Intelligent Experiments through Real-time AI for EIC

- *Fast Data Processing and Autonomous Detector Control for High-Energy Nuclear Experiments (Fast-ML)*

Ming Liu

Los Alamos National Lab

ePIC AI Townhall

07/16/2026

From Markus

During Each Flash Talk, Consider

- What problem is the project addressing?
- How does it connect to ePIC workflows?
- What data, software, and infrastructure does it use?
- What is its level of integration in the ePIC software stack?
- What support or collaboration could help move it forward?

Broader Goal: Identify opportunities to:

- Integrate AI tools into the ePIC software stack and simulation campaigns.
- Share tools, data, expertise, and infrastructure.
- Build connections across related projects.

What & Why: SRO, Lessons Learned and Plan for EIC

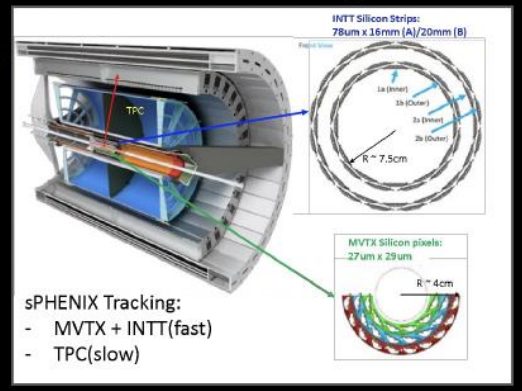
Cristiano Fanelli @Tue

AI applications for real time event processing

M. Liu, Intelligent Experiments through real time AI

ML on FPGA

- Fast Data Processing and Autonomous Detector Control for sPHENIX and Future EIC Detectors



Identify D/B hadrons with real-time ML

- Topology of D/B decays
- Monitor collision vertex
- Feedback for improvement

The challenges:

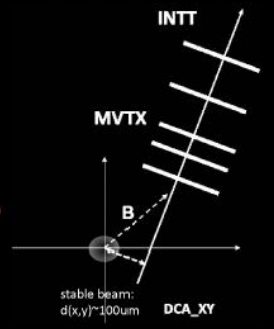
Very high p+p collision rate: ~3MHz

Low rate of rare signals: ~150Hz (beauty for eg)

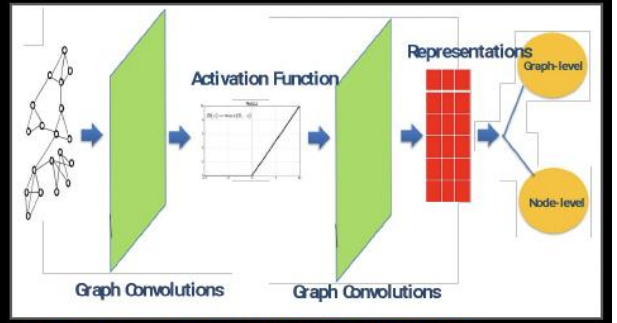
Limited DAQ trigger bandwidth: ~15 kHz

(or 0.5% of p+p collisions)

No effective conventional triggers available



hls 4 ml



J. Kvapil et al, JINST 19.02 (2024): C02066.

Priorities moving forward

Salvatore Fazio @Tue

Top most near term priority for the reminder of 2026

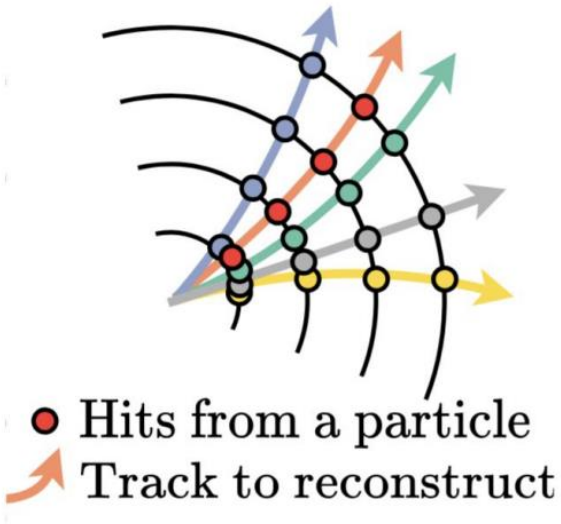
- Successfully land our 6+ **NIM-A papers**

Mid-term priorities

- Refine studies on **machine background effects** on physics
- Support base-lining of detector design with dedicated physics performance studies

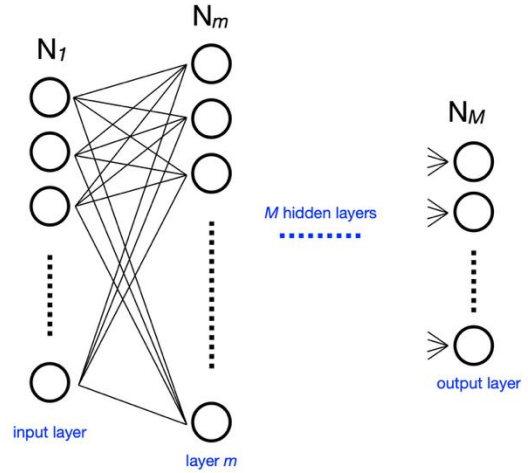
Why Fast-ML?

- High data throughput from modern detectors in high-energy experiments
  - O(1~10)TB/s @detectors, CMS, ATLAS, ALICE, **sPHENIX, EIC** …
  - Very large data volume (~100PB/year)
    - **for EIC, could be background dominant**



- **Our goal** - use AI/ML based algorithms at **hit level**:
  1) reject beam background to improve SRO
  2) tag important rare or calib/monitor events in real-time
  3) allow fast data filtering for rapid offline analyses

sPHENIX as the first test ground, ultimately for EIC in 2030s



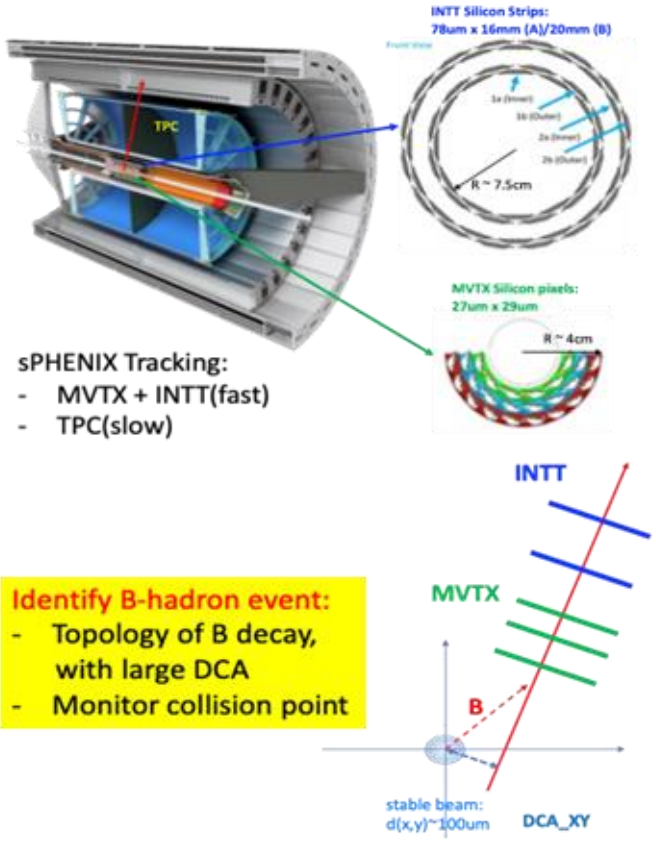
Real-time AI

Our Approach for sPHENIX SRO

- Heavy flavor event AI-trigger demonstrator

Two half-barrels

Selective streaming real-time AI and autonomous detector control:
Developed a demonstrator for p+p running - generalizable for applications in experiments at the EIC



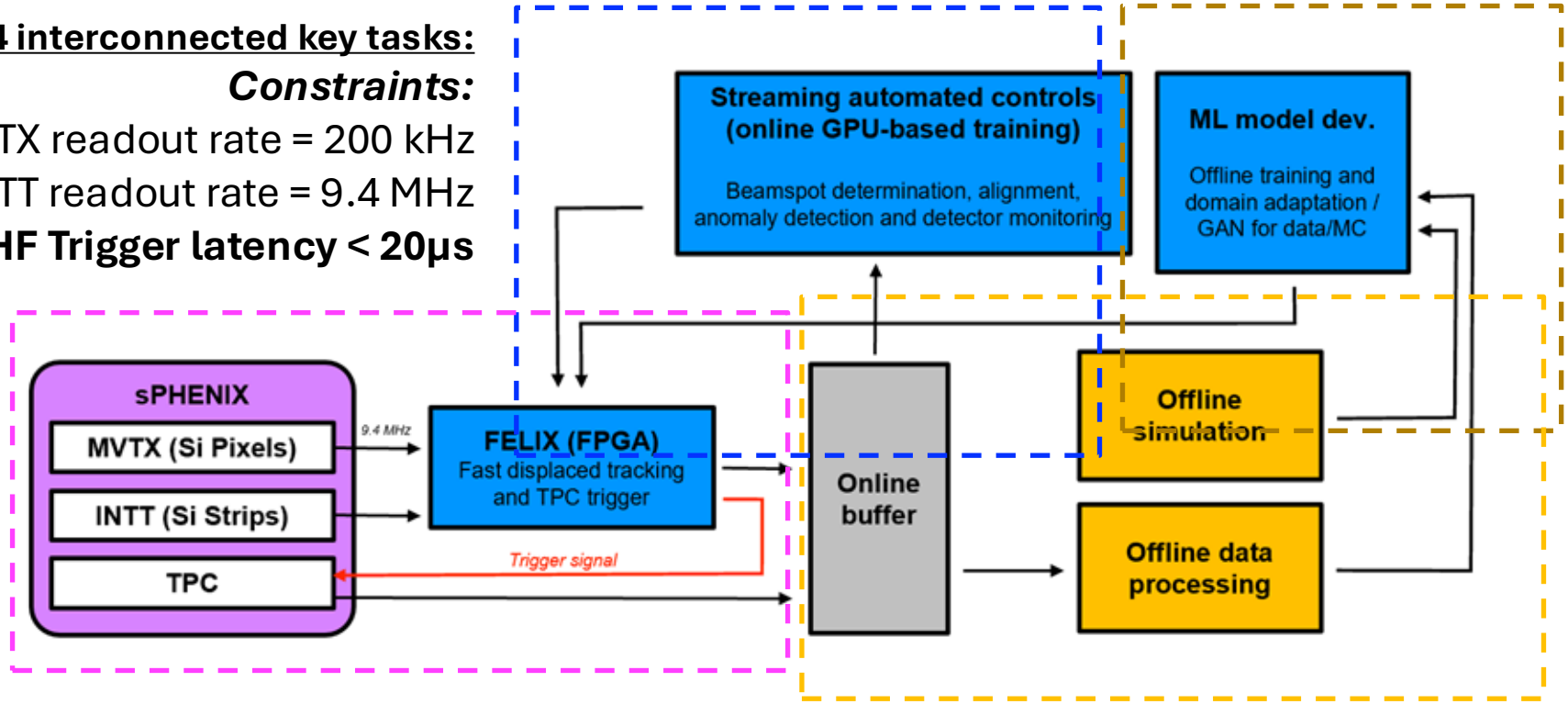
<u>4 interconnected key tasks:</u>

Constraints:

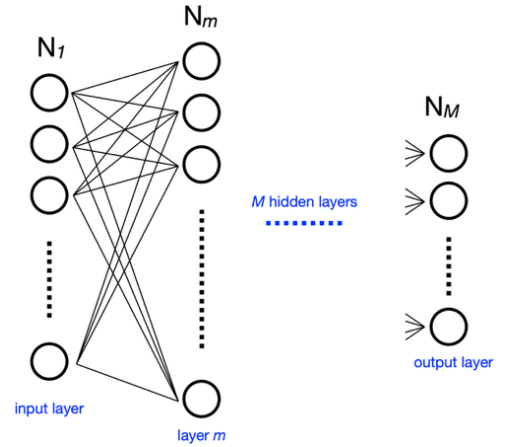
MVTX readout rate = 200 kHz

INTT readout rate = 9.4 MHz

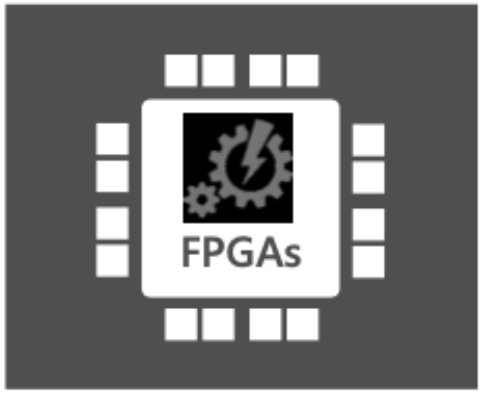
HF Trigger latency < 20μs



3 Major Areas of R&D



- ❑ Physics/detector simulations and AI/ML algorithm development
- ❑ Translate AI/ML algorithms into hardware language FPGA code with data processing latency and hardware resource constraints
- ❑ Deploy FPGA algorithms in a demonstrator system in sPHENIX



Good lessons learned from sPHENIX operation with real beam

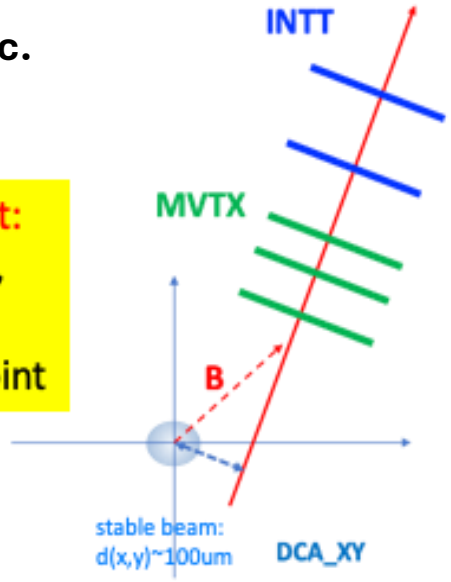
- **p+p, Au+Au in 2023-2025**
- **Next – EIC, early 2030s**

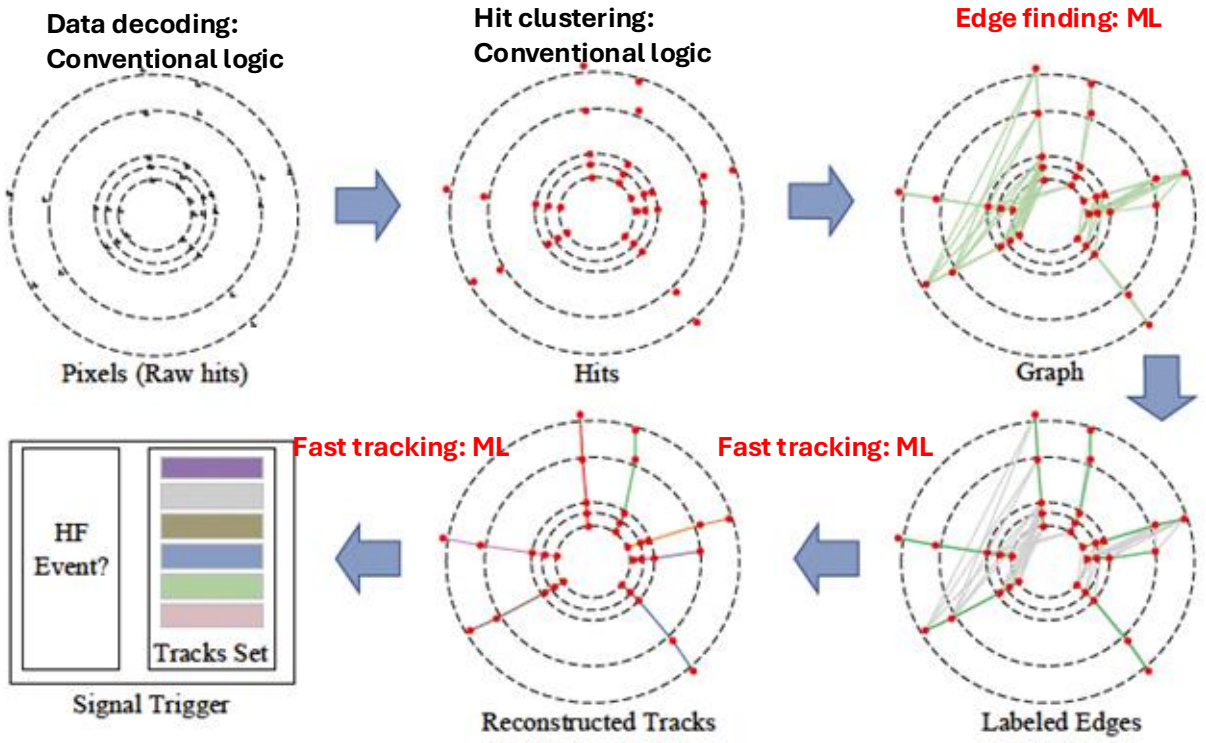
GNN based Real-Time HF Trigger on FPGA

- AI HF-Trigger algorithms not sensitive to small IP shift!

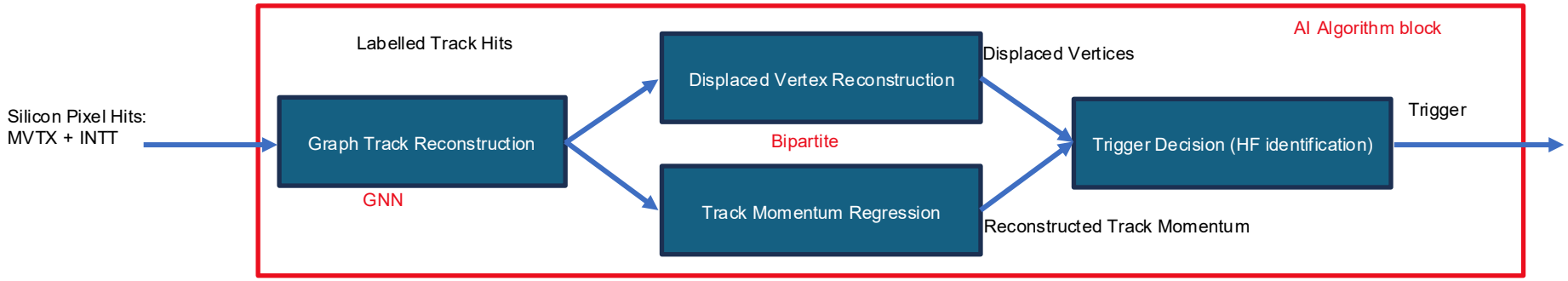
Features: DCA, pT etc.

Identify B-hadron event:
- Topology of B decay, with large DCA
- Monitor collision point



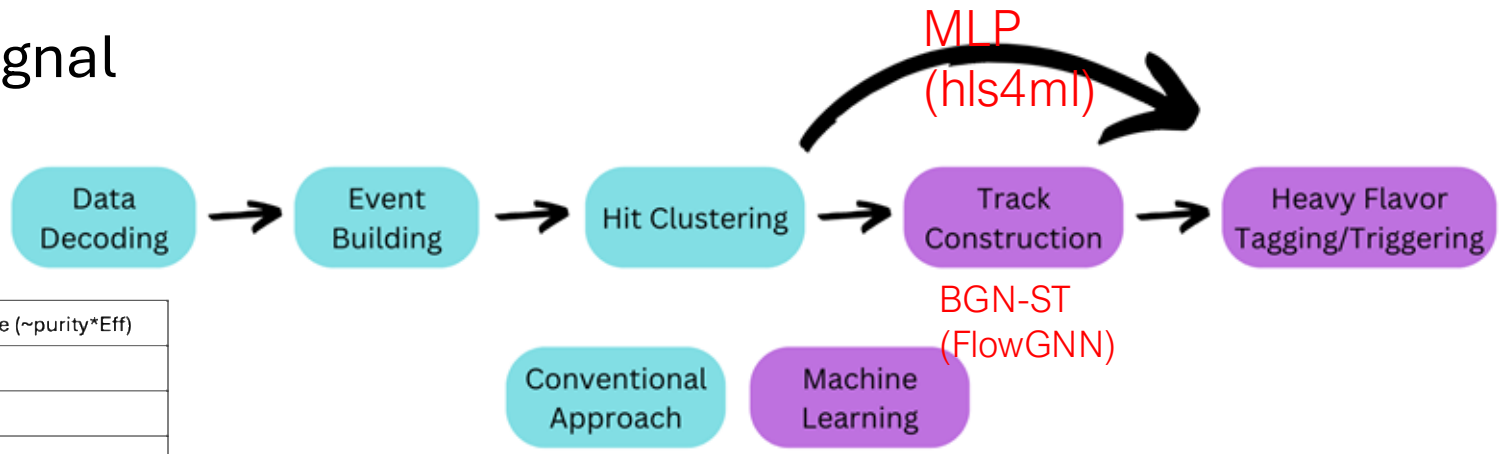


sPHENIX full simulation data



FPGA Ready Algorithm Summary

- Hit decoding and clustering - conventional algorithms
- *Event building, collect hits from the same collisions – MVTX(slow) + INTT(fast)*
- "Track" reconstruction - using GNNs in two parts
  - Edge candidate generation - connect clusters (nodes) with edges, with geometric constraints
  - Edge candidate classification - using graph convolutional network (GCN) (arXiv: 1609.02907)
  - Construct final "tracks"
- Use a least squares method to perform p_T prediction from track curvature
- Tagging of the heavy flavor signal

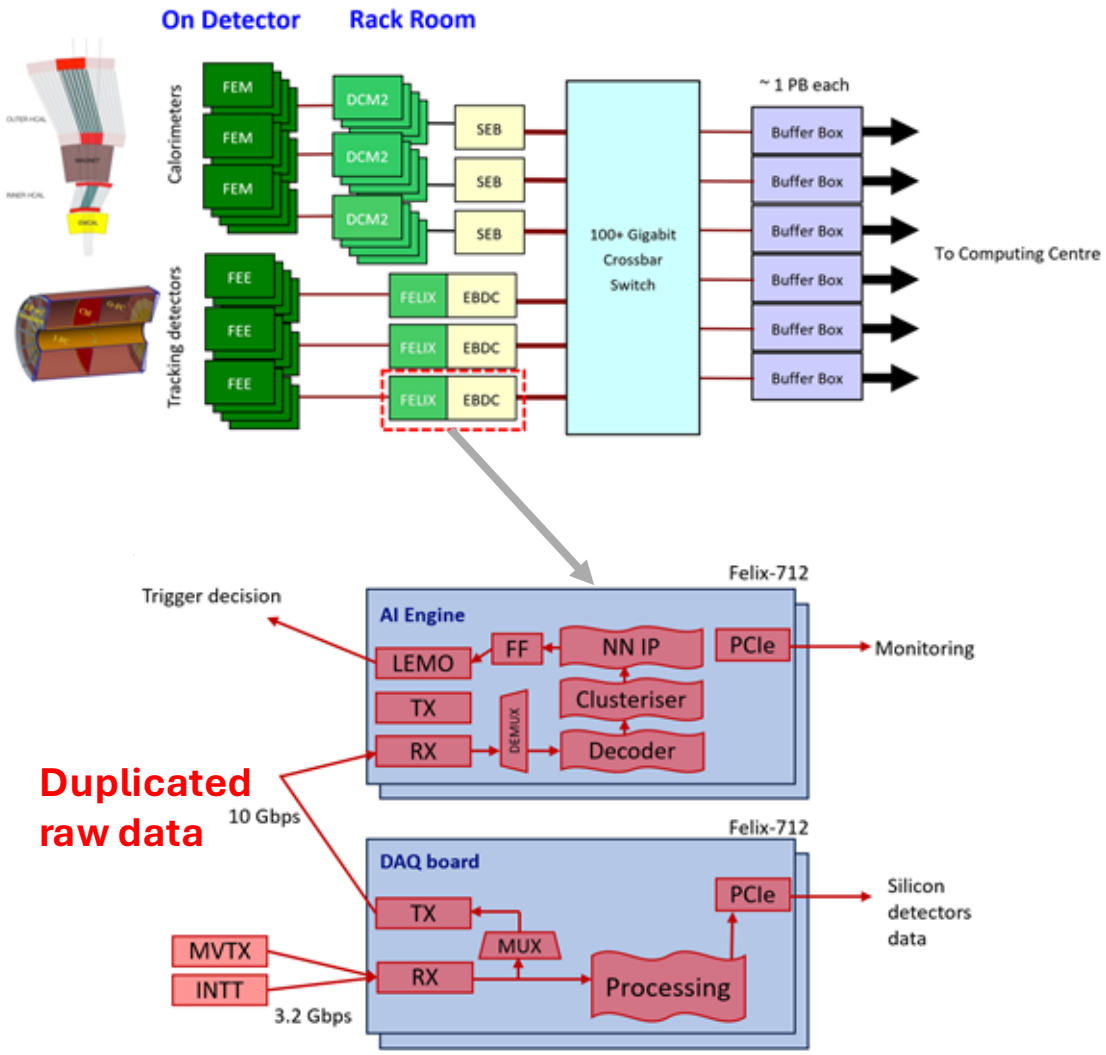


Model Configuration	Precision (Purity)	Recall (Efficiency)	F1-Score (~purity*Eff)
No Pileup	92%	90%	91%
No Pileup, FPGA-ready	79%	87%	83%
Pileup (20)	80%	73%	76%

Also an alternate implementation, taking the clusters directly without explicit track reconstruction.

Readout and HF AI-Trigger Implementation in FPGA

- The sPHENIX tracking detectors use FELIX-712 PCIe-based boards
  - Contain an AMD/Xilinx Kintex UltraScale FPGA (xcku115-flvf1924-2-e)
- To the readout DAQ boards, add AI Engine boards to perform the B-tagging using AI (FELIX-712)
- Exploring implement graph neural networks (GNNs) with two approaches:
  - FlowGNN (arXiv: 2204.13103)
  - hls4ml (arXiv: 1804.06913)

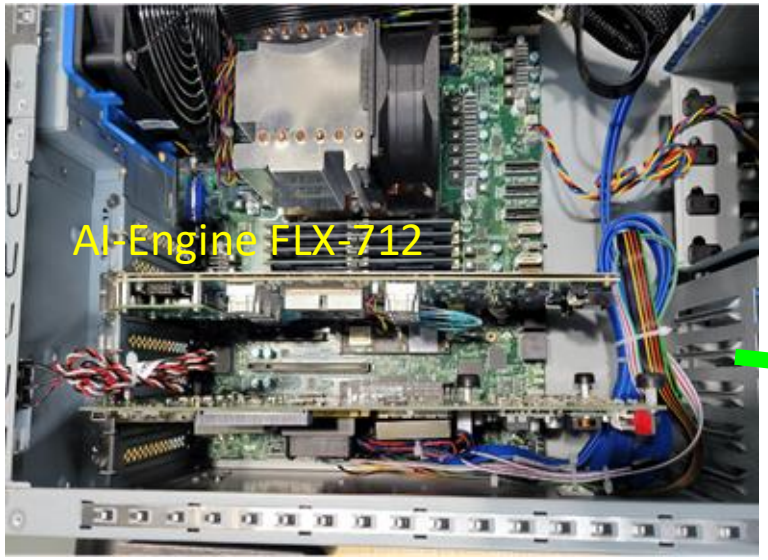


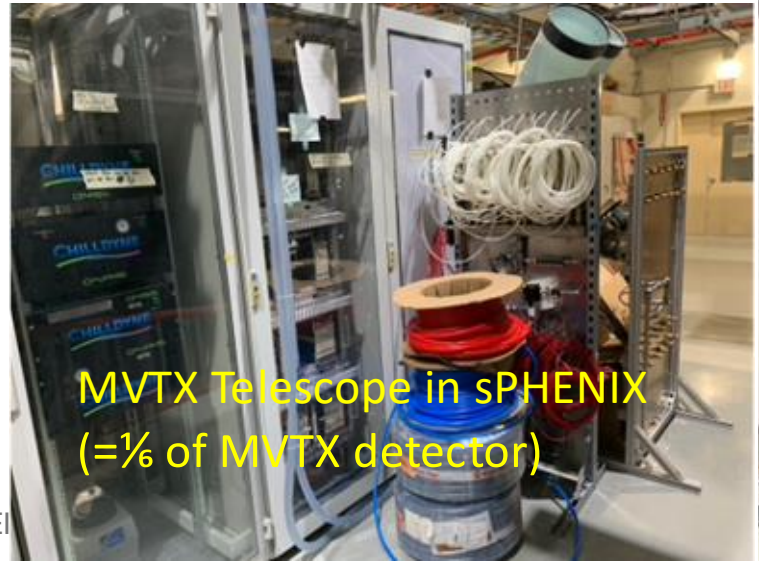
A Small Scale Demonstrator:

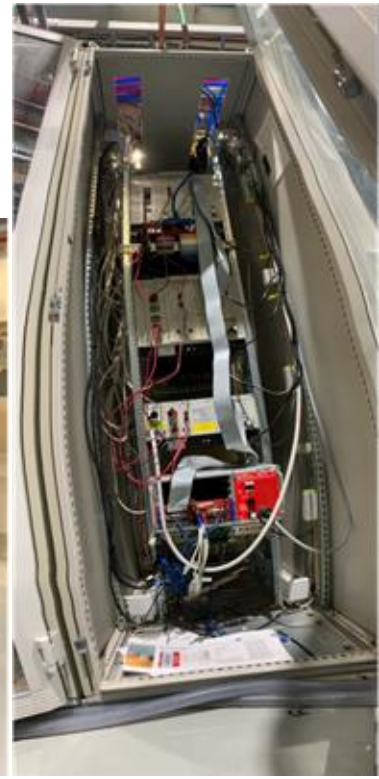
- MVTX Telescope Communication

Due to very tight sPHENIX operation schedule, we didn't get the opportunity to integrate AI/ML system into the sPHENIX DAQ, instead, used MVTX telescope in the sPHENIX counting house for the system test

- ❑ FELIX-712 was designed as sPHENIX readout board, the PCIe is used to receive data from the optics
  - ➢ We use this to save the timing (Bunch Crossing ID) and trigger decision from the AI
  - ➢ We have configured the PCIe uplink (normally used just for configuration) to load real detector data to the board, in order to have a controlled validation environment
- ❑ Successfully received and decoded data from single stave of the MVTX 8-stave telescope (MVTX = 6 x Telescope)
- ❑ Added ILA via Xilinx virtual cable for additional debugging and monitoring







EIC - Prepare for Unexpected

- lessons learned from sPHENIX data taking and implications for future EIC and other experiments

New ideas being developed to address new challenges ...

Unexpected Challenge First Observed in sPHENIX 2023 Au+Au Runs

- Challenge of full streaming readout in high beam backgrounds!

- ❑ **NO problem in p+p collisions**
- ❑ **Major beam-related background with Au beam**
  - ➢ Events with just one bunch of beam
  - ➢ Related to beam halo induced particles hitting large number of sensor pixels in the MVTX detector sensors

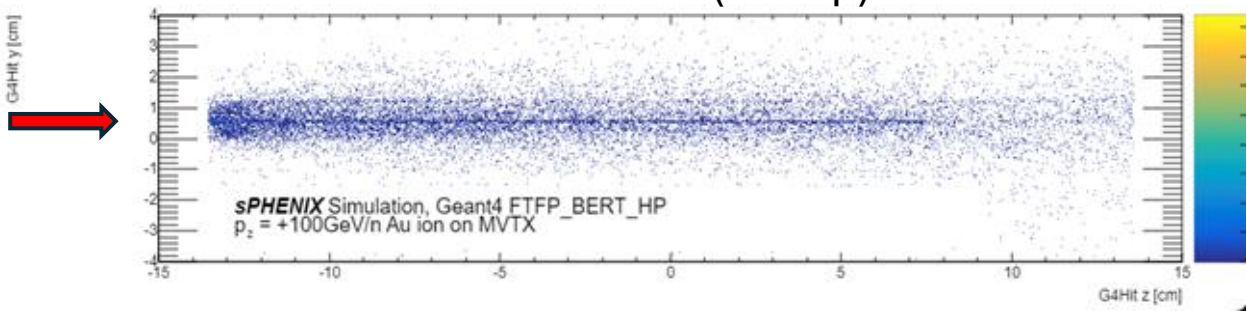
Date >> DAQ bandwidth! (>10^3)

EIC: phase-1, e+A program

- ❑ Could face similar high backgrounds with ion beam

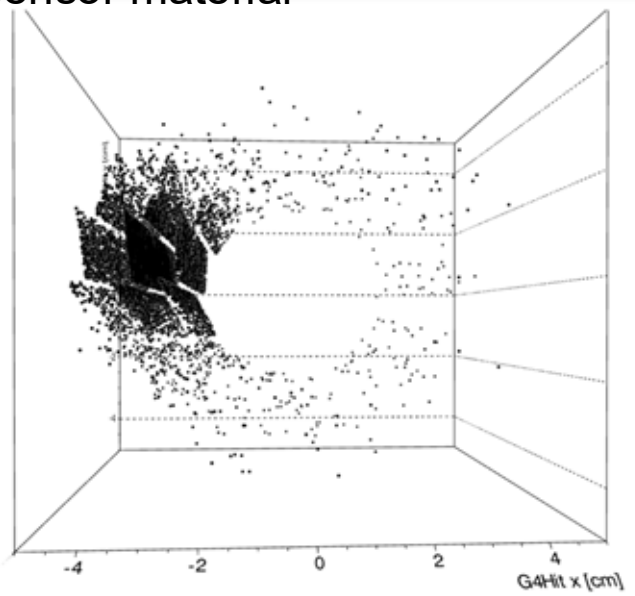
Smart data management highly desired on/near the detectors for full streaming readout in high background environment

One silicon stave (9-chip)



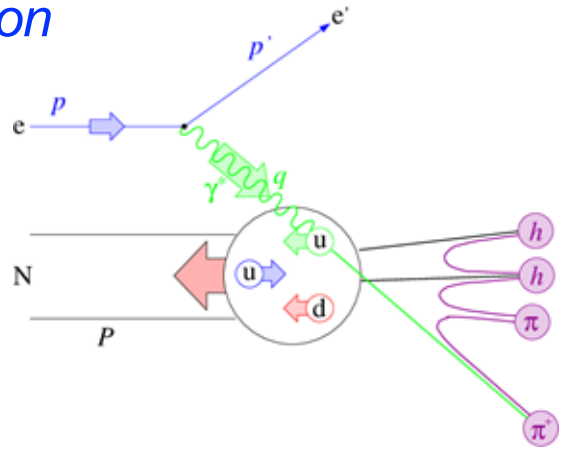
GEANT Simulation:

Single 100 GeV Au ion striking the end of the 50um thick MVTX silicon sensor material



Proposal: Fast-ML for EIC

- DIS-electron identification in real-time with beam background suppression

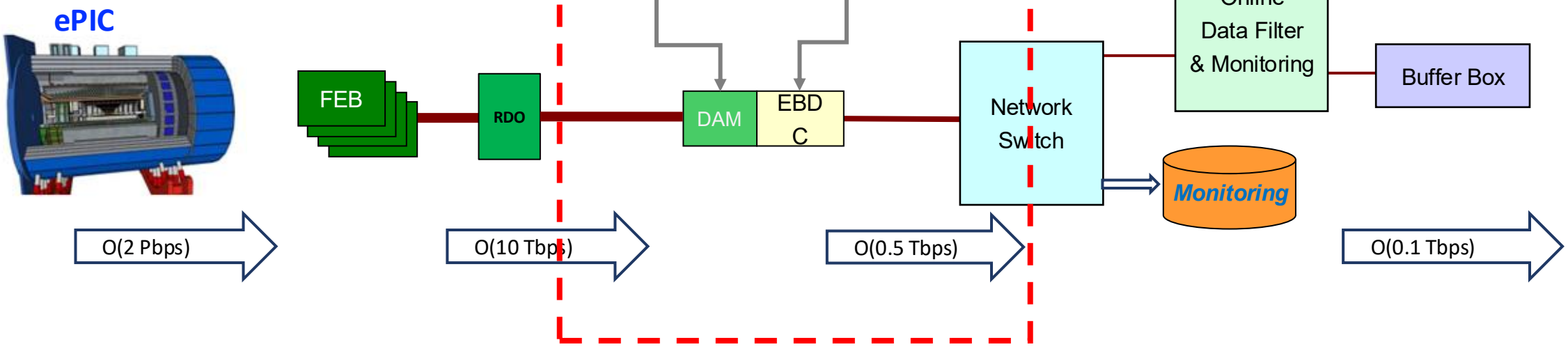


Selective streaming readout for AI-Engine:

- ❑ tag DIS-electron to define DIS event ID
  - ➢ EMCal + Trker + ePID
  - ➢ DCA~0
- ❑ With AI noise suppression on chip (AI-on-Sensor)!

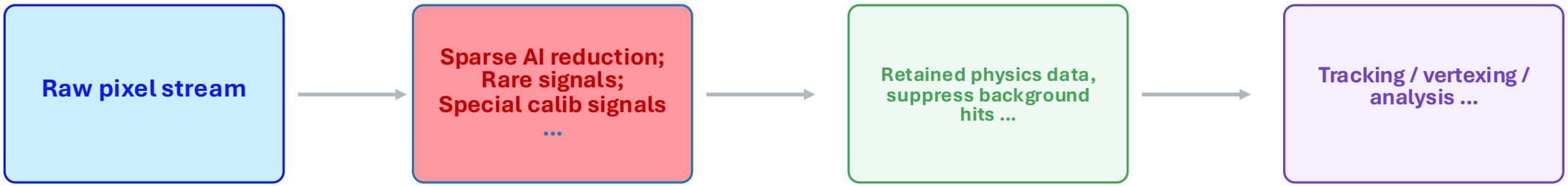
Agentic AI:
- active beam background rejection;
- rare/calib signals (J/Psi, D^0,SVT etc.)

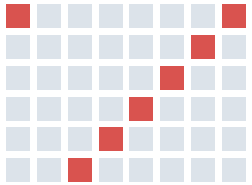
SRO + AI/ML Fast Data Processing:
- DIS e-tagger: event ID
+ other rare process, HF-tagger etc. ...



AI-native detector intelligence at the sensor-data level

Agentic AI tools





Act early: before backgrounds propagate through the full data chain.

Preserve what matters: signal hits, local topology, timing, and uncertainty-aware summaries.

> This is **not** “ML after reconstruction.” It is a physics-preserving data-reduction layer that becomes part of the detector data path.

Source: proposal main text, pp. 1–7

Could use more helps … and welcome to join us!

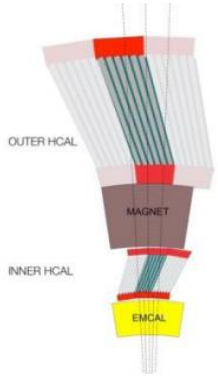
- **Time-structured beam related backgrounds for SVT (and other subsystems)**
  1. Synchrotron radiations from e-beam
  2. Heavy ion beam halo … sPHENIX Au-beam like background

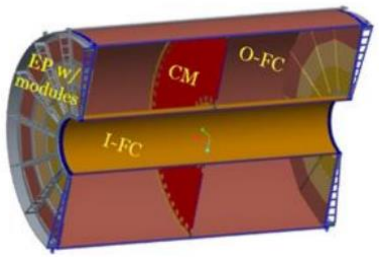
- **DAQ integration related:**
  - Maximum latency allowed for SRO, subsystem dependent?

- **Offline filtering/small data stream for fast calibration/monitoring analyses**
  - Detector performance, pi^0
  - Physics signals, J/Psi, D^0 etc.
  - Selected DIS events …

Backup slides for discussions

sPHENIX Readout and Trigger Distribution





On Detector

Rack Room

Front-End Module/Electronics

Data Collection Module

oHCal

iHCal

EMCal

FEM

FEM

FEM

DCM2

DCM2

DCM2

SEB

SEB

SEB

TPC

INTT

MVTX

FEE

FEE

FEE

FELIX

EBDC

FELIX

EBDC

FELIX

EBDC

100+ Gigabit Crossbar Switch

Buffer Box

Buffer Box

Buffer Box

Buffer Box

Buffer Box

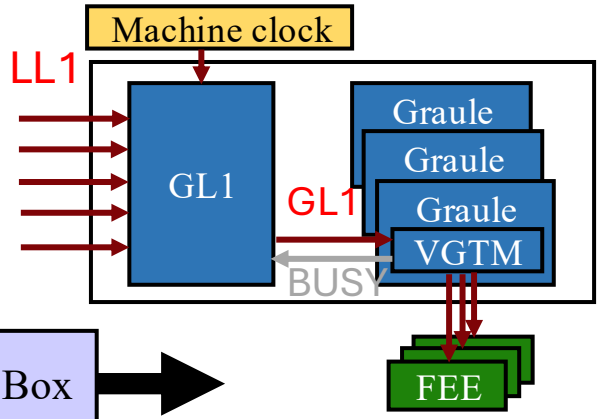
Buffer Box

~1 PB each

To Computing Center

Tigger inputs

- 8 LEMO, 56 fiber
- EMC, iHCal, oHCal, MBD,sEPD, ZDC



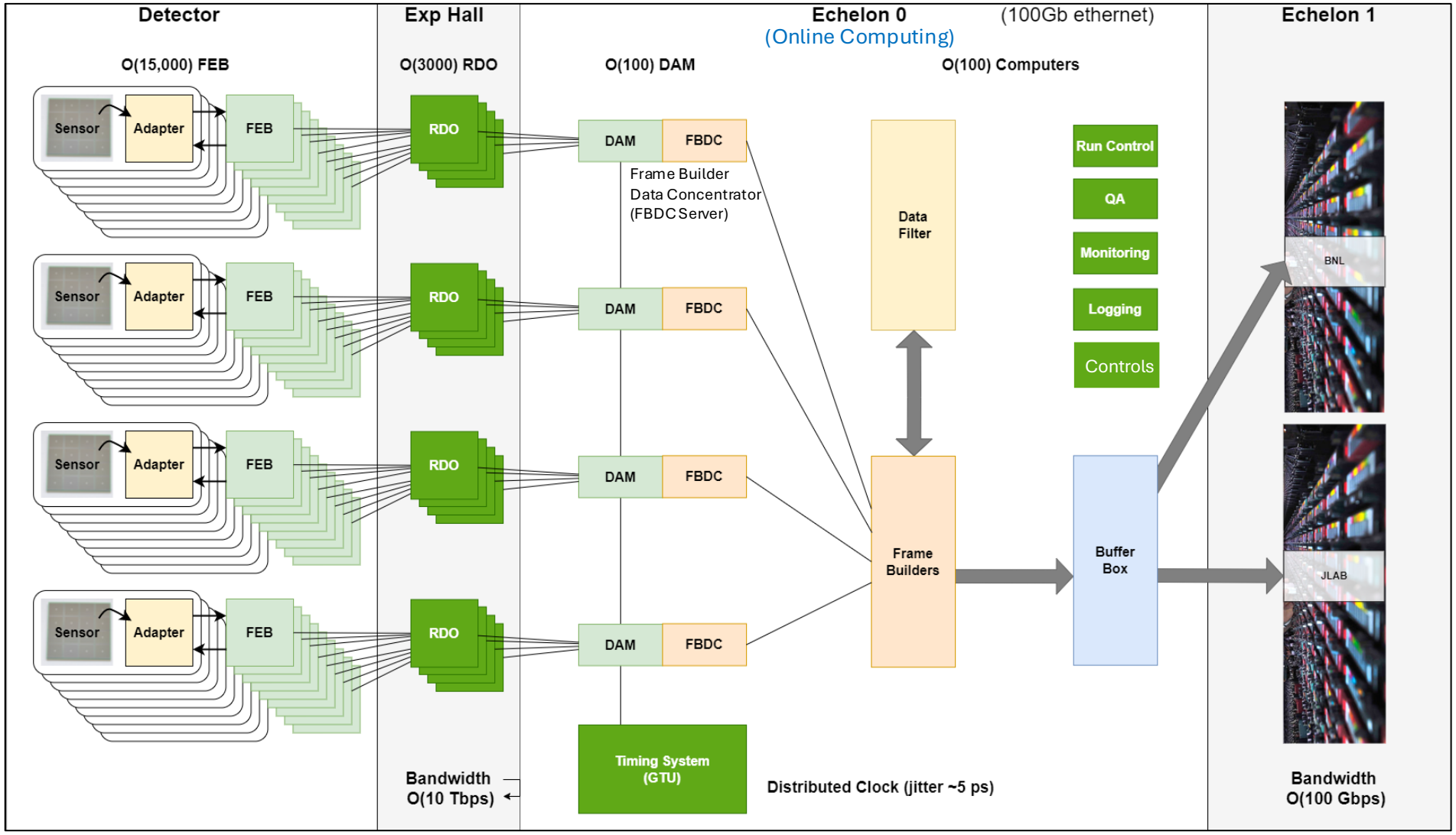
SRO(MVTX+INTT) -> ML -> HF Trigger

ePIC Streaming DAQ Architecture

EIC Specs

- Bunch Crossing
  ~10.2 ns (98.5 MHz)
- Interaction Rate
  ~ 2 us (500 kHz)
  (one interaction per ~200 bunch crossings)
- High Luminosity
  -- up to 10^{34} $cm^{-2}s^{-1}$
  -- Low occupancy physics signal
  -- Background and noise significant for some sub-systems

A big unknown: beam backgrounds, could easily overwhelm the DAQ system!



Objectives: establish the minimum AI elements needed

1

Realistic EIC benchmarks

Benchmark datasets, baselines, and evaluation protocols for inner silicon pixel streams under synchrotron and beam-related backgrounds.

“realistic beam background with time structures”

2

Physics-preserving sparse models

Deployment-aware hit-level and cross-layer models that suppress/compress background while preserving tracking, vertexing, and physics utility.

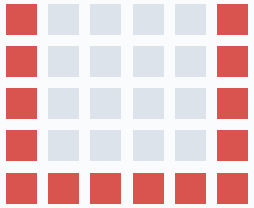
3

Bounded agent-assisted workflow

Constrained optimization, quantization/compression sweeps, synthesis-report parsing, and physics-aware regression testing.

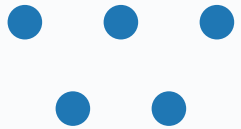
Hierarchical sparse model from chips to physics

Chip/tile hit sets



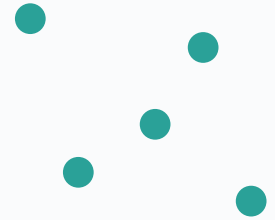
variable-size local hits

Sparse encoders



set transformer / sparse encoder

Cross-layer graph



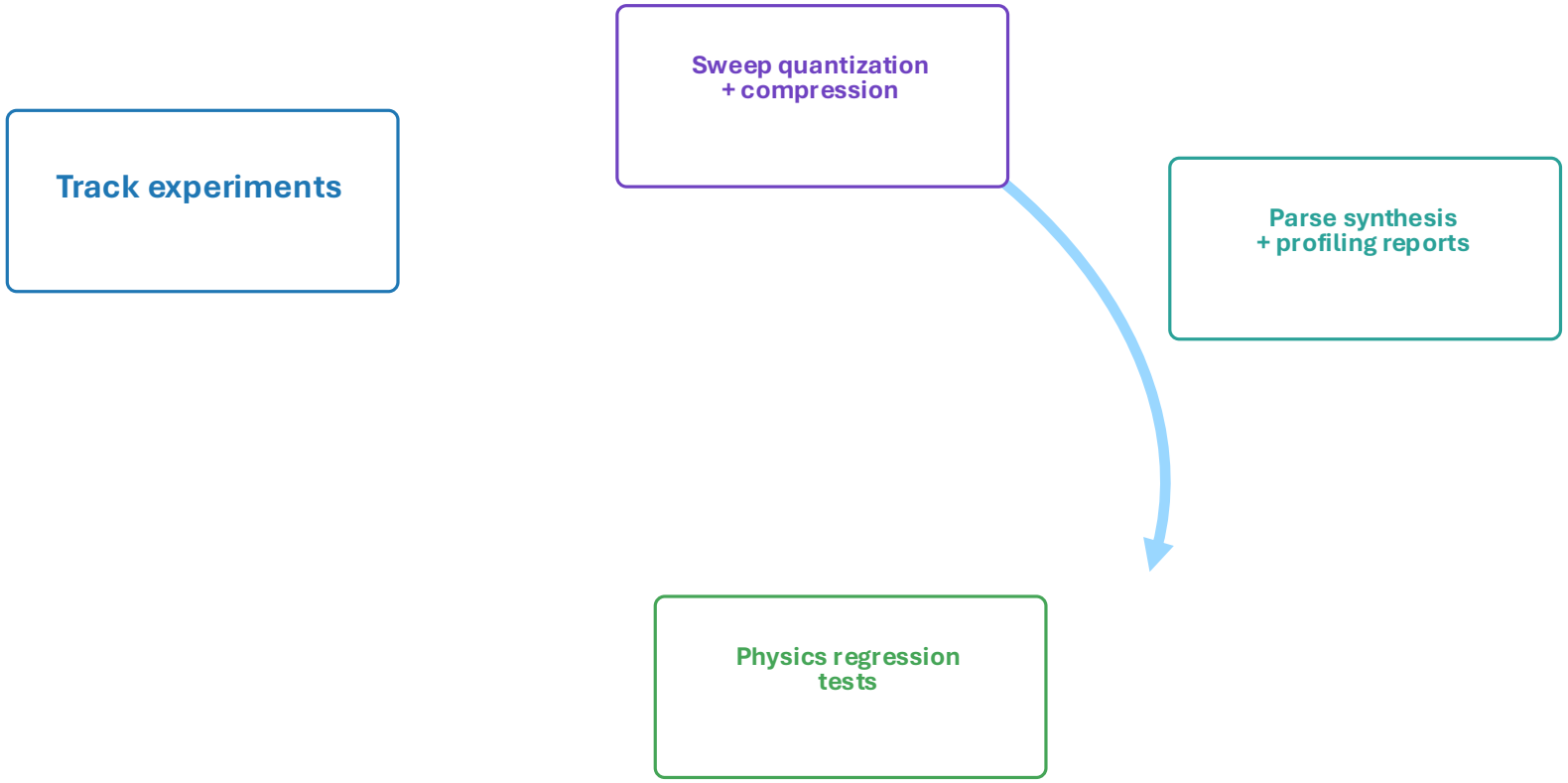
geometrically viable chip pairs

Outputs to DAQ + analysis

- Hit-retention scores
- Learned summaries
- Sparse graph objects

> Model objective: suppress/compress background-dominated pixel data while retaining information needed for downstream tracking, vertexing, and physics analysis.

Bounded agentic loop for faster deployment optimization



Human-in-the-loop constraints

Predeclared physics retention, latency, and resource thresholds reject bad candidate models automatically.

Deployment path

GPU, FPGA, and hybrid options are screened early; one primary path is selected by Month 3 for the integrated demonstration.

Trigger Performance Metric

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Edge candidates are created from hits using geometric criteria
  - Geometric criteria produces roughly O(n) hits, even with pileup date (usually ~2x as many edge candidates as there are hits)
- GNN classifies edge candidates based on hits information
- GNN also performs tracking de-pileup using fast INTT hits
- GNN trained to prioritize preserving edge candidates arising from trigger particles

An efficient, low-parameter counts, FPGA-ready effective tracking algorithm

Model Configuration	Precision (Purity)	Recall (Efficiency)	F1-Score (~purity*Eff)
No Pileup	92%	90%	91%
No Pileup, FPGA-ready	79%	87%	83%
Pileup (20)	80%	73%	76%

The Latency Constrains for ML-base Algorithm

- ❑ The TPC buffer can hold up ~30 us of data before receiving a readout trigger
- ❑ Detector readout delay, fiber transmission delay, data encoding/decoding
  - ➢ MVTX readout window ~8us
  - ➢ Interaction Region (IR) ->Counting house ~0.3 us (100 m cables)
  - ➢ FELIX data forward, decoder buffers ~0.6 us (@240 MHz)
  - ➢ Global level 1 Trigger decision latency + counting house -> IR ~0.3 us
  - ➢ … **~10 us**
- ❑ The goal is to achieve **~10 us** latency for the trigger algorithm

FPGA Resource Utilization

green: PCIe
purple: decoder
yellow: clusterizer
turquoise: local to global
brown: hsl4ml aggregate
pink: hsl4ml predict

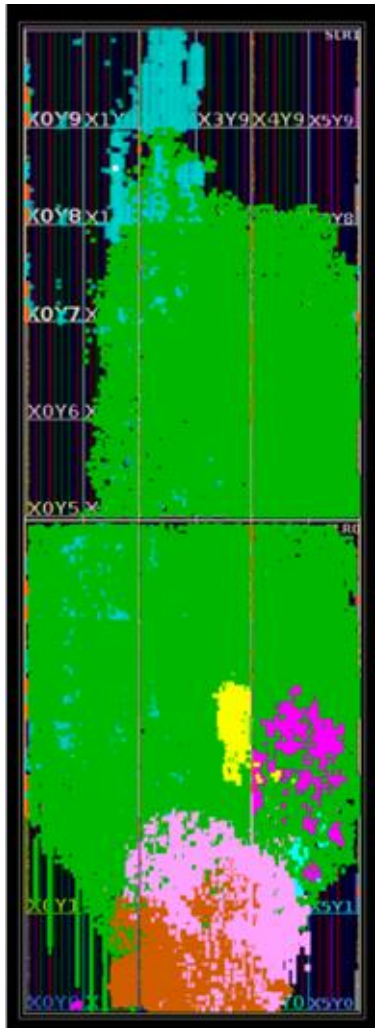
- Currently we have single stave implementation to validate modules
  - 3 decoders, 1 clusteriser, 1 transformation

	LUT (663K)	FF (1.3M)	BRAM (2K)	DSP (5.5K)
1-stave	163K (24.5%)	359K (27.6%)	1K (50%)	525 (9.5%)
8-staves	232K (35%)	412K (31.6%)	1.2K (60%)	581 (10.5%)

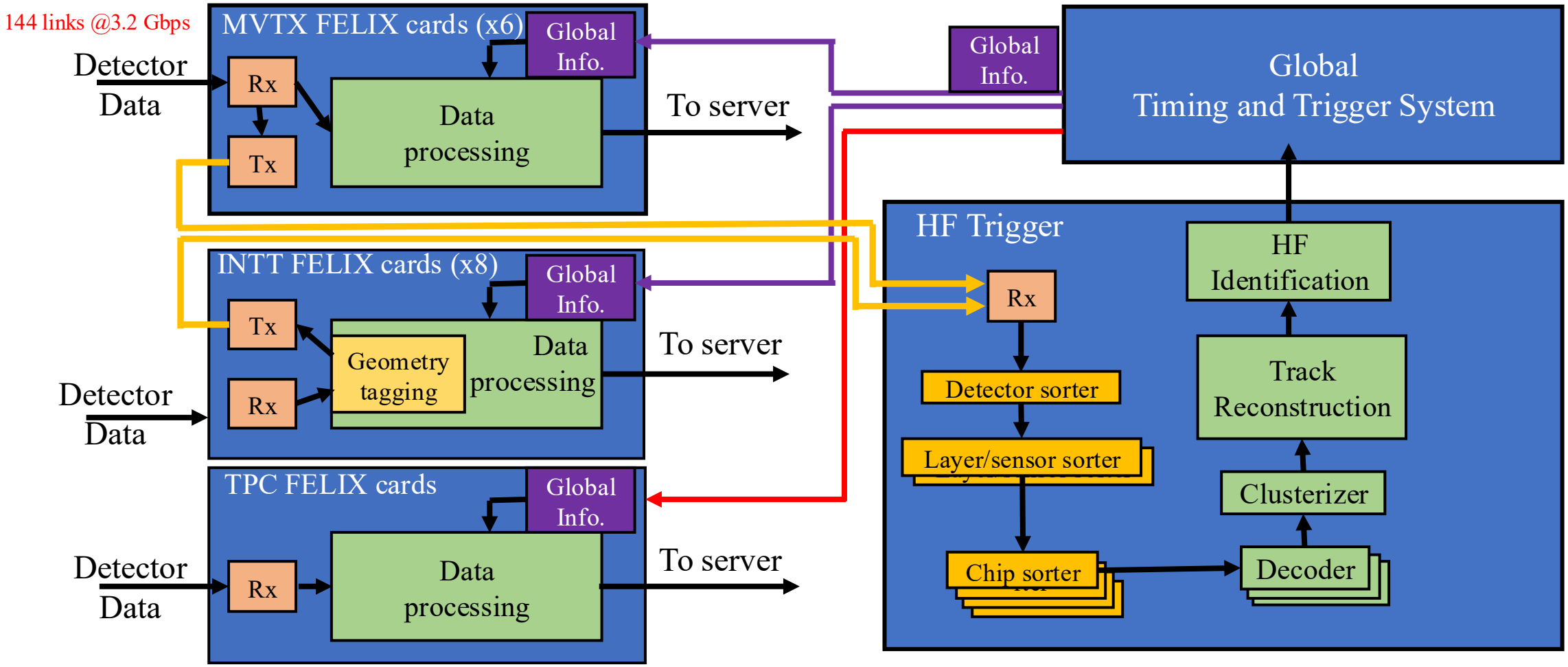
- Target is 72 decoders, clusterisers, and transformations
  - Current projection:

	LUT (663K)	FF (1.3M)	BRAM (2K)	DSP (5.5K)
Infrastructure	87K (13.1%)	196K (14.8%)	879 (40%)	-
Decoder	98K (14.7%)	91K (7%)	432 (21%)	-
Clustering	267K (40%)	213K (16.4%)	-	-
Transformation	25K (3.8%)	22K (1.7%)	540 (27%)	576 (10.4%)
AI module (FlowGNN)	194K (29%)	214K (16.4%)	406 (20%)	488 (8.8%)
AI module (hls4ml)	40K (6.1%)	45K (3.5%)	31 (1.5%)	517 (9.4%)

Need to reduce



(III) Full Demonstrator HF Trigger Diagram



Fast Data Processing and Autonomous Detector Control for High-Energy Nuclear Experiments *(Fast-ML)*

- ❑**A joint effort of NP, HEP, CS and EE**
  - **- LANL, MIT, FNAL, NJIT, GIT, ORNL et al**
- ❑**Physics simulation and AI-ML algorithms**
- ❑**Firmware implementation**
  - ***- hls4ml*, FlowGNN etc.**
- ❑**Demonstrator deployment**

DOE NP AI/ML projects:
- 2022-2023;
- 2024-2025