

Fast-ML for sPHENIX and EIC

Ming Liu

Los Alamos National Laboratory

For the sPHENIX/EIC Fast-ML Collaborations

MENU Conference 2026 @SBU

Oct. 19-23, 2026

Outline

- **Motivation: data challenge**

- Real-time AI for high data throughput challenges in HENP experiments
- Improved performance and fast physics analysis

- **Two recent R&D efforts**

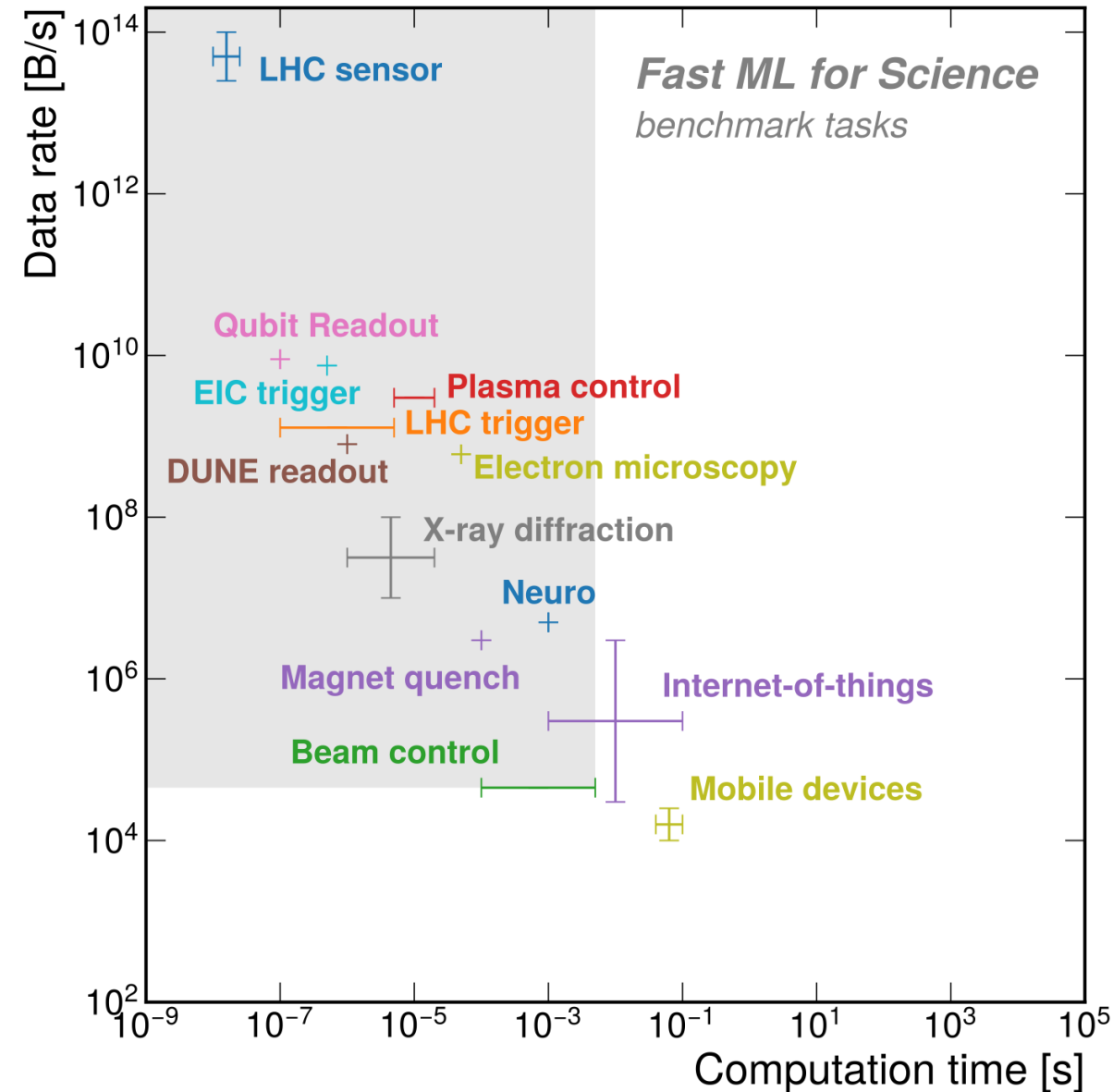
- sPHENIX AI-based HF Trigger demonstrator

*“Intelligent Experiments through Real-Time AI: Fast Data Processing and Autonomous Detector Control for High-Energy Nuclear Experiments”,
2022-2023, 2024-2025*

- Agentic AI for ePIC streaming

*“Genesis phase-I: Agentic AI for Real-time Expedited Discovery from High-complexity EIC Data Stream”,
2026 - 2027*

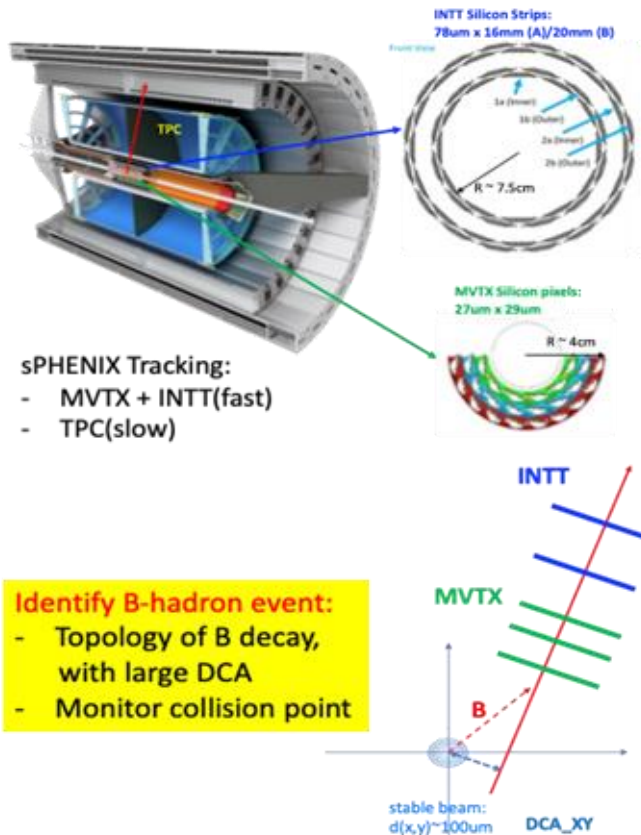
- **Future opportunities**



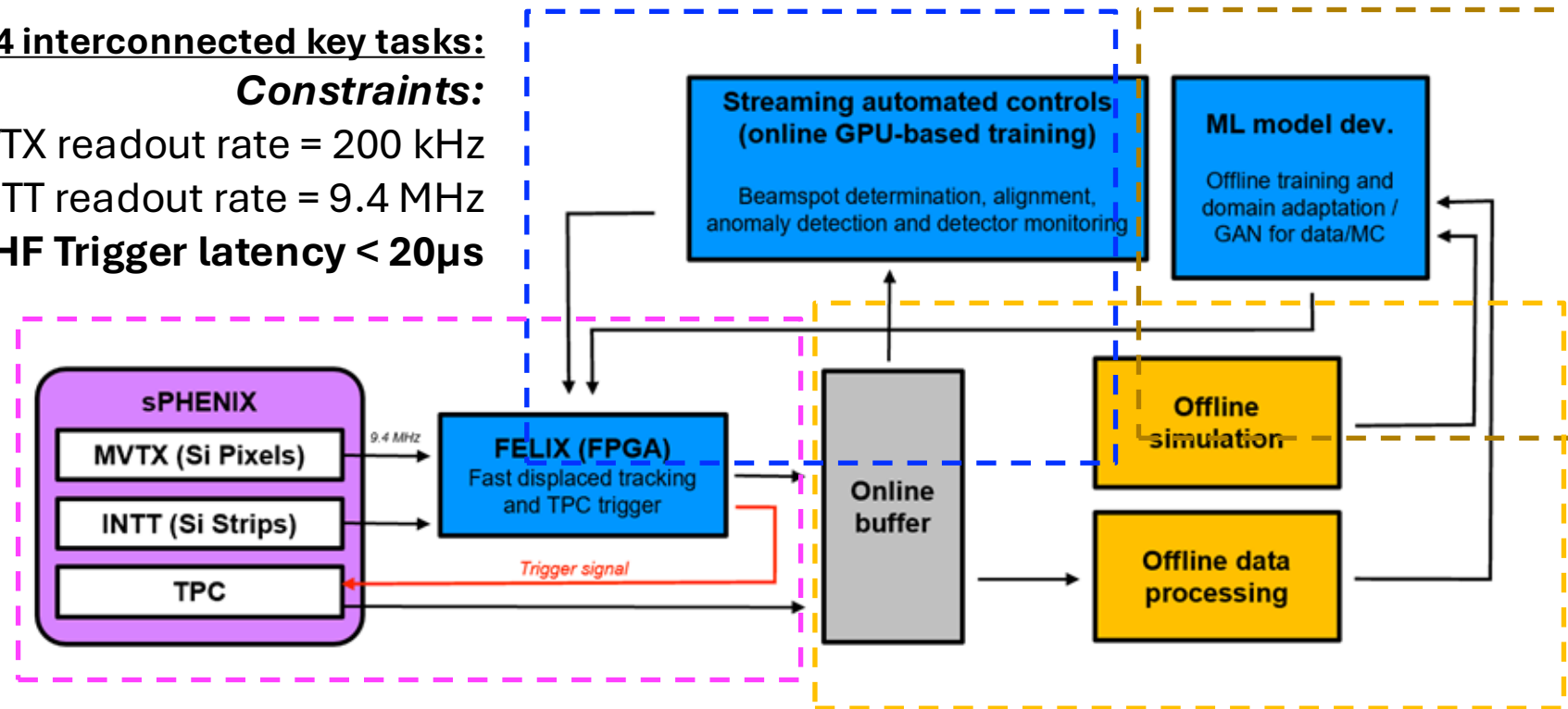
HF AI-trigger demonstrator for p+p collisions

Selective streaming real-time AI and autonomous detector control:

Deliver a demonstrator for p+p running for sPHENIX - generalizable for applications in experiments at the EIC



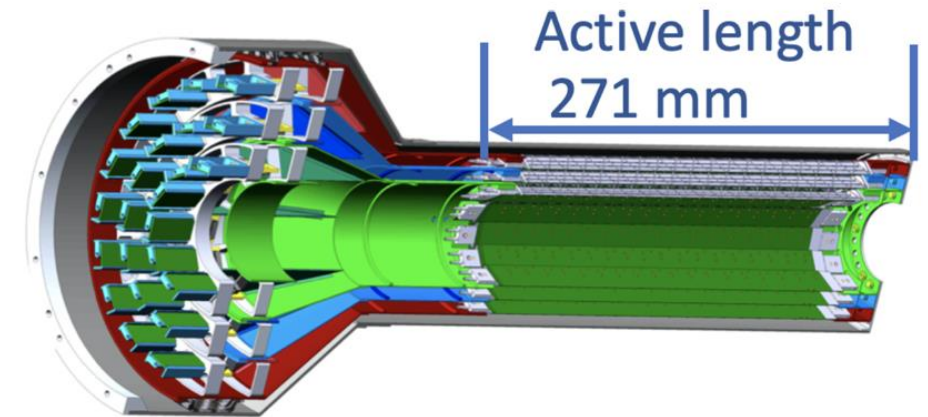
4 interconnected key tasks:
Constraints:
MVTX readout rate = 200 kHz
INTT readout rate = 9.4 MHz
HF Trigger latency < 20μs



MVTX and INTT: first full streaming detectors @RHIC

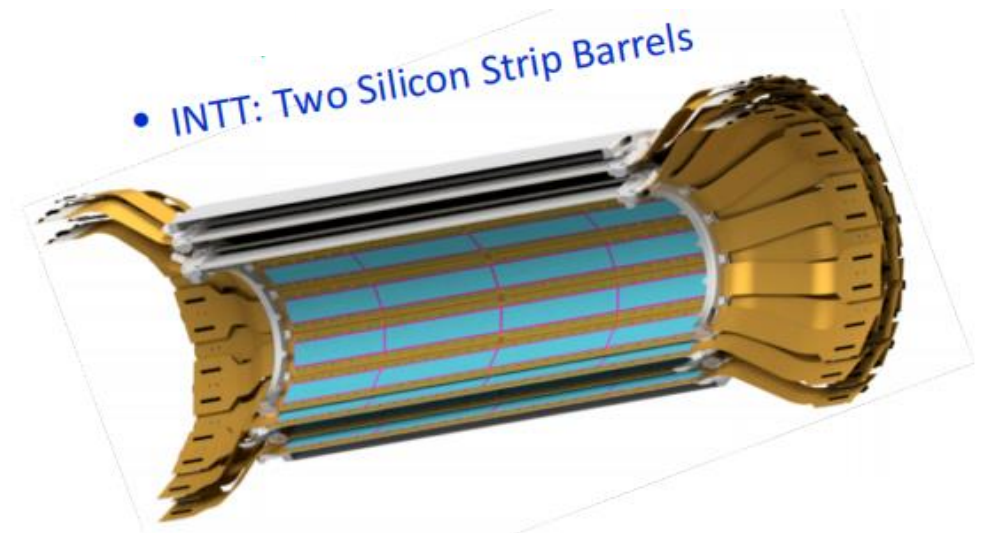
❑ MVTX – Monolithic-active-pixel-sensor based vertex detector

- Pitch: $27\text{ }\mu\text{m} \times 29\text{ }\mu\text{m}$
- Time resolution: $5\sim 10\text{ }\mu\text{s}$ (adjustable)
- 3 layers, 48 staves: $\sim 230\text{M}$ pixels channels



❑ INTT – micro-strip tracking detector

- Pitch: $27\text{ }\mu\text{m} \times 16$ (or 20) mm
- Time resolution: $\sim 50\text{ ns}$ ($< \text{BCO } 106\text{ns}$)
- 2 layers, 56 ladders

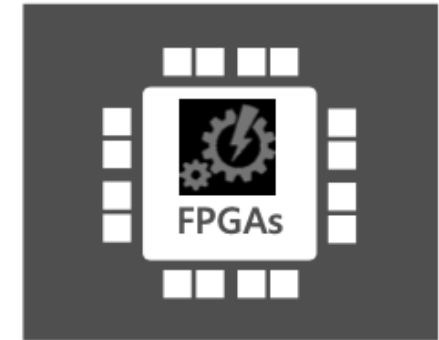
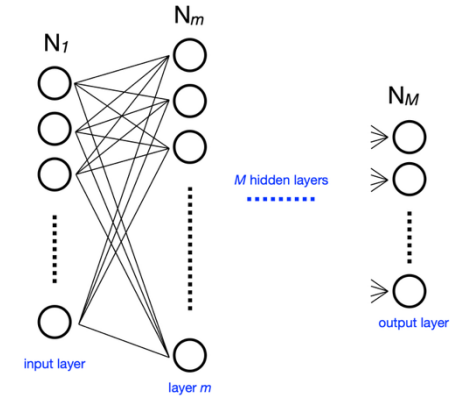


3 Successful R&D focus areas

- ❑ Physics/detector simulations and AI/ML algorithm development
- ❑ Translate AI/ML algorithms into hardware language FPGA code with (1) data processing latency and (2) hardware resource constraints
- ❑ Deploy FPGA algorithms in a demonstrator system in sPHENIX

Lessons learned from sPHENIX streaming operation with real beams, p+p, Au+Au in 2023-2025

Potential implications for the EIC (mid 2030s)



GNN Bipartite architecture design

Several other models are explored also

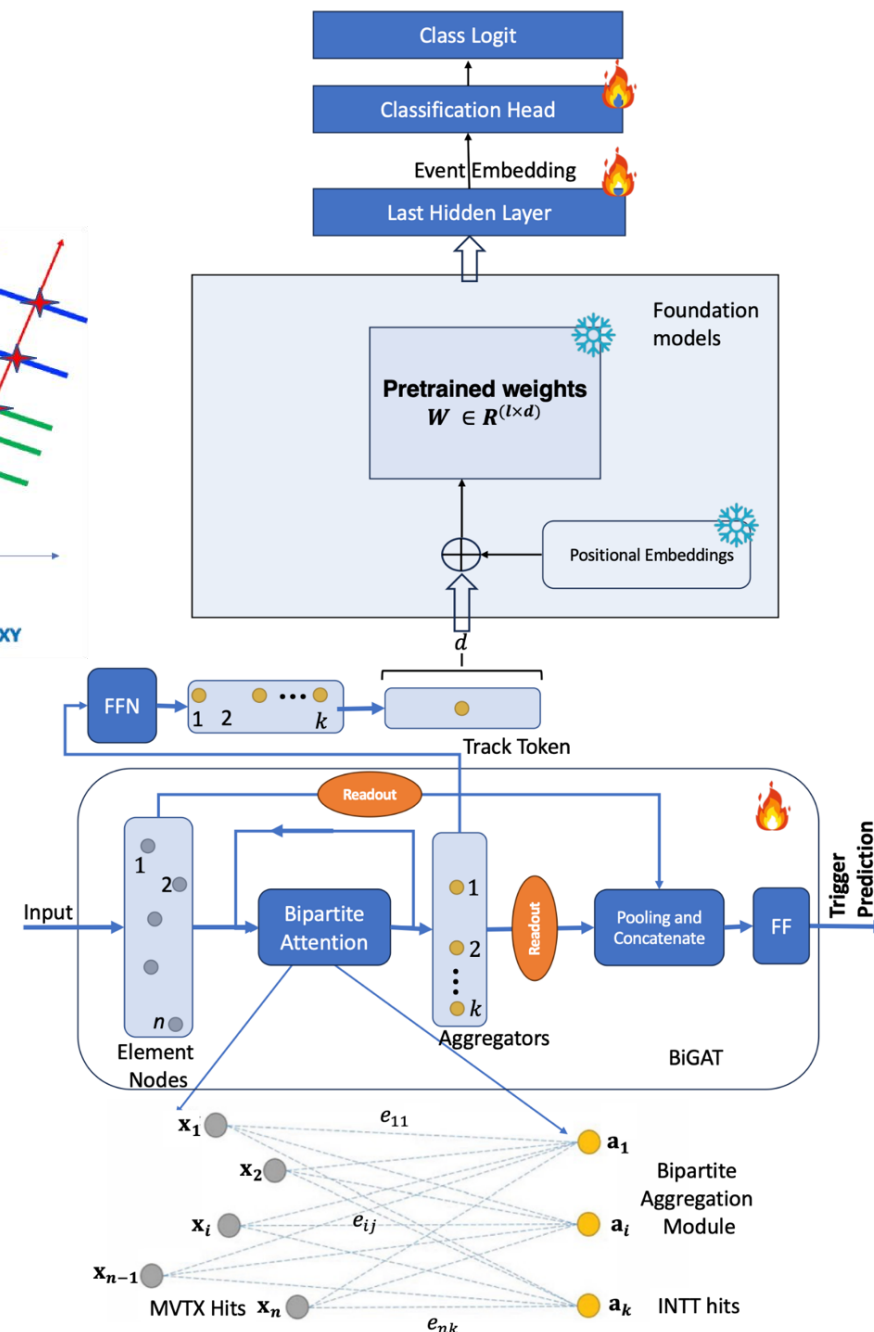
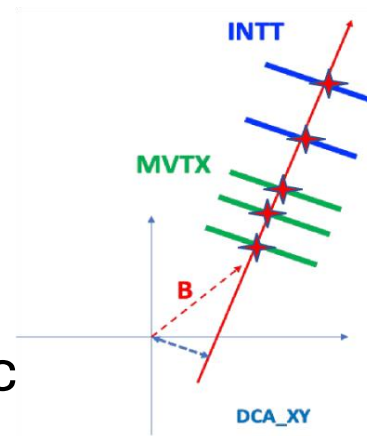
❑ Align with Detector Hardware

- MVTX v.s. INTT (50ns time stamps)
- Use INTT as anchors to reject out-of-sync MVTX hits

❑ Tracking between INTT and MVTX

❑ Bipartite graph represents all feasible connections between MVTX hits on the left and INTT hits on the right.

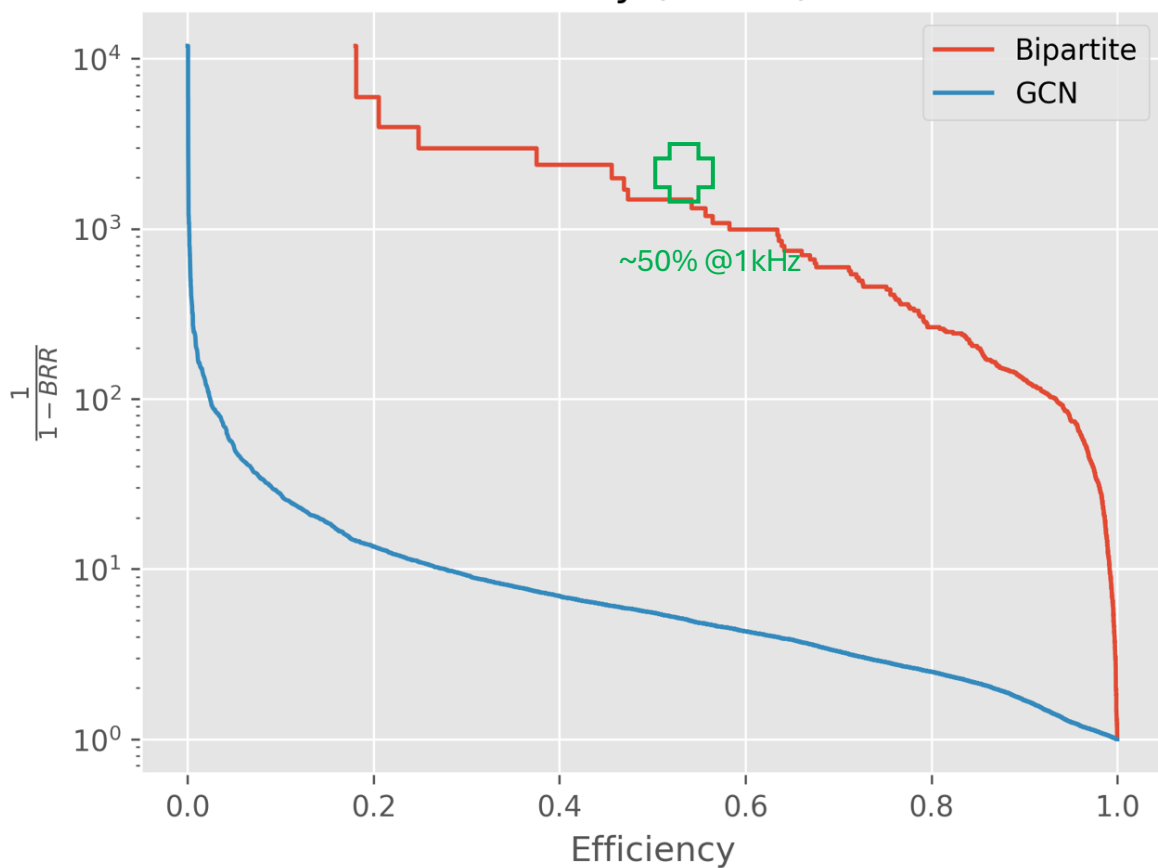
❑ Frozen Foundational Attention Networks



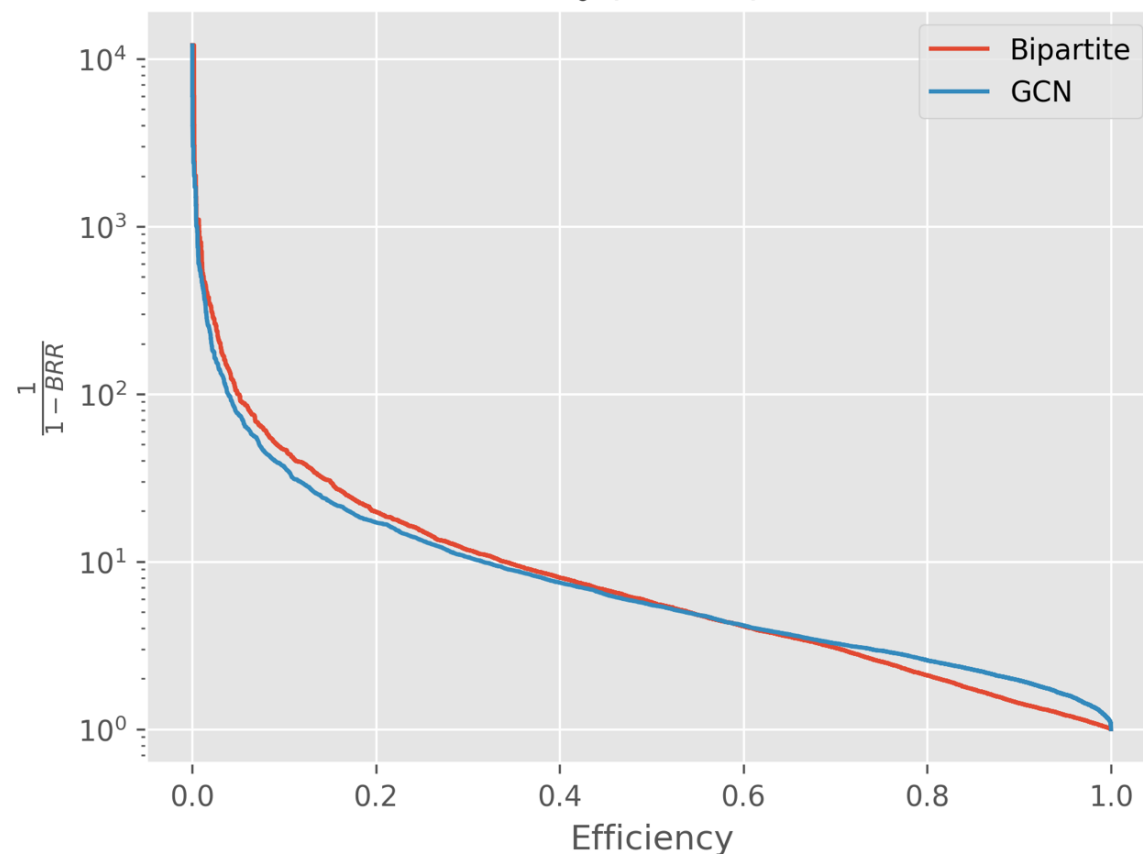
Signal tagging efficiency vs background reject rate (BRR)

Simulations: Charm and Beauty hadrons in p+p, w/ 20 MB pileup events per readout frame ($\sim 10\mu\text{s}$)

Beauty (W=20)

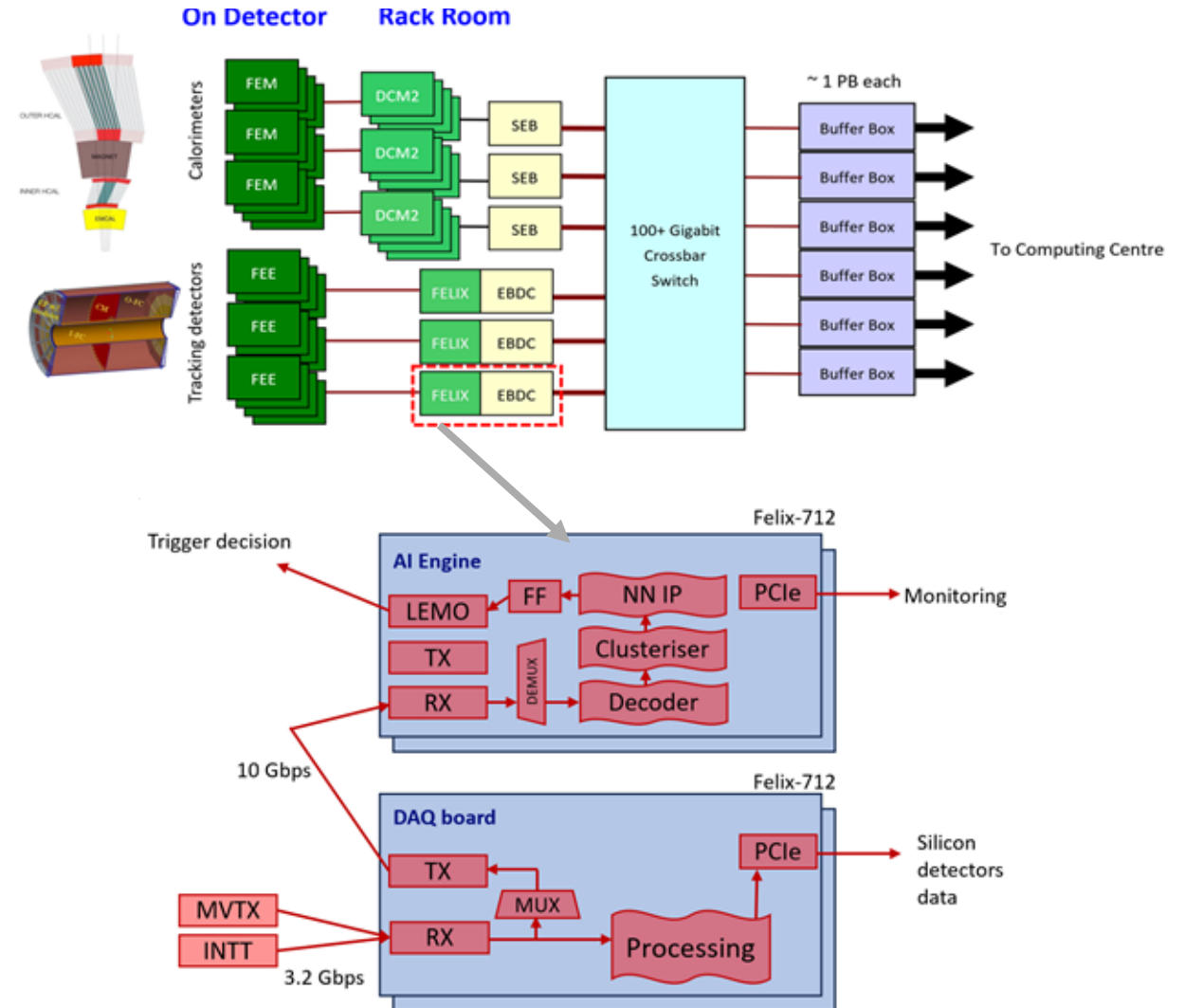


D_0 (W=20)



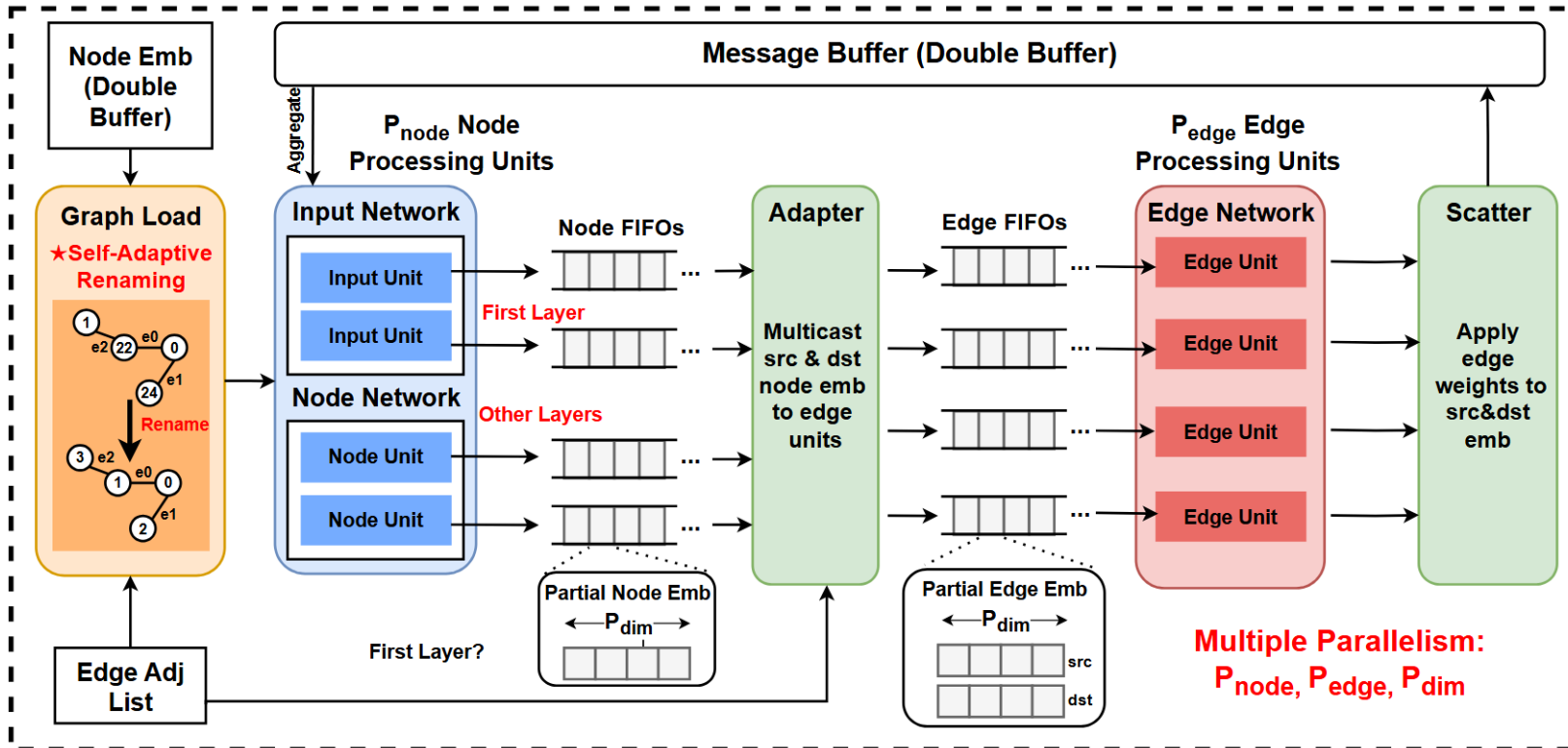
Readout and HF AI-trigger implementation in FPGA

- ❑ 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](https://arxiv.org/abs/2204.13103))
 - GNN-BiPartite model
 - [hls4ml](https://arxiv.org/abs/1804.06913) ([arXiv: 1804.06913](https://arxiv.org/abs/1804.06913))



Approach #1: FlowGNN

- ❑ FlowGNN is a flexible architecture for Graph Neural Network acceleration on FPGAs
- ❑ Manual implementations, from PyTorch $\rightarrow$ C $\rightarrow$ Verilog, using High Level Synthesis (HLS)
 - Version 1: Track construction only:
 - **8.82 us per graph** (Freq. 285 MHz), tested with: 92 nodes, 142 edges
 - Version 2: from Hits $\rightarrow$ Clustering $\rightarrow$ Triggering:
 - **9.2 us per graph** (Freq. 180 MHz), Tested with: 92 nodes, 142 edges



Approach #2: GNN Bipartite model

✅ The overall latency is **reduced by 14.6%** on FPGA

⚠ However:

- BRAM usage increased a lot because of more parameters
- The clock frequency drops from 285 MHz → 198 MHz (30% reduction) due to heavier resource utilization
- More latency is spent on data movement for parameter loading

📝 Planned:

- More aggressive quantization, Model pruning if possible

GCN (previous results) @ 285 MHz				Bipartite Model (new results) @ 198 MHz			
DSP	FF	BRAM	Latency	DSP	FF	BRAM	Latency
488 (5.4%)	214K (8.2%)	406 (20.2%)	8.82 us / graph	359 (3.9%)	344K (13.2%)	1766 (87.6%)	7.53 us / graph
							14.6% ↓

#3: *hls4ml* Initial implementation (MVTX-only MLP)

- ❑ The MLP-layerwise model has been synthesized for the FPGA
- ❑ The model consists of two parts
 - The first part, called the **aggregation step**, collects all the clusters. It is called for each cluster in a bunch crossing. This needs a high throughput: initiation interval every 1 clock cycle, **117 ns latency**
 - The second part, called the **prediction step**, is called once per bunch crossing, to make a prediction based on the ingested clusters: 63 clock cycles, **308 ns latency**
- ❑ The two models are synthesized separately, with the FPGA utilization for the FELIX 712 given below, using Vitis HLS and Vivado 2024.1.

	Aggregation step	Prediction step
LUT	23 587 (3.56%)	16 582 (2.50%)
FF	15 129 (1.14%)	31 226 (2.35%)
DSP	19 (0.34%)	498 (9.02%)
BRAM	0 (0%)	30.5 (1.41%)

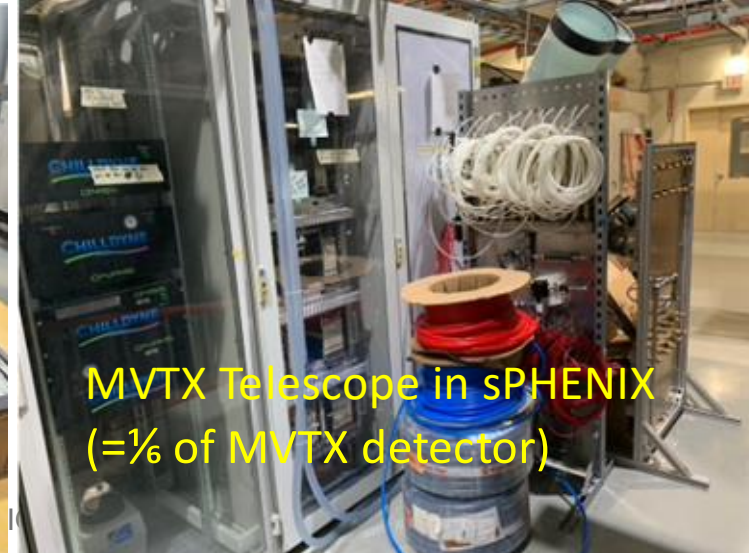
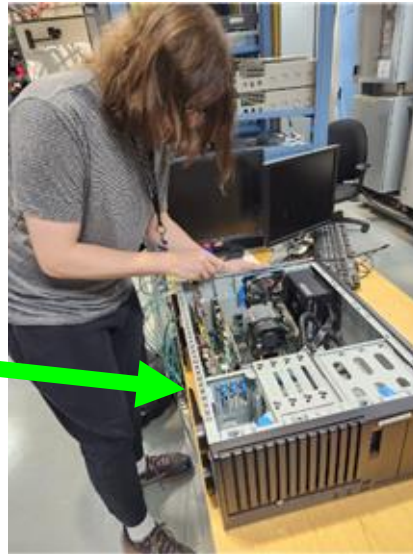
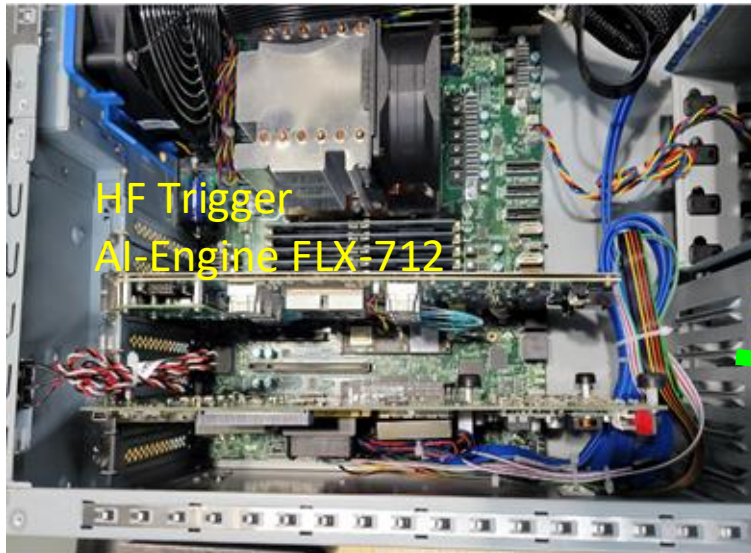


A small scale demonstrator:

- with 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
 - Save the timing (Bunch Crossing ID) and trigger decision from the AI
 - Configured the PCIe uplink (normally used just for configuration) to load real detector data to the board, for 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



sPHENIX HF Trigger Demonstrator R&D Team

❑ 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

- *FPGA, GPU, CPU etc.*

✓ 2022 – 2023

✓ 2024 – 2025

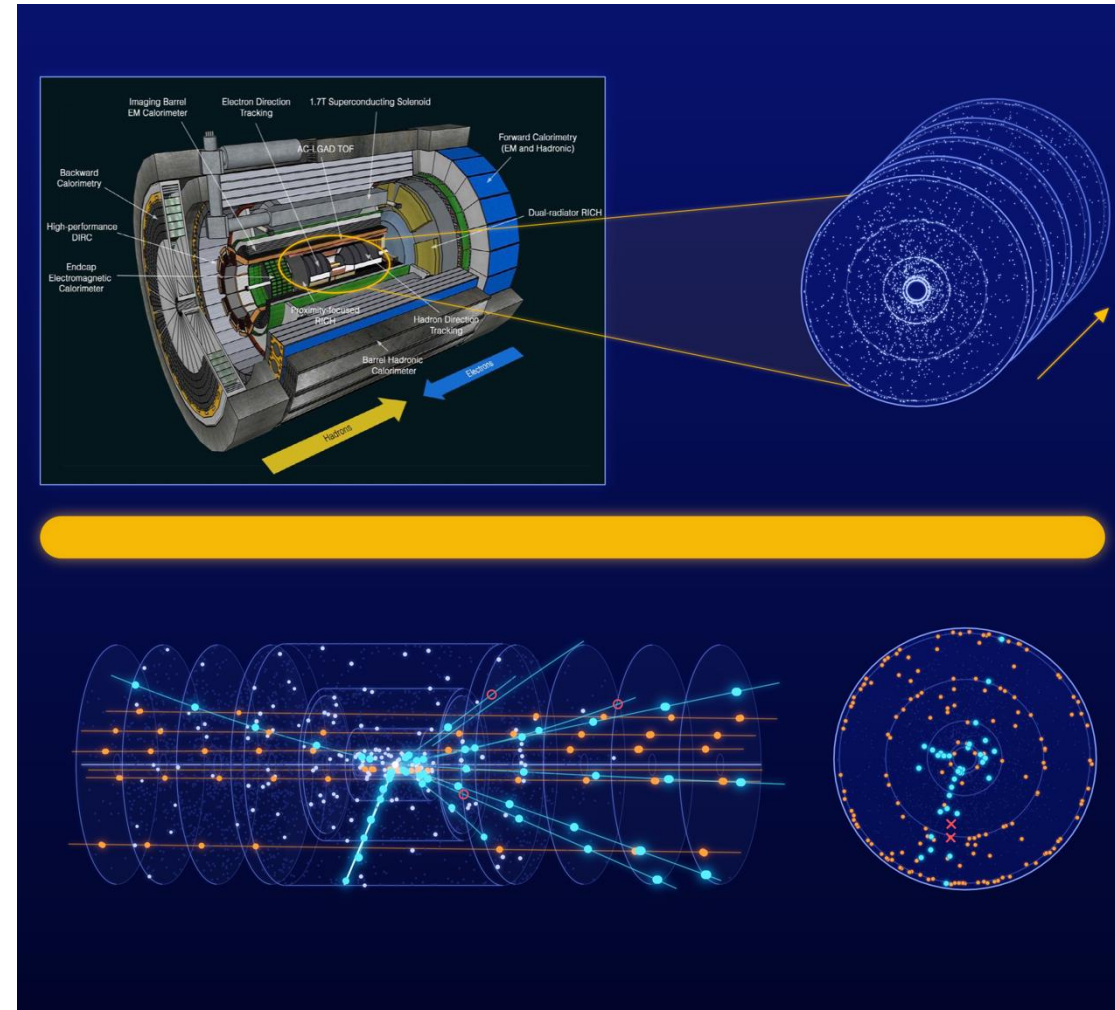


Project Title: Agentic AI for Real-time Expedited Discovery from High-Complexity EIC Data Streams

Genesis phase-I project: 2026 - 2027

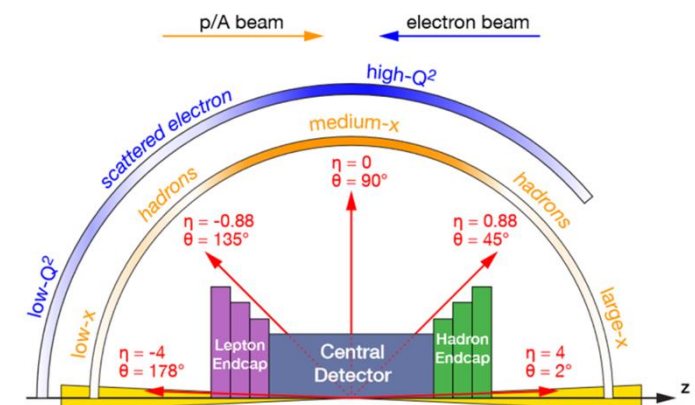
- EIC beam background rejection with AI in real-time

Senior/Key Personnel for Lead Institution and All Partner Institutions	
Institution Name	Senior/Key Personnel including the Director
Purdue University (Lead)	PI: Wei Xie (Director); co-PIs: Miaoyuan Liu, Ruqi Zhang
Fermi National Accelerator Lab	co-PIs: Nhan Tran, Jovan Mitrevski
Los Alamos National Lab	co-PI: Ming Xiong Liu; Senior/Key Personnel: Xuan Li
Massachusetts Institute of Technology	co-PI: Philip Harris
New Jersey Institute of Technology	co-PI: Dantong Yu



The EIC beam background challenges

<https://wiki.bnl.gov/EPIC/index.php?title=Background>



rates in kHz	5x41 GeV	5x100 GeV	10x100 GeV	10x275 GeV	18x275 GeV	Vacuum
Total ep	12.5 kHz	129 kHz	184 kHz	500 kHz	33 kHz	
hadron beam gas	12.2kHz	22.0kHz	31.9kHz	32.6kHz	22.5kHz	10000Ahr
	131.1kHz	236.4kHz	342.8kHz	350.3kHz	241.8kHz	100Ahr
electron beam gas (Bremsstrahlung scatterings)	2181.97 kHz	2826.38 kHz	3177.25 kHz	3177.25 kHz	316.94 kHz	10000Ahr
electron beam gas (Coulomb losses - w/ collimators)		116.57 kHz		29.56 kHz	0.86 kHz	10000Ahr
electron intrabeam (Touschek losses - w/ collimators)		1112.3 kHz		233.5 kHz	0.55 kHz	
DIS eA	kHz	kHz	kHz	/		
hadron beam (Au) gas	7.36kHz	10.3kHz	10.3kHz	/		10000Ahr
	79.1kHz	110.7kHz	110.7kHz	/		100Ahr
electron SR		36608 MHz		36608 MHz	3324 MHz	

Genesis Fast-ML for EIC streaming readout

- *Physics (DIS-electron, HF, Jets etc.) identification in real-time with background suppression*

Selective streaming readout for AI-Engine:

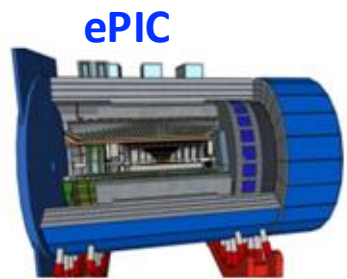
- ❑ AI SVT background/noise suppression near detectors (DAM/EBDC)
- ❑ tag DIS-electron to define DIS event ID
 - EMCal + SVT + ePID ...
 - DCA~0

SRO + AI/ML Fast Data Processing:

- **DIS e-tagger: event ID**

+ other rare process, HF-tagger etc.

- **Beam backgrounds ...**



O(2 Pbps)

FEB

RDO

DAM

EBD
C

Network
Switch

Online
Data Filter
& Monitoring

Buffer Box

Monitoring

Echelon-0

O(10 Tbps)

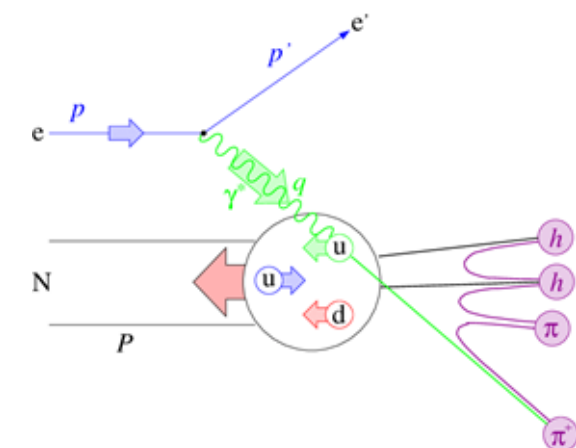
O(0.5 Tbps)

O(0.1 Tbps)

AI-on-SVT

e-tagger + Evt-ID

Adaptive
Learning



Bounded agentic AI

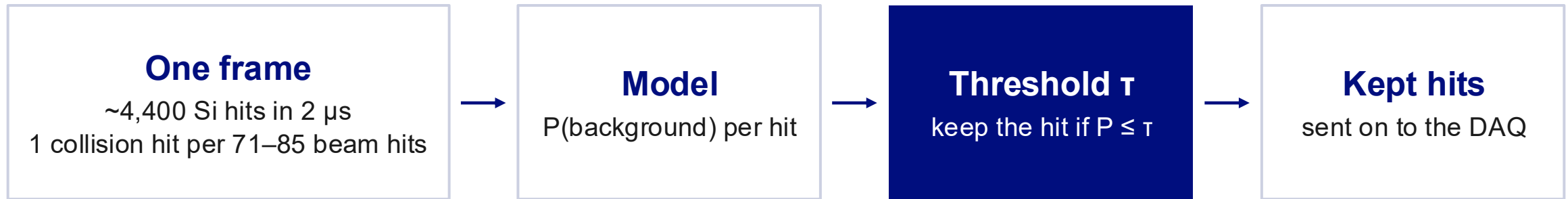
Automate selected model-to-deployment optimization steps under explicit physics and hardware constraints, with expert oversight.

Sensor-to-Physics Pipeline



Agentic loop — constrained quantization/compression sweeps · synthesis-report parsing · physics-aware regression testing · feeds back into model & deployment selection.

Per-hit score and a frozen threshold



How τ is set

τ = the 95th percentile of P over all collision hits of the 800 validation frames

Frozen, then applied unchanged to the 800 test frames

Reported on the 800 test frames

Retention ϵ_s = collision hits kept / all collision hits

Reduction R = $1/\epsilon_B$, with ϵ_B = beam hits kept / all beam hits

Gate, criterion 1: $R \geq 10$ at $\epsilon_s \geq 95\%$

Model families and 3-hit rejection geometry

Rule

No learning: keep a hit if its best 3-hit segment points to the vertex

Trees

Boosted decision trees; the full tree uses 27 numbers per hit: layer, r , $|z|$, charge + 9 triplet features

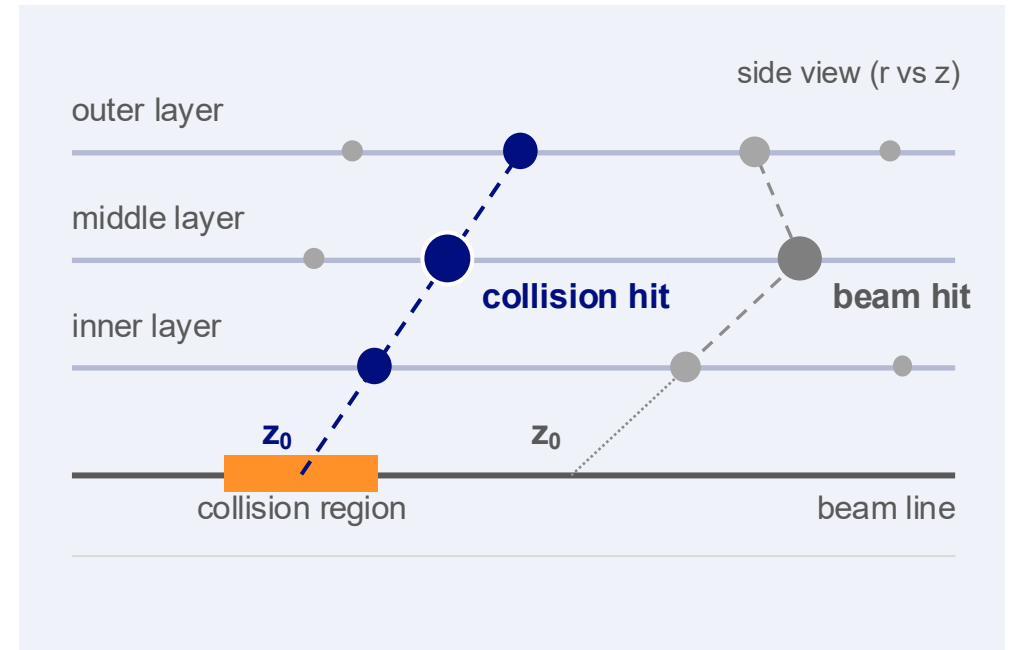
Weakly supervised

Weakly supervised tree: position + charge, trained on frame labels (overlay frame vs clean collision event)

EdgeConv

Graph network: each hit exchanges learned messages with its neighbors, 3 rounds

Triplet features: three questions per hit



Three key results and the benchmark deliverable

1 No fine silicon timing

20× | 24×

fewer beam hits at 95%
collision-hit retention

gate: 10×

2 No hit labels

12.5× | 16×

weakly supervised tree: frame
labels only, level with hit labels

3 What decides it

Readout

fine silicon time worth ~4×

Background

retention holds up to ×30

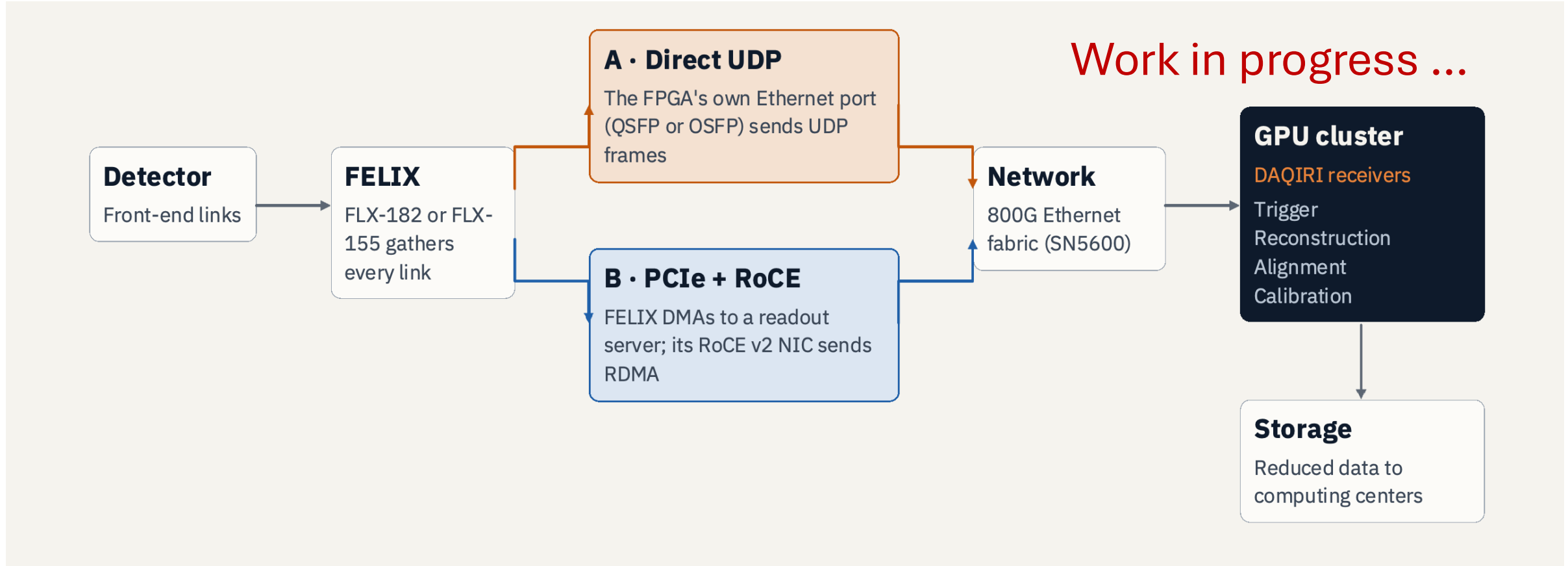
Geometry

the outer barrel is the weak spot

$Q^2 \geq 100$ | $Q^2 \geq 10$, on 800 test frames each; threshold frozen on validation frames
Fine silicon time = a per-hit silicon time stamp as the simulation stores it (labeled 10 ns, 8 ns rms), real 2 μ s MAPS do not give it



FPGA and GPU streaming development path



Two network options: direct FPGA UDP, or PCIe DMA followed by RoCE v2 RDMA.

The bench we ran today scales to FLX-155A, 800G NICs and an 800G switch

Speed and Latency: Hardware Considerations

TODAY · OCTOBER 2026 · 2 × 100G INTO ONE GPU NODE · MEASURED

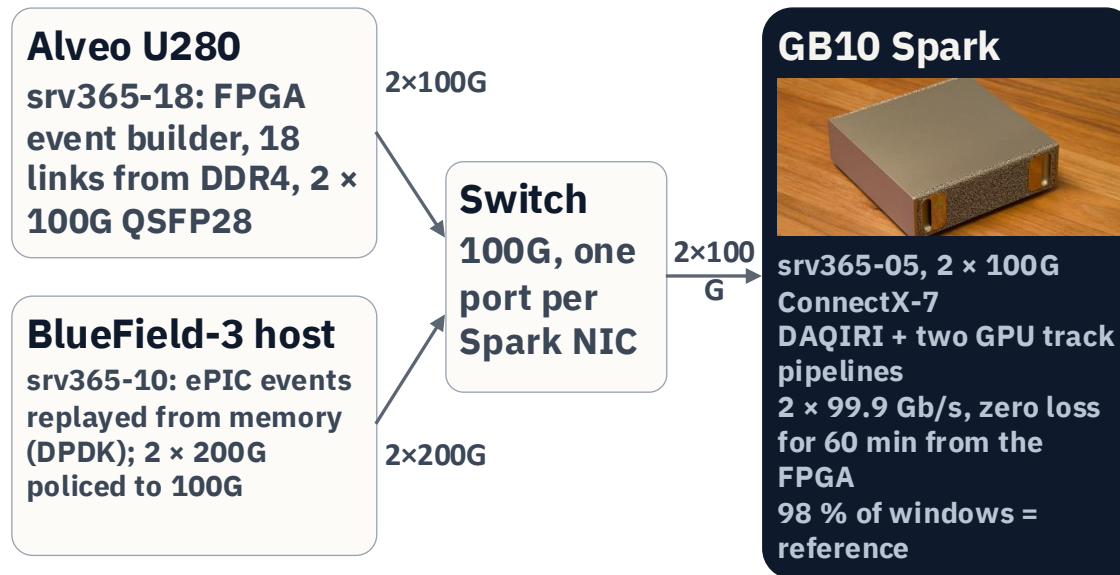
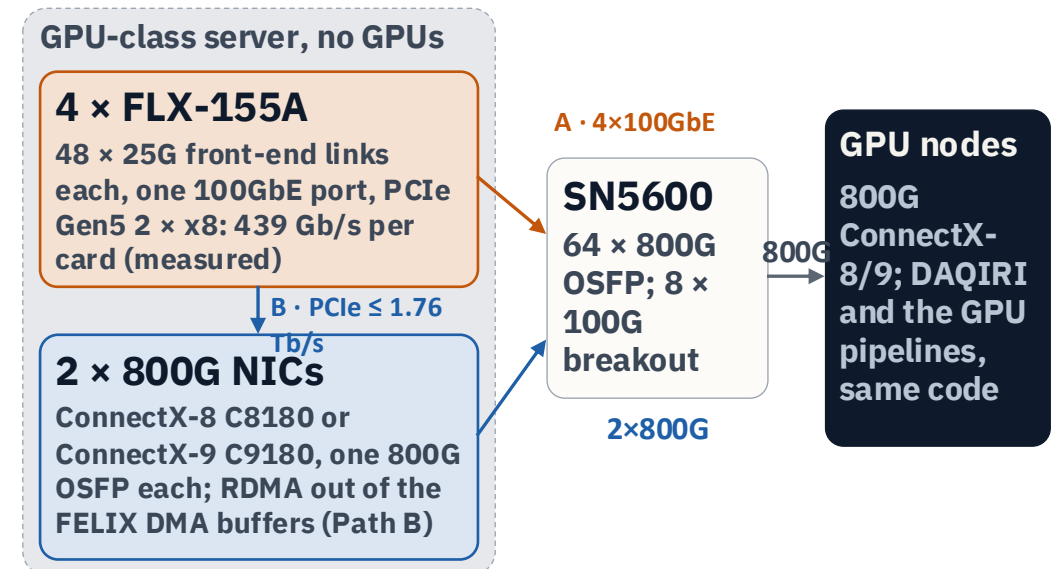


Photo (cropped): Daniel Lu, Wikimedia Commons, CC BY-SA 4.0, creativecommons.org/licenses/by-sa/4.0

Same chain, from the FPGA: 199 Gb/s with zero loss for an hour, every frame counted at every stage.

No untested assumption: FLX-155A is measured, ConnectX-8/9 and SN5600 are off the shelf, and DAQIRI with the GPU pipelines ran today at line rate on real ePIC events.

PROPOSED · 4 × FLX-155A · 2 × 800G OUT · STOCK PARTS

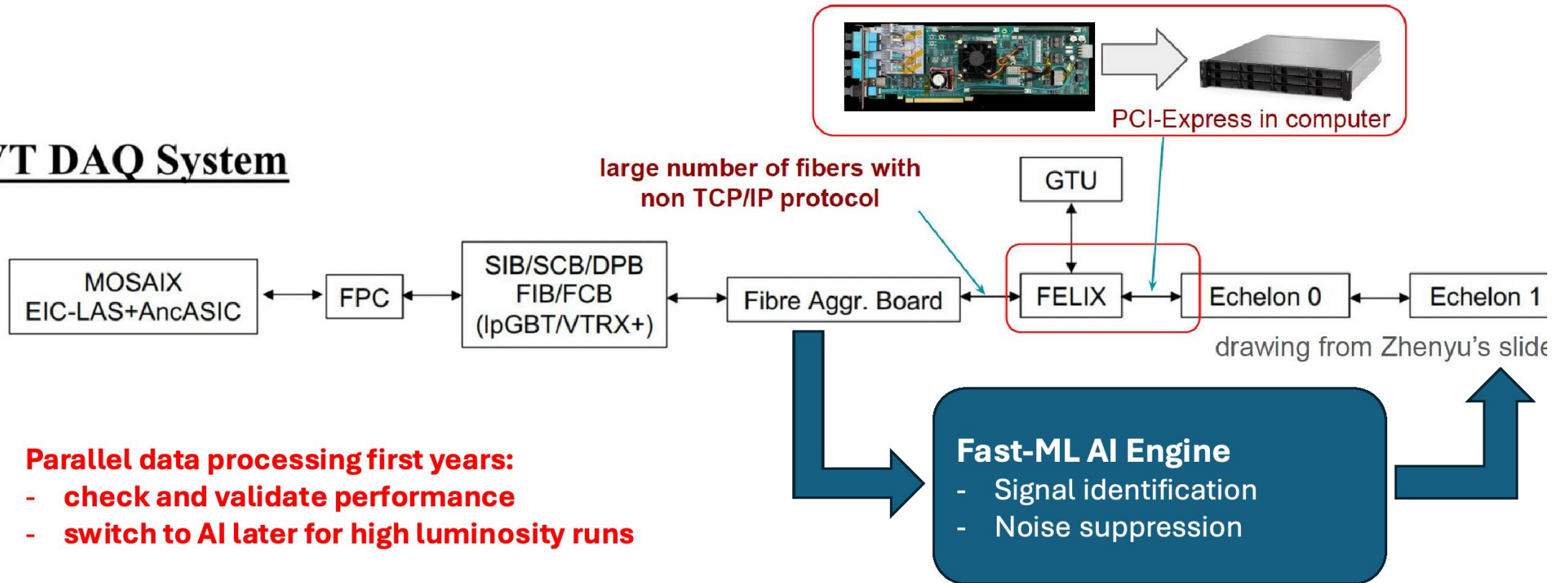


Swap in FLX-155A, the SN5600 and 800G ConnectX-8/9: the receiver, frame format and GPU pipelines do not change.

Phase-II proposal consideration

- SVT demonstrator and parallel AI processing

SVT DAQ System



Co-develop with ePIC DAQ. Validate in parallel before promotion for high-luminosity operation.

Summary and outlook

- Successful real-time HF AI-Trigger demonstrated for p+p in streaming mode
- Phase-I ties realistic SVT streams to physics-preserving ML and bounded model-to-deployment optimization, in progress
- Benchmark v1 trees reach 20–24× beam-hit reduction, with 95% signal retention.

Next: close downstream-physics checks, outer-barrel performance, calibration under shift and measured hardware latency.

Prepare for Genesis Phase-II proposal development, join us!