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

Jakub Kvapil for the FastML team

TWEPP 2023 Topical Workshop on Electronics for Particle Physics
Geremeas, Sardinia, Italy

Los Alamos National Laboratory (LANL)
Fermi National Laboratory (FNAL)
Massachusetts Institute of Technology (MIT)
New Jersey Institute of Technology (NJIT)
Oak Ridge National Laboratory (ORNL)
Georgia Institute of Technology (GIT)

LA-UR-23-31115

FastML - Who are we?

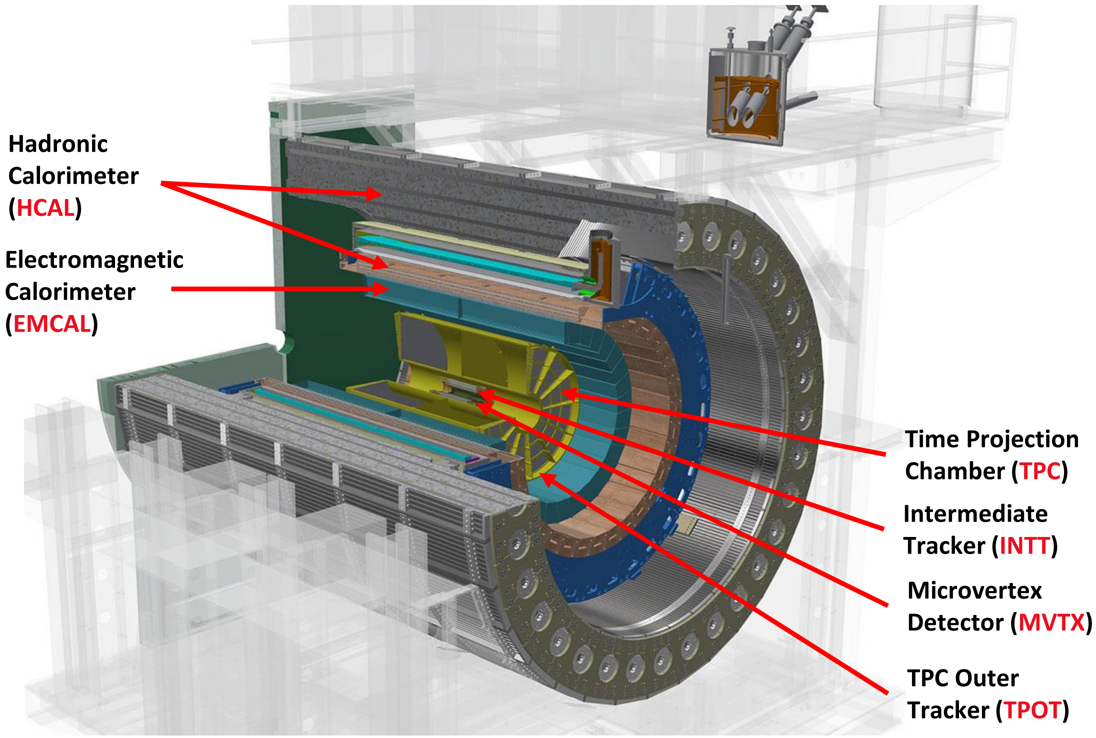
- Cross-discipline group of sPHENIX and LHC physicist, engineers, and computer scientists working on firmware-based ML applications data selections
  - sPHENIX is benefiting from a 2020 Department of Energy (DOE), USA funding call
- The mission
  - Efficiently extract critical and strategic information from large complex data sets
  - Address the challenges of autonomous control and experimentation
  - Artificial Intelligence for data reduction of large experimental data

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

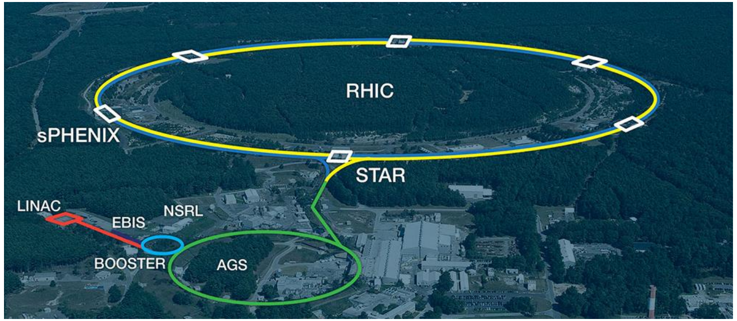
A proposal submitted to the DOE Office of Science

April 30, 2021 @ renewed for 2 more years in 2023

sPHENIX experiment



- Located at RHIC accelerator at BNL (USA)
- ~56 MHz accelerator clock with ~9.3 MHz BC
- Running period 2023-2025
- ~4m long, ~5m high, 1000 tons
- Tracking detectors (MVTX, INTT, TPC, TPOT) and calorimeters (EMCAL, HCAL)
- 1.4 T Magnetic Field, $|\eta| \leqslant 1.1$
- Tracking detectors capable of streaming readout, but unable to save all TPC data.
- 15 kHz designed Trigger Rate



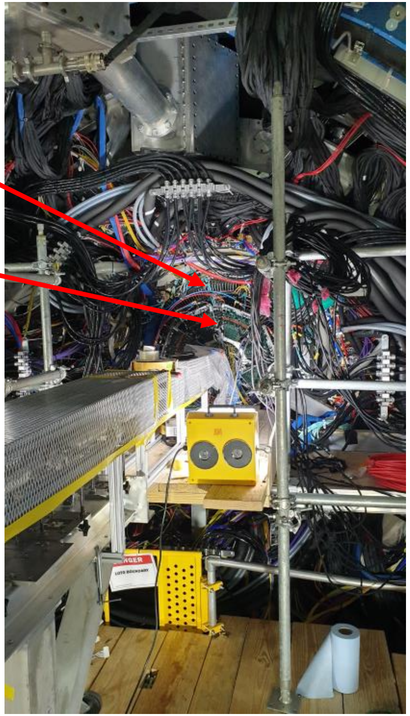
MVTX and INTT

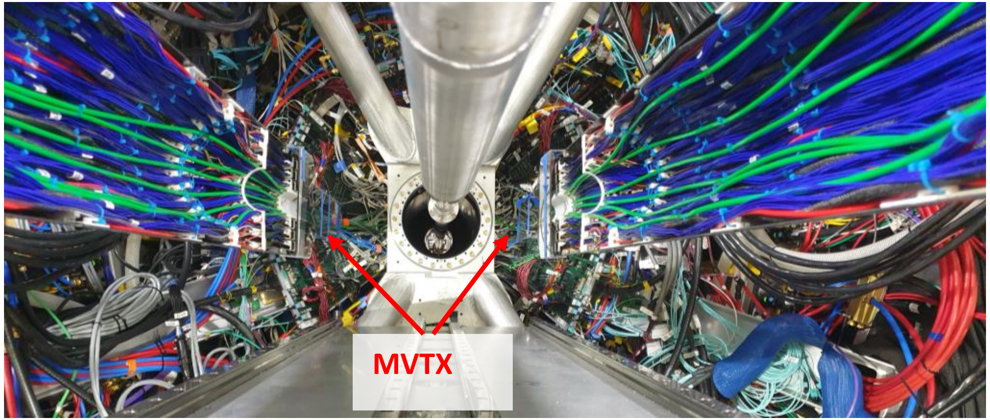
MVTX - Active area ~1685 cm^2

- Based on ALICE ITS2 ALPIDE chips, with ATLAS FELIX backend
  - Monolithic Active Pixel Sensors
  - Very fine pitch (27 μm x 29 μm)
  - Event Time resolution ~ 5 μs
  - 3 layers, 48 staves total, 9 chips per stave ~ 230M total channels

TPC

INTT





INTT

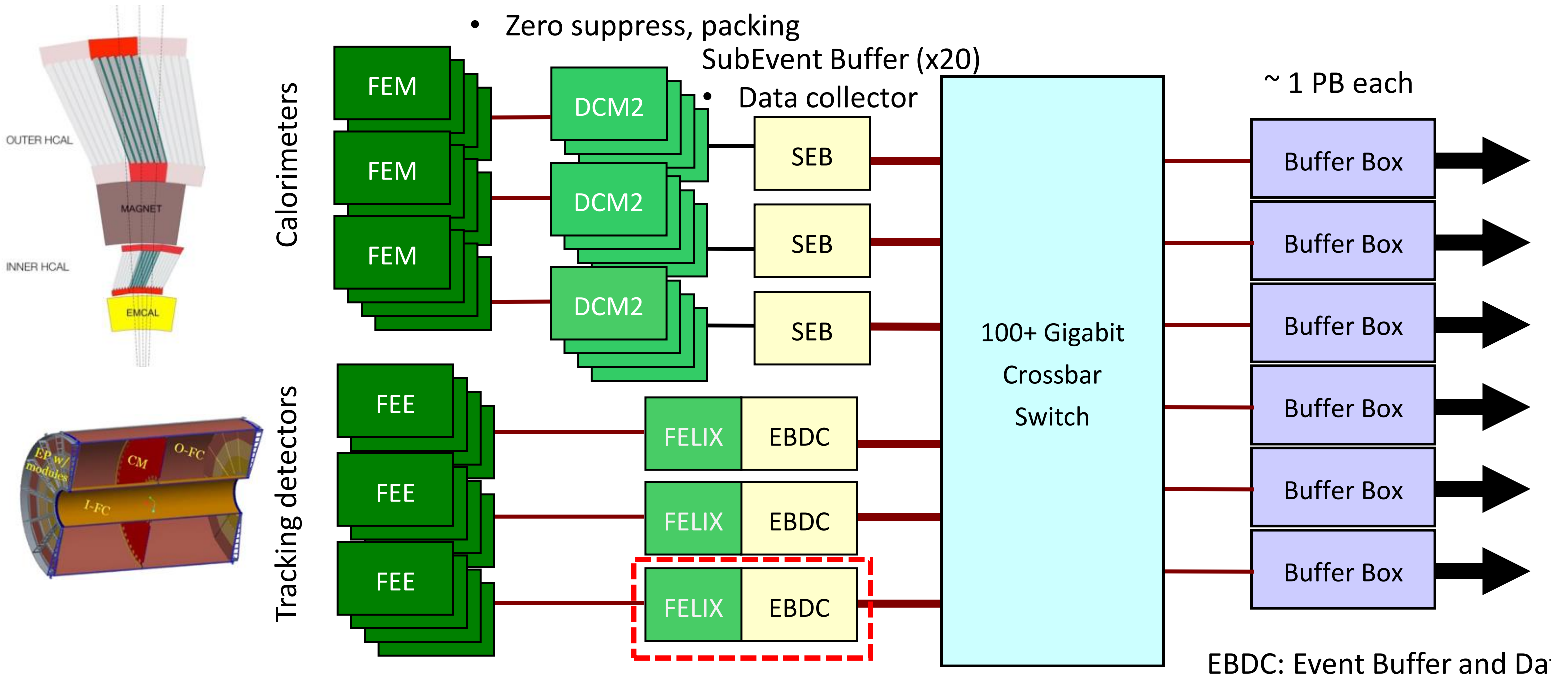
- Silicon Strip Detector
  - Hamamatsu silicon modules
  - Pitch 78 μm x 16 (or 20) mm
  - Excellent Time resolution ~100 ns
    (100 ns is the RHIC BC time)
  - 2 layers, 56 ladders total, 360k channels

sPHENIX Readout and AI-ML HF Trigger Integration

On Detector **Rack Room**

Front-End Module/Electronics

Data Collection Module



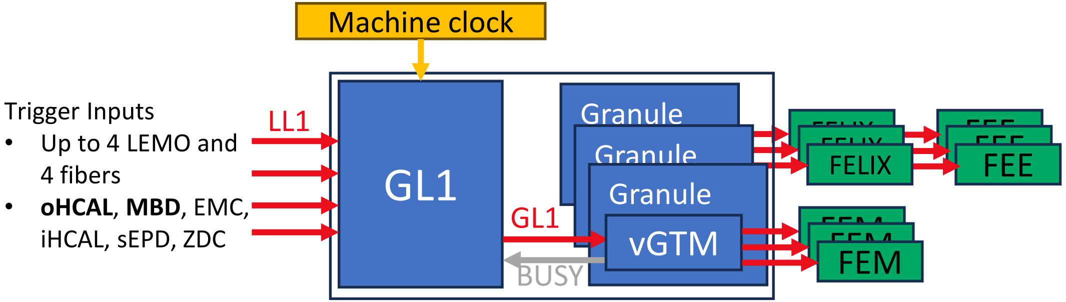
FELIX (MVTX+INTT) -> AI/ML -> Trigger

EBDC: Event Buffer and Data Compressor (x40)

- 6x MVTX, 8x INTT, 24x TPC

The timing and trigger distribution

- The Global Level 1 Trigger (GL1) and the machine clock is distributed via Granule Timing Module (GTM)
  - GL1 trigger is used by calorimeters and the TPC
  - GL1 transmits clock and trigger to the vGTM, which then transmits it to the FEE
    - vGTM is the adapter to a given detector
    - GL1 is maintaining the BUSY received from vGTM



Motivation – Heavy Flavour

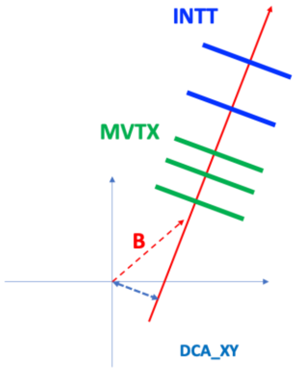
- Integrate the AI-based heavy flavour trigger system demonstrator into the sPHENIX experiment for p+p run in 2024 to R&D its feasibility, requirements, and constrains
  - Heavy-flavour (HF) events are very rare ~1% of Minimum Bias (MB) events at RHIC energy
  - RHIC collision rate is around 2-3 MHz, sPHENIX readout 15 kHz (DAQ - 300 Gb/s)
    - Trackers are Streaming Readout (SRO) capable, but can't save all TPC data
  - 10% trigger-enhanced SRO increases HF MB rate ~ 300 kHz
  - **ML HW tagging aims to sample remaining 90% of the luminosity using the tracklet reconstruction from the silicon trackers**

- The aim is to deploy future system on Electron-Ion Collider (EIC)
  - AI-based electron tagging with streaming readout to identify the (non)interesting Deep-Inelastic-Scattering (DIS) processes in the e+p/A collisions.
    - based on the measured scattering electron energy and direction

The ML algorithm – TrackGNN

ECML PKDD 2022, Sub 1256

- Based on Graph Neural Network (GNN)
  - Detector and physics knowledge improves prediction
  - Based on PyTorch and PyTorch Geometric
- Initial training on simulated data from MVTX and INTT
  - On GPU - NVIDIA Titan RTX, A500, and A6000
- Topological selection of HF signals on FPGA
  - Tracking and clustering must be done on FPGA
- Beam-spot and anomaly detection on GPU based feed-back system

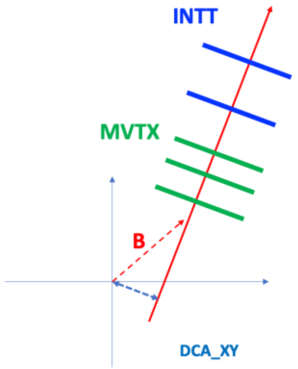


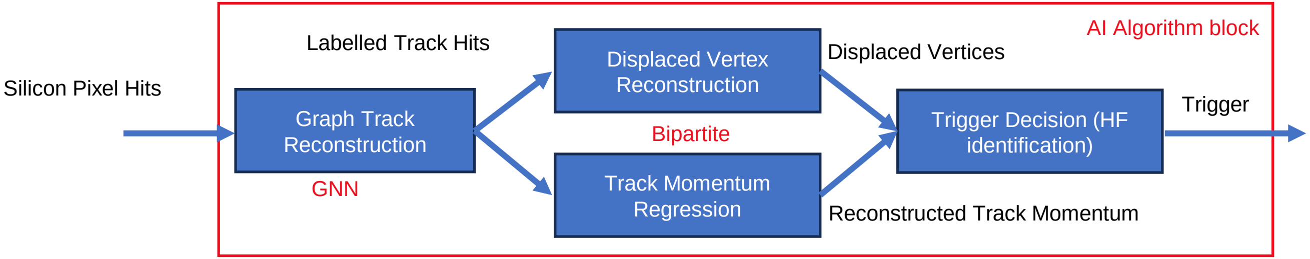
- We propose a novel method to treat the events as track graphs instead of hit graphs. This method is driven by the physics (transverse momentum)
  - Estimate momentum based on silicon hits -> 15% improvement on trigger decision

The ML algorithm – TrackGNN

ECML PKDD 2022, Sub 1256

- Three stages of event processing
  1. Hits clustering
  2. Track reconstruction + outlier hits removal
     - by connecting the hits across different detector layers into hit pairs.
     - apply geometric constraints and down select the hit
  3. Trigger detection
- Graph Neural Network to solve
  - the track reconstruction problem
  - the inter-track adjacency matrix prediction
  - the graph level trigger detection





The ML algorithm – TrackGNN

ECML PKDD 2022, Sub 1256

- Challenges
  - To provide an end-to-end solution that uses raw detector readout hit information to make trigger decisions for data collection.
  - To design a neural network compatible with the given detector readout and capable of learning a broad spectrum of physics properties
- using low-level hits to build the high-level trigger decision.
- Growing sub-field of geometric deep learning

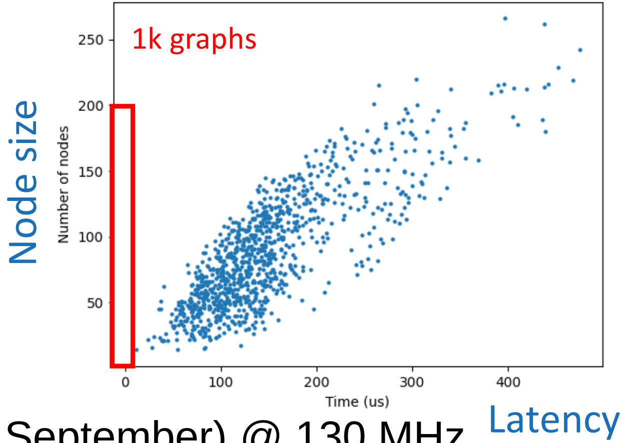
$D^0 \to K^- \pi^+$ sample

1% signal/background ratio			0.1% signal/background ratio		
Background Rejection	Efficiency	Purity	Background Rejection	Efficiency	Purity
90%	72.5%	7.25%	90%	78%	0.78%
95%	48.9%	9.78%	95%	50%	1.0%
99%	15.0%	15.0%	99%	17%	1.7%
99.33%	10.5%	15.74%	99.33%	11.0%	1.65%

Generation of the GNN IP core – two parallel efforts

1. Team lead by the Georgia Institute of Technology (GIT)
   - Direct translation of the sPHENIX TrackGNN model to IP using HLS
   - Model
     - 5 layers, each layer: 64 dim 4 layers for node and 64 dim 4 layers for edge embedding
   - Goal: 100-200 nodes, 200-500 edges
   - Implementation
     - 100 nodes, 140 edges
     - Measured Start-to-end latency
       - 150 us @ 130 MHz, 130 us @ 180 MHz
     - Still needs 10-20x speedup!

Utilization (Alveo U280)	
LUT	308K (23.7%)
FF	378K (14.5%)
BRAM	1025 (50.8%)
DSP	1426 (15.8%)



   - Fast-paced development 380 us (25th August) -> 150 us (4th September) @ 130 MHz
     - Attempts to increase clock to 300 MHz failed on timing constrains
     - Detailed latency breakdown and parallelism exploration ongoing
     - Might require model changes

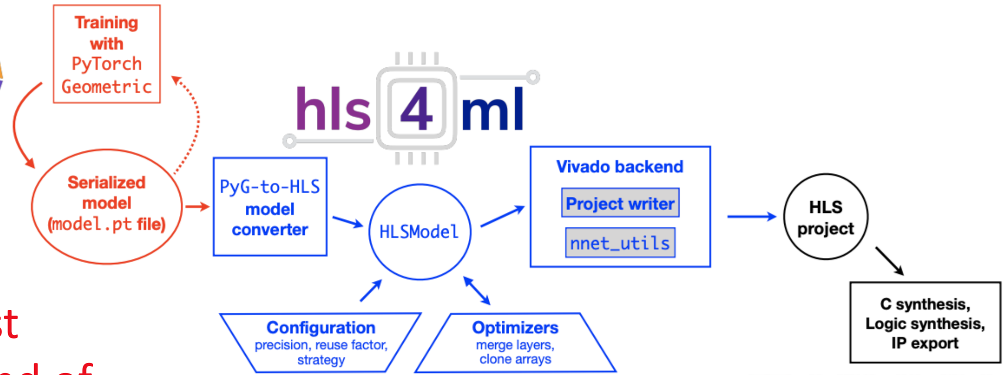
Close discussion between model developers and FPGA engineers

Generation of the GNN IP core – two parallel efforts

arXiv:2112.02048
arXiv:2103.05579

2. Team lead by the Massachusetts Institute of Technology (MIT) and Fermilab (FNAL)
   - Based on High Level Synthesis for Machine Learning (hls4ml), a generalized python framework for machine learning inference in FPGAs

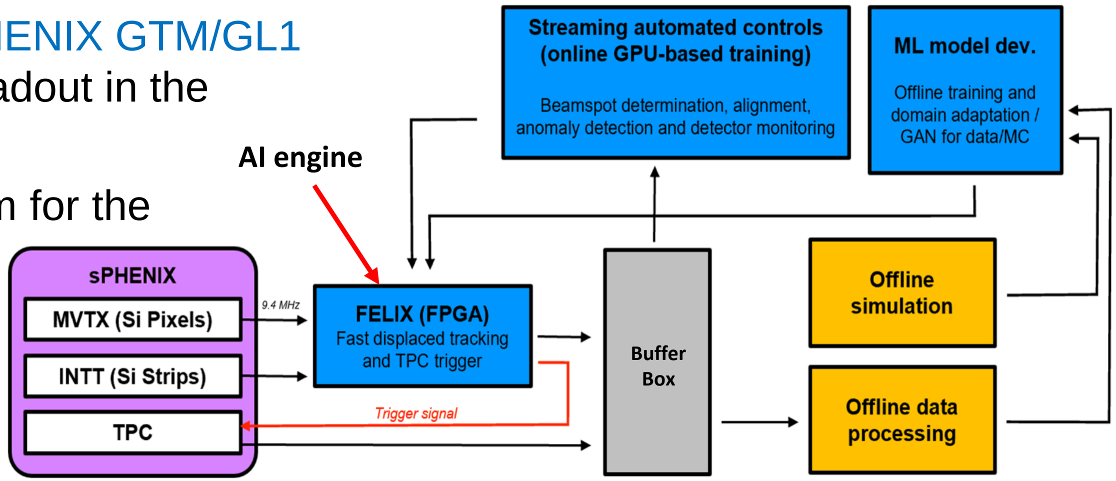
- Third main upgrade underway, focusing on 3 examples
  - Example 1: Tri-muon reconstruction with the LHC (muon endcaps)
  - Example 2: Heavy flavor tracking at sPHENIX
  - Example 3: Silicon strip tracking at LHC



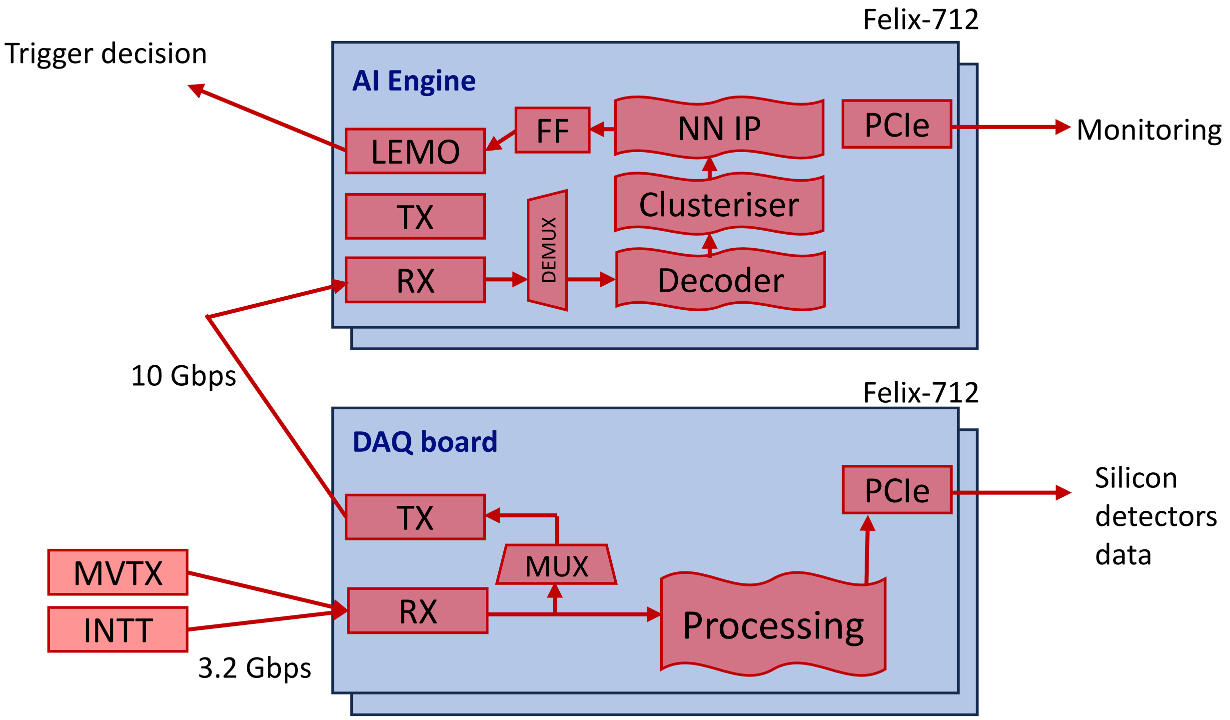
Initial translation just started, expected first version of the TrackGNN model on FPGA end of October 2023

The DAQ–AI Data Flow

- Motivation to use FELIX board:
  - To reuse the PCIe implementation (16-lane Gen-3) and software tools provided by the FELIX developers
  - on-board FPGA is a Kintex Ultrascale XCKU115FLVF1924-2E
- The decision signal of heavy flavor event from the AI-Engine will be sent out via the LEMO connectors to the sPHENIX GTM/GL1 system to initiate the TPC readout in the triggered mode
- GPU based feed-back system for the beamspot monitoring



The firmware design - data flow



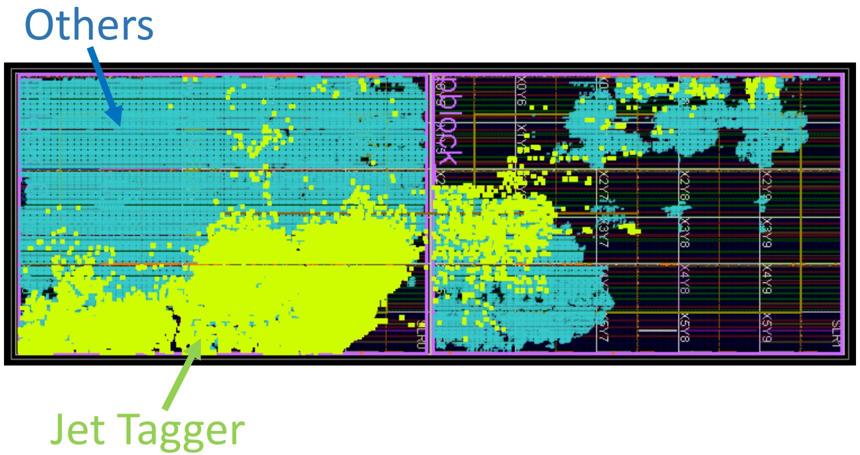
- MVTX 144 links @ 3.2 Gbps and INTT raw data stream will feed two AI engines (one for each hemisphere)
  - 24 links for MVTX and 24 links for INTT per AI engine
  - 8b10b protocol with links driven @ 10Gbps
    - tested up to 14 Gbps, with external loopback measurement at FELIX with BER < 10^{-16}
- Raw MVTX and INTT data packets:
  - 1 MVTX packet @5 us strobe
    - ~10 pp collisions (MB events) @2MHz pp collisions
  - 50 INTT packets @ 100 ns strobe
- Data needs to be decoded, clustered, time aligned and feed the neural network IP

Very challenging project to fit in the FPGA resources!

The PCIe utilization

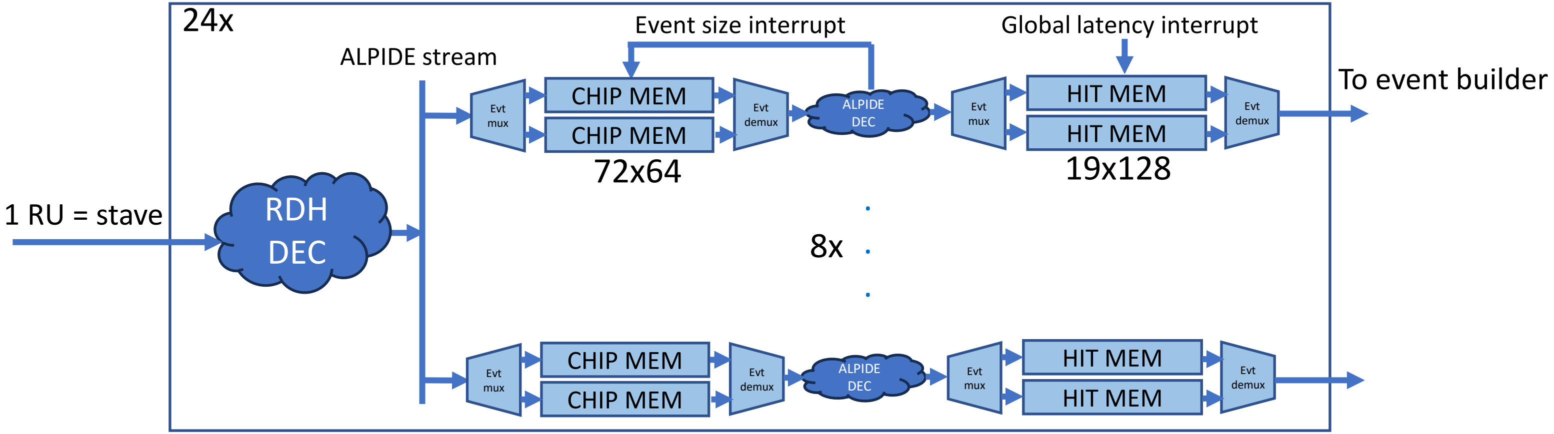
- Initial implementation at FELIX-711 (rm-4.11) by FNAL group
- Aim to use FELIX FW implementation of PCIe and its Software tools
- We use this standard well-understood benchmark model “Jet Tagger” (arXiv:1804.06913) to test the workflow
  - QKeras and converted to hls4ml to create an IP
  - 16 inputs (expert variables) and 3 dense hidden layers with 64, then 32, then 32 neurons
- Current efforts to extract and only use the Wupper module (PCIe) to lighten to logic and keep more resources for the AI IP code

Post-implementation utilization (FELIX-711)			
	FELIX-711	PCI (Wupper)	Jet Tagger
LUT	241K (36.3%)	28K (4.26%)	83K (12.5%)
FF	310K (23.41%)	76K (5.75%)	50K (3.76%)
BRAM	635 (29.4%)	91 (4.22%)	195 (0.09%)
DSP	72 (1.3%)	0 (0%)	72 (1.3%)



MVTX decoder

- Initial implementation of the FPGA-based MVTX decoder
- Max 128 hits per chip stored (expected physics ~50, issues with beam background?)
  - Maximum latency 532 ns @ 240 MHz
- The MVTX data latency depends on the actual collision time and hit occupancy
  - To provide a fixed latency to the GTM a BC information from INTT is used
  - An interrupts to event size/processing time are in place not so exceed the maximum latency
    - Separate memory per MVTX event to fast clear the data



The latency constrains for the TrackGNN

- The TPC buffers can hold up to 30 μs of data
  - The goal of this project is to aim for 10 μs collision-trigger latency to capture the TPC stream
- The Calorimeter buffers can hold up to 6.4 μs of data
  - Can we improve the latency down to 5 μs to also capture the calorimeter stream?

- The latency breakdown

1. MVTX readout window 5 μs – not fixed interaction-readout latency!
2. IR -> Counting house ~0.3 μs (81 m fibres)
3. FELIX -> AI data forward, decoder buffers ~ 0.6 μs (@240 MHz)
4. Clusterizer + tracking + Trigger decision (currently 130 μs for TrackGNN model!)
5. AI -> GTM -> TPC FELIX (negligible, all three sits in Counting house)

Summary

- The TrackGNN model has been developed and tested on HF event simulation for sPHENIX
  - provides good precision while analyzing two hemispheres independently
- IP core generation by two teams
  - Huge progress and improvement of the utilization and latency
  - Might need to reassess the model used to fit within FPGA resources and latency
- FLX-712 boards to serve as AI engine installed in sPHENIX counting house
  - Final push to finalize development of each FPGA component and placing them together
- Design and test the feed-back system
- A new FLX-182 board arrived to BNL which will be the base for EIC development
  - The backup plan to use it for the sPHENIX TrackGNN model (the FPGA is 3x bigger)
  - Probe the possibility of using off-the-shelf card (Alveo etc.)

Thank you for your attention

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.

The timeline

- July 2023
  - FELIX-712 and FELIX-182 setups installed at sPHENIX Counting house
- October 2023
  - TrackGNN IP core should be optimized and Implemented
  - Discussing between physicist, model developer, and FPGA engineers to meet the physics goals and constrains of the triggering system
- November 2023
  - Cosmic stream from the MVTX sent to the AI engine – tuning of the decoder parameters
- December 2023
  - Cosmic stream from the INTT sent to the AI engine – tuning the alignment and event builder
- January 2024
  - First pp beam at RHIC, final adjustment of the AI engine, performance studies

Motivation – The challenges

- Real-time selection of rare decays of HF particles
  - requires continuous monitoring and adjustment of the
    - beam trajectory ("beam spot") – in time periods of seconds to hours, the position and shape can change (this will affect the HF the topology)
    - detector alignment, conditions and anomalies
- Adapt AI to continuous learning and changing conditions -> adaptive learning
  - Development of real-time autonomous closed loop adaptive learning system

Predicted timeline

2021
- Project funded by DOE FOA
- Initial simulations constructed
- First data for algorithm training

2022
- MVTX & INTT SRO
- Fast tracking & trigger algorithms in place
- Initial FPGA bitstream synthesis
- GPU feedback machine R&D

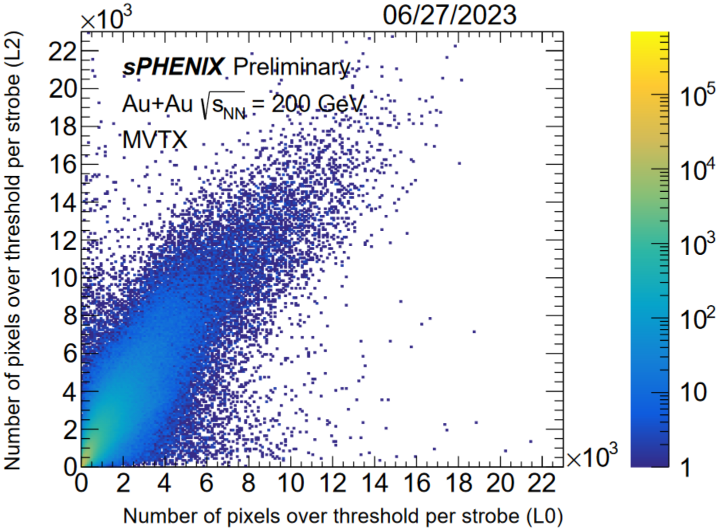
2023
- Refine interface between system and detectors
- Improve algorithms with latest data stream and commissioning info
- Pre-commissioning

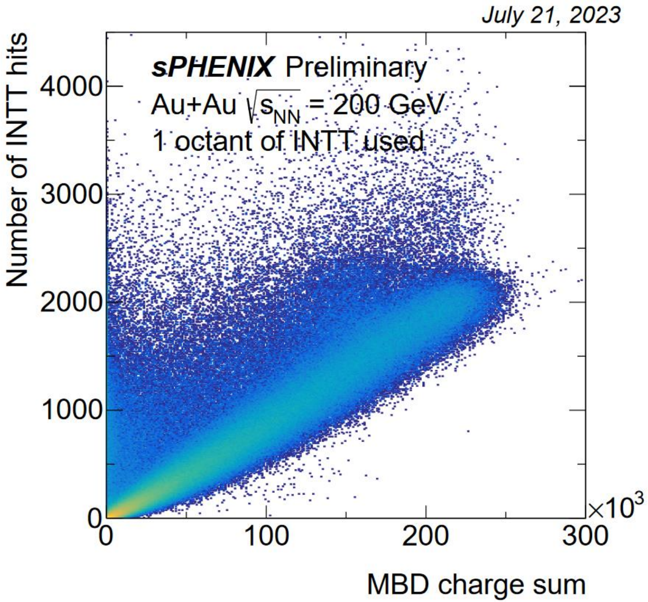
We are here!

2024
- Deploy device at sPHENIX pp/pA run
- EIC preliminary TDR (CD2)

Later
- Final design for EIC TDR (CD3)
- Take advantage of new technology if required
- Deploy device at EIC

MVTX and INTT commissioning performance





- Timing in detectors on good track

From sPHENIX to ePIC: Streaming + AI/ML DAQ

from Jo's talk at ePIC collab. mtg