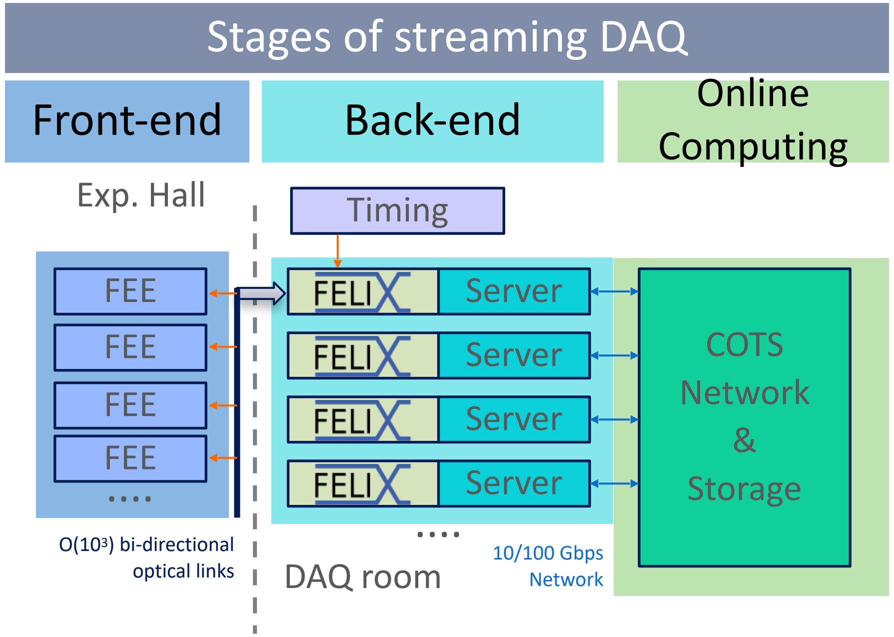
Streaming DAQ and real-time AI driven by NP science
A new and paradigm shift for experiments in next NP LRP



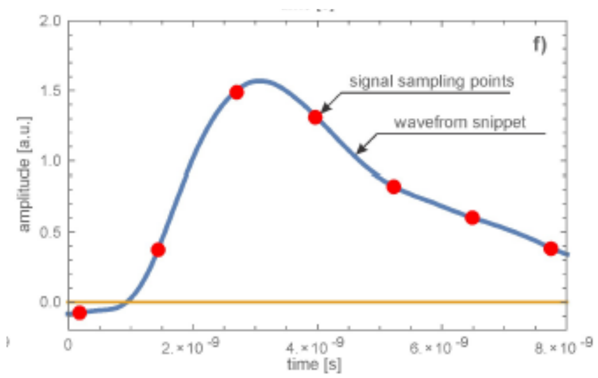


Physics need → Streaming DAQ → Opportunity for real-time AI → Enhanced physics program

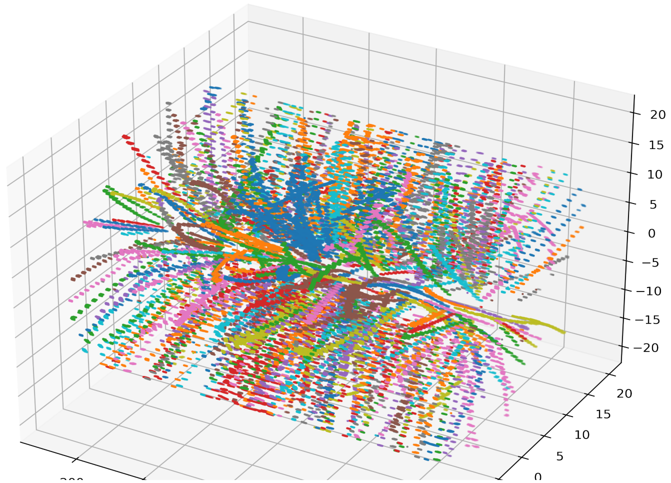
AI-based reduction at scale



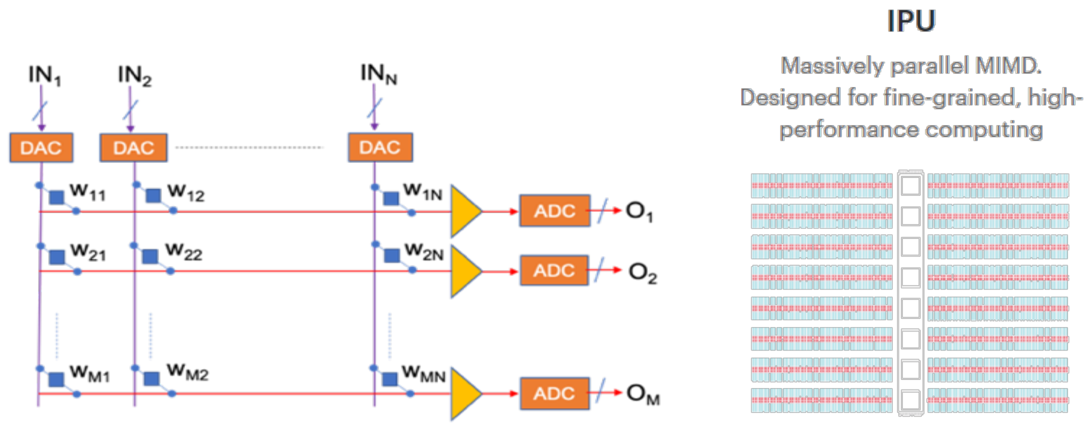
- Opportunities in all stages of EIC streaming DAQ, including
  - Waveform feature extraction at front end
  - Full detector lossy compression and noise filtering at online computing
- Exploring novel non-Von Neumann architecture NN processor, designed for high throughput NN inference
- On road to EIC, sPHENIX provides a streaming DAQ testbed with comparable data rate



[arXiv:2204.13223]



[Y. Huang, ICMLA21, RT22]



Non-Von Neumann architecture NN processor

Development of ML FPGA filter for particle identification and tracking in real time

Sergey Furletov
Jefferson Lab

Team :
F. Barbosa, L. Belfore, N. Branson, C. Dickover, C. Fanelli, D. Furletov,
S. Furletov, L. Jokhovets, D. Lawrence, D. Romanov

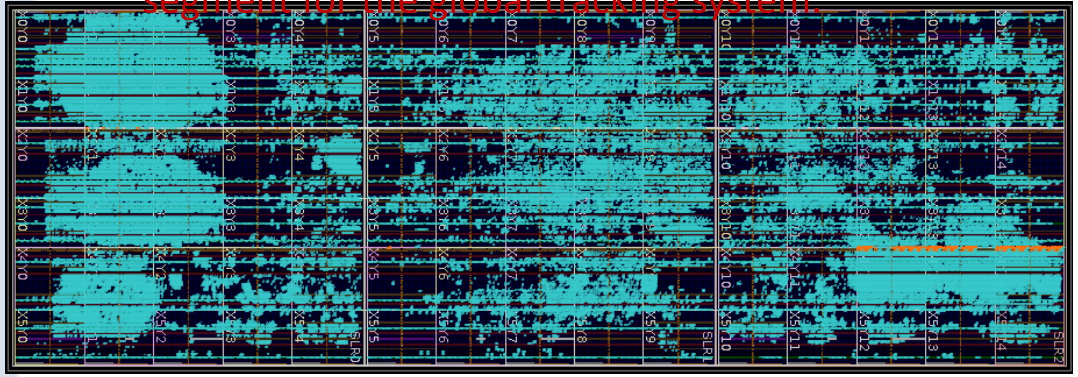
2nd workshop on Artificial Intelligence for the EIC

13 Oct 2022

A test bench for GEMTRD tracking and PID on FPGA

Jefferson Lab
Thomas Jefferson National Accelerator Facility

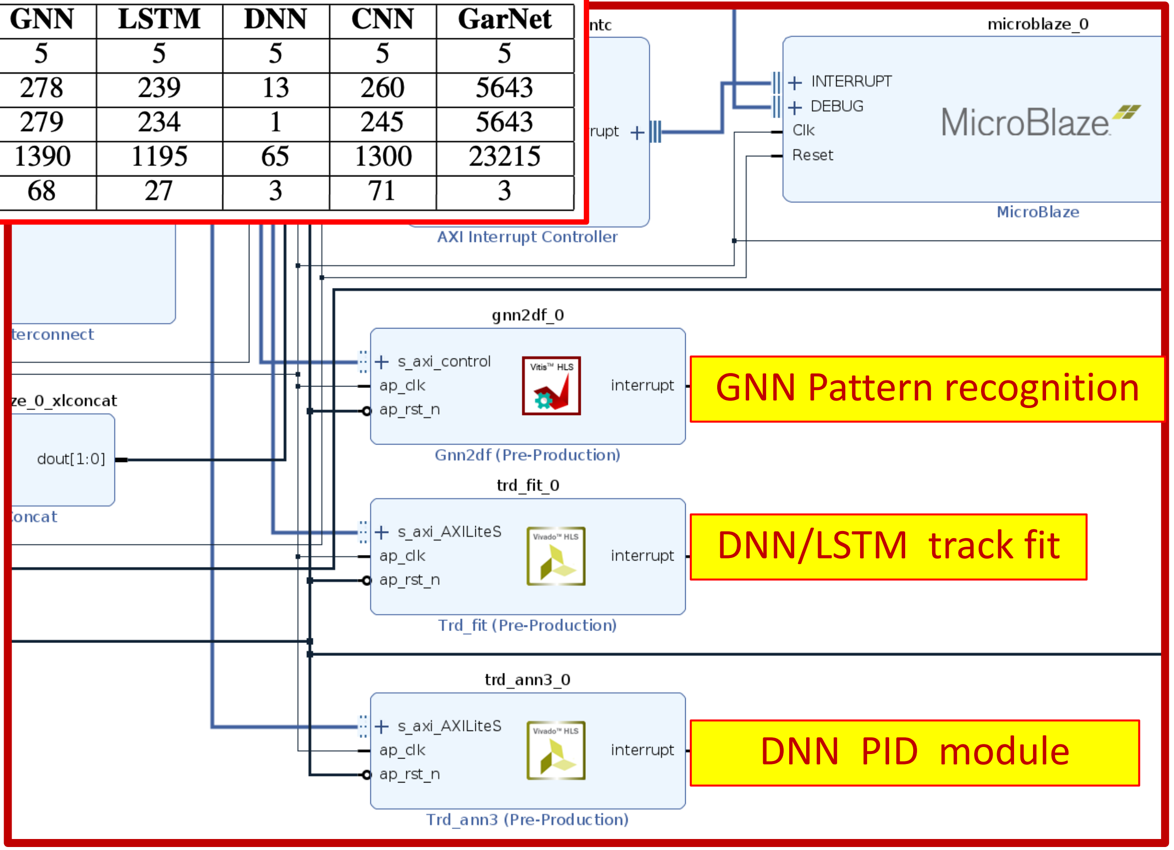
- *The e/pion separation in the GEM-TRD detector is based on counting the ionization along the particle track.*
- *For electrons, the ionization is higher due to the absorption of transition radiation photons*
- *So, particle identification with TRD consists of several steps:*
  - The first step is to cluster the incoming signals and create "hits".
  - The next is "pattern recognition" - sorting hits by track.
  - Finding a track
  - Ionization measurement along a track
  - As a bonus, TRD will provide a track segment for the global tracking system.



- *Several version of IPs were synthesized and tested on FPGAs.*
- *The logic test was performed with the MicroBlaze processor.*
- *We are currently working on a fast I/O interface to get data directly from the detector.*

FPGA IP SYNTHESIS SUMMARY.

	GNN	LSTM	DNN	CNN	GarNet
Clock, ns	5	5	5	5	5
Latency, clocks	278	239	13	260	5643
Interval, clocks	279	234	1	245	5643
Latency, ns	1390	1195	65	1300	23215
Utilization DSP (%)	68	27	3	71	3

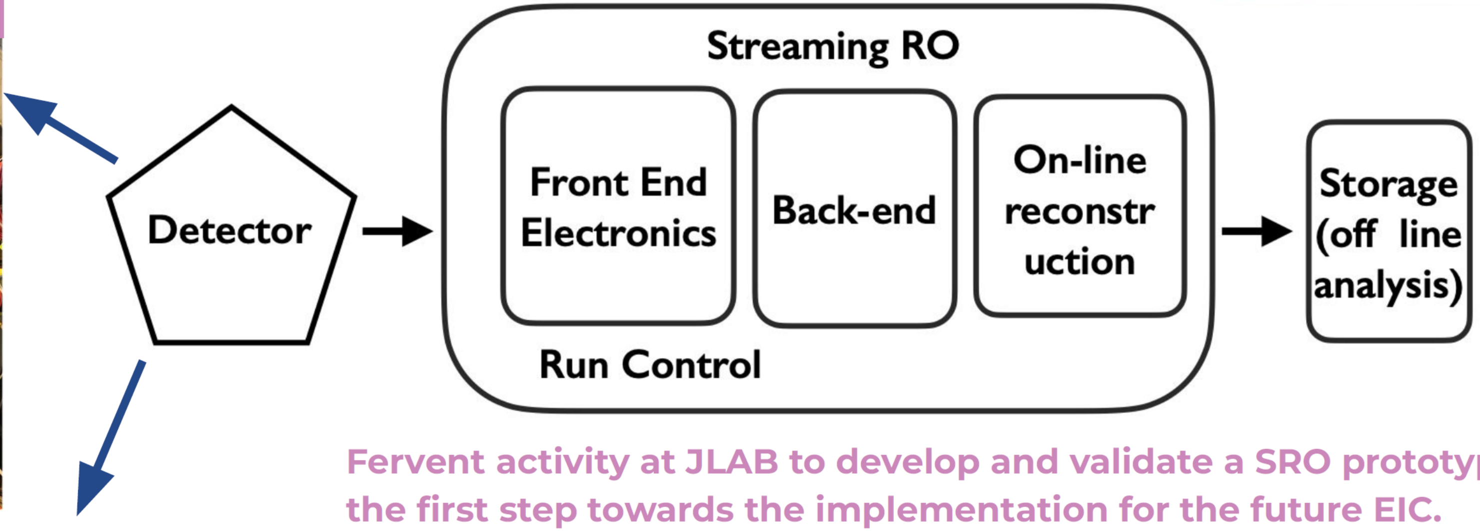


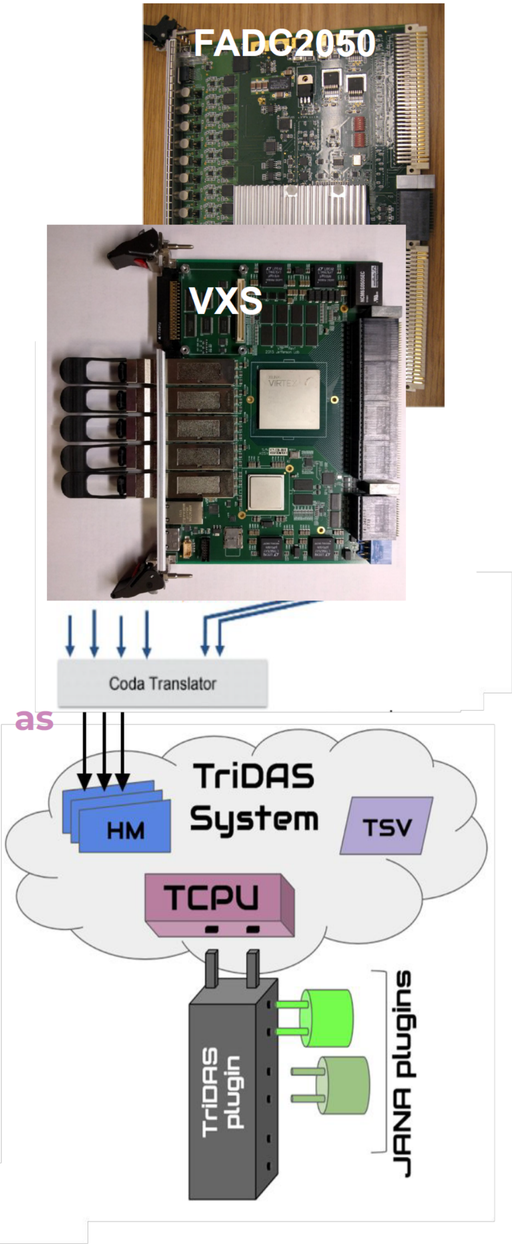
Streaming Readout DAQ @ JLAB

Mariangela Bondi

Jefferson Lab tests a next-generation data acquisition scheme

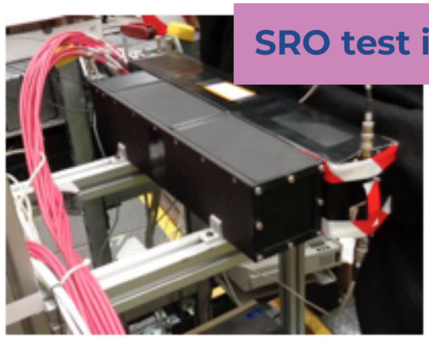
SRO test in HALL B





Fervent activity at JLAB to develop and validate a SRO prototype as the first step towards the implementation for the future EIC.

SRO test in HALL D



SiPM(left) & PMT(right) cal. prot.

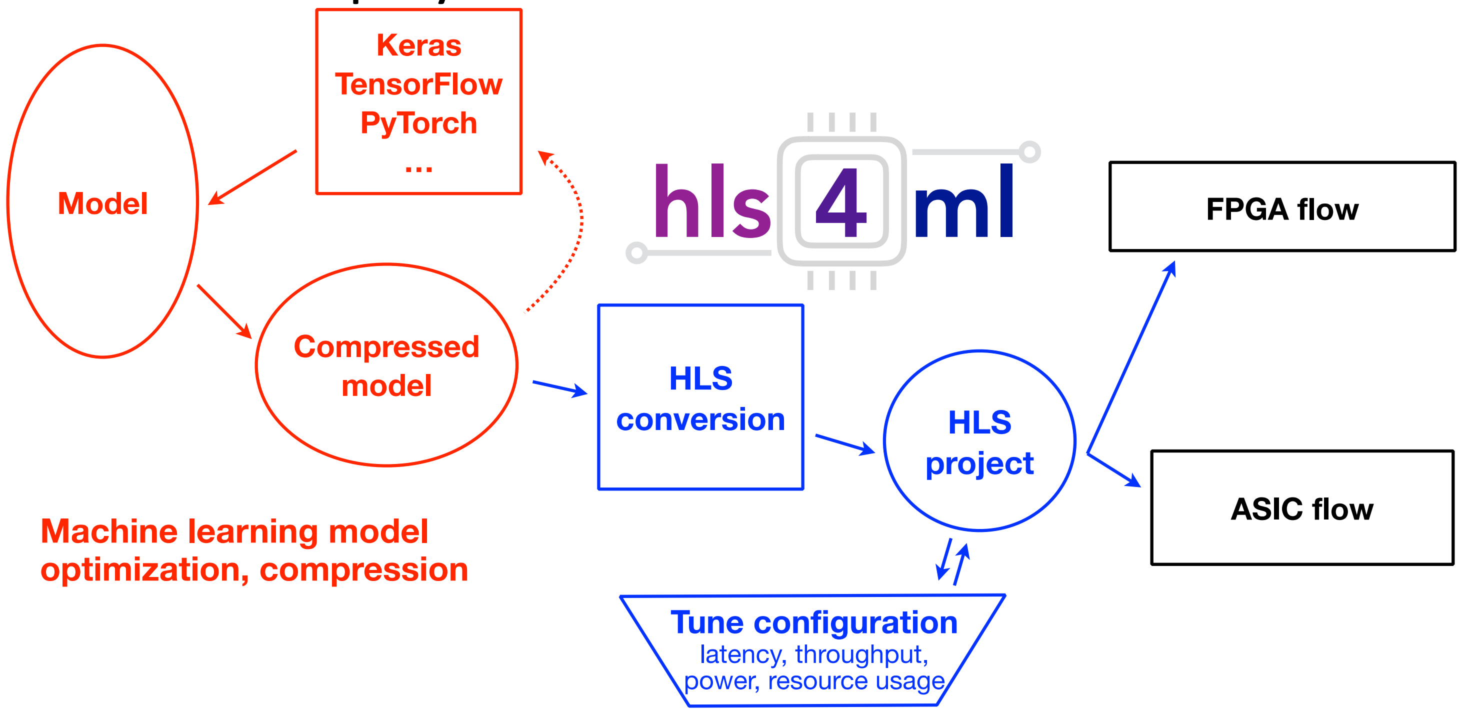
Waveboard

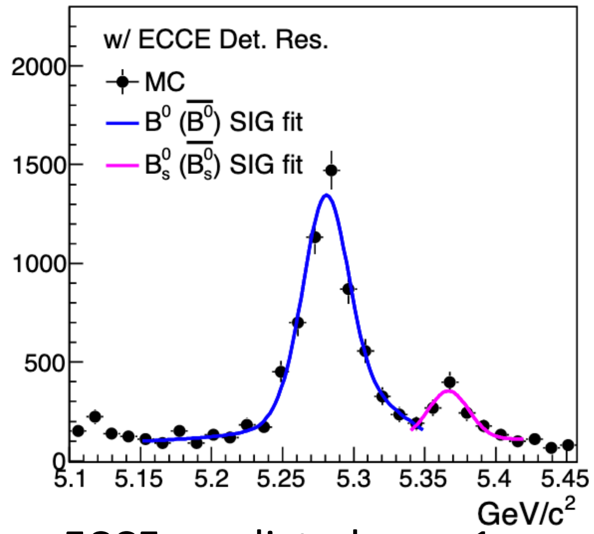
- Test in HALL D
  - EIC ECAL prototype
  - **GOAL**: compare Trigger-less to Triggered
- Test in HALL B:
  - 10.4 GeV electron beam on thin Pb/Al target
  - CLAS12 Forward tagger: ECAL + Hodo
  - Goal:
    - **DAQ performance**
    - **First AI application in streaming readout on real data tested online**
    - **Physics channel identification: pi0 production**

[arXiv.2207.10632]
[JINST 13 P07027]

ML for Heavy Flavor Identification, C. Dean

- Fast ML-based pre-selections can increase signal-to-background ratio
- We can reach rarer physics in year-1 compared to conventional selections alone
- Initial selection models show > 90% selection accuracy
- Model will be deployed on FPGA





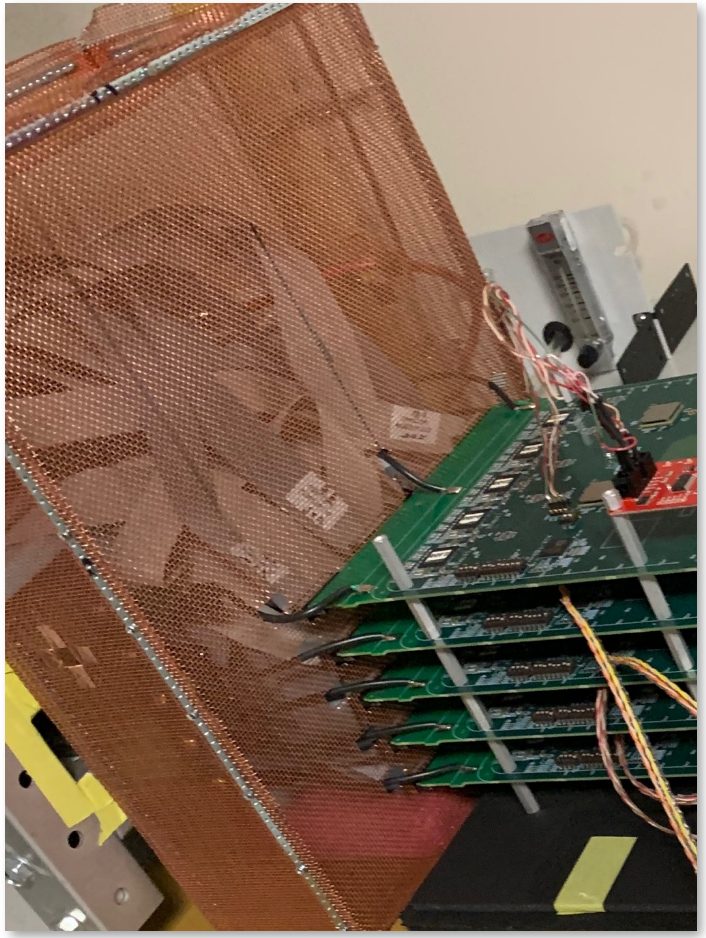
ECCE-predicted year-1 b-hadron fit with conventional selection

AI for streaming readout: an architectural perspective

Summary of talk

- Streaming readout architecture in general and DL algorithm in particular fit well on programmable devices such as FPGAs
- Technologies (HW) and methodologies (SW) to use them are getting complex and complex (HLS4ML, AI Engines, . . .)
- Some ideas can help riding technology: data compression on the edge to reduce bandwidth, knowledge distillation to obtain simpler networks, quantization aware training to exploit more efficient data types.
- A readout system could be composed of many processing nodes: we need a framework to distribute computing tasks and orchestrate them for real time processing.
- APEIRON project has been briefly presented: a system of multiple interconnected FPGAs with applications developed according to a dataflow programming model.
- We are able to map the directed graph of tasks and network channels with direct coupling among them.

Address Challenges of Autonomous Control and Experimentation



INDRA-ASTRA

Develop a prototype for a fully automated, responsive detector system as a first step towards a fully automated, self-conscious experiment.

R&D integrated with streaming readout and AI/ML efforts at Jefferson Lab

Status

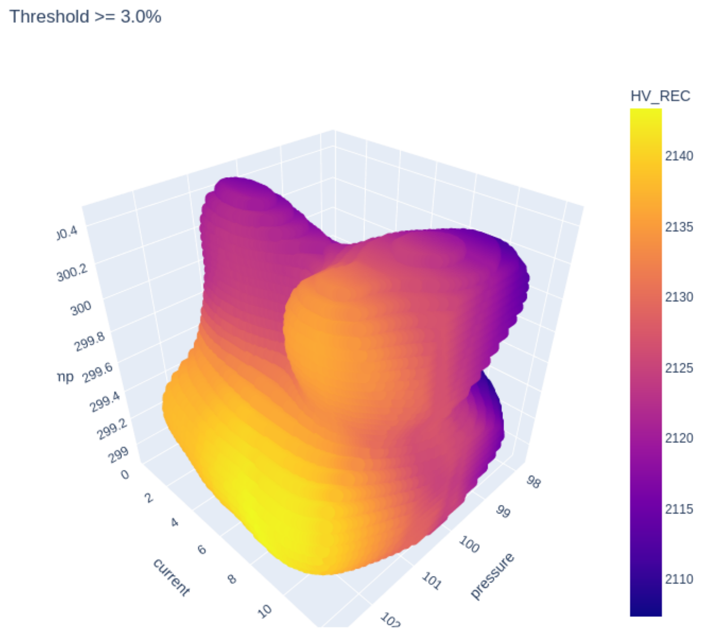
- Developed method for autonomous calibration of DIS experiments using baseline calibrations and autonomous change detection.
- Developed ADWIN2 and multiscale method for autonomous change detection.
- Versatile multiscale method can be used to increase reliability of data and find and fix issues on time.

Summary

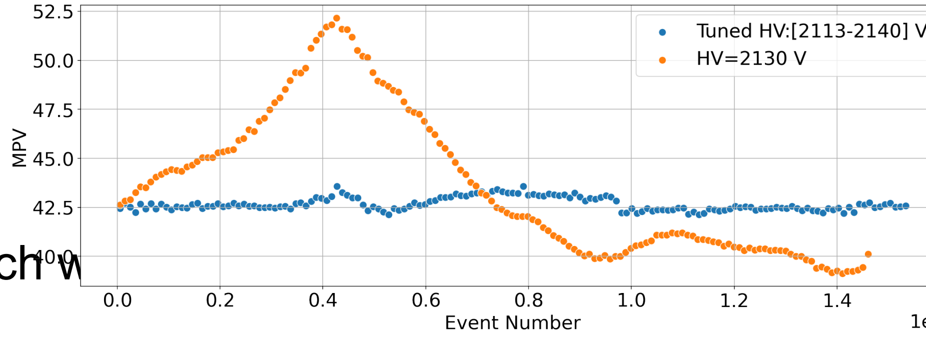
AI Experimental Control (and calibration) --GlueX Central Drift Chamber

Thomas Britton

- Successfully deployed an AI system to control and calibrate the GlueX CDC
  - ~ 1% error in predicting gcf (<< promised 5%)
  - Able to stabilize gcf via HV control
  - The system is reconfigurable on the fly
  - Uses estimated UQ in an actionable way
    - Lots of subtilties and avenues for research
      - Complex 3D surface
      - Limited data in some dimensions
        - Islands of data



- Looking to finish TtoD
  - HV dependence so a bit of bootstrapping probably required



- Probing other detectors and systems to apply these techniques
  - Potential self-learning to enable eminently deployable system which will only control when "confident" and efficiently learn when not

Jefferson Lab