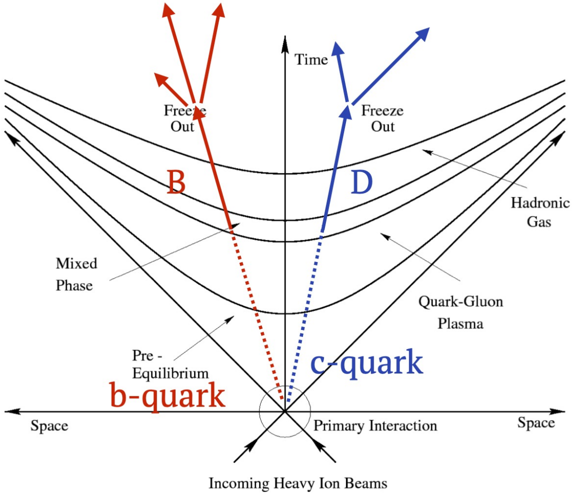
Machine Learning based fast data processing and autonomous trigger for sPHENIX and EIC detectors

Y. Corrales Morales
Los Alamos National Laboratory
For the Fast-ML Team

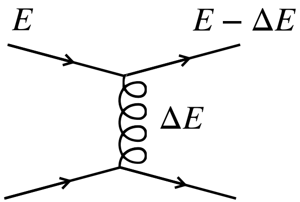
RHIC AUM22 workshop on Accelerating RHIC science with Machine Learning

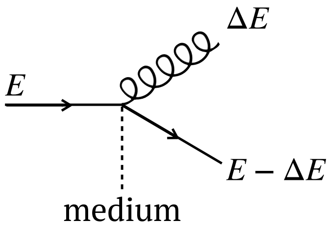
Physics Motivation



Quantitative improvement in the characterization of the QGP properties by a high precision measurement of rare probes over broad p_t range.

- Heavy flavours (i.e. charm and beauty quarks) are mainly produced in hard-scattering processes in shorter time scales compared to the QGP formation time
- HF probe the entire space-time evolution of the system, losing energy by interacting with the medium constituents via elastic scatterings and gluon radiations

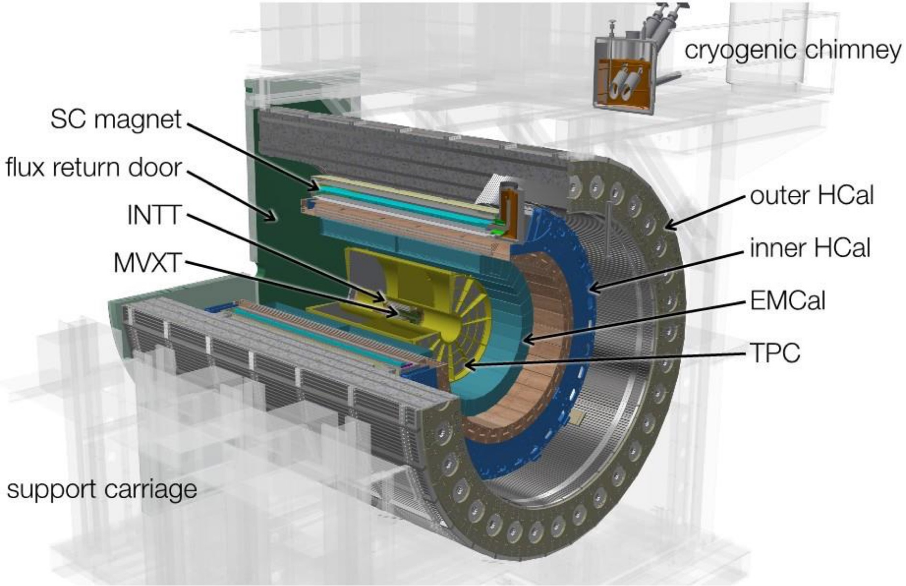




- Properties of in-medium energy loss studied via the nuclear modification factor R_{AA}

$$\boxed{R_{AA} = \frac{1}{\langle N_{coll}^{AA} \rangle} \frac{dN_{AA}/dp_T}{dN_{pp}/dp_T}}$$

The sPHENIX detector

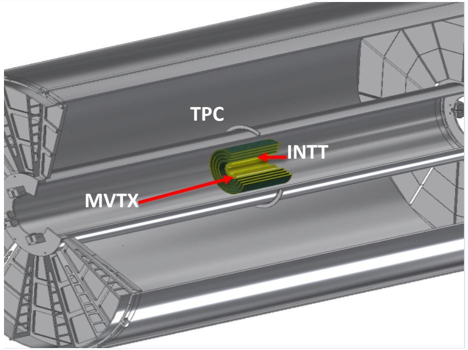


- Hermetic detector designed to study heavy flavor and jet physics in Heavy Ion Collisions at RHIC:

First run year	2023
$\sqrt{s_{NN}}$ [GeV]	200
Trigger Rate [kHz]	15
Magnetic Field [T]	1.4
$\lvert\eta\rvert$	$\leqslant$1.1
$\lvert z_{vtx}\rvert$ [cm]	10
N(AuAu) collisions*	$1.43\text{x}10^{11}$

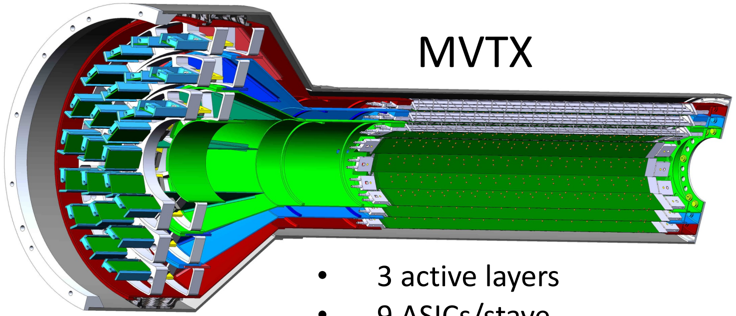
* In 3 years of running

Tracking at sPHENIX



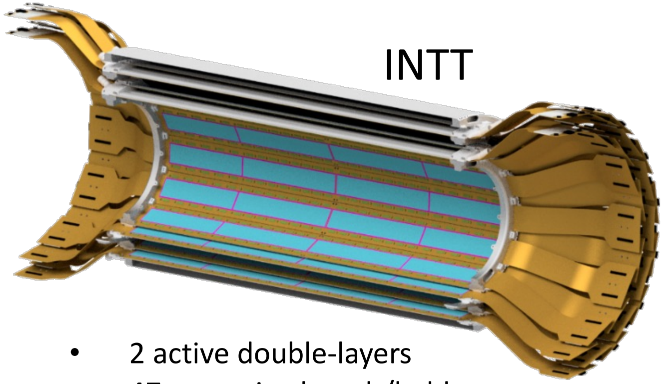
- Tracking consists of 3 sub-detectors:
  - Pixel Vertex Detector (MVTX)
  - Intermediate Silicon Tracker (INTT)
  - Time Projection Chamber (TPC)
- MVTX and INTT are both capable of streaming readout
- Combined tracking to r = 10.3 cm

MVTX



- 3 active layers
- 9 ASICs/stave
- 27 cm active length/stave
- Pixel detector, 27um x29um

INTT



- 2 active double-layers
- 47 cm active length/ladder
- Silicon strip detector

sPHENIX HF constraints

sPHENIX

- sPHENIX has great tracking and calorimetry
- However, limited by calorimetry backend readout rate (15kHz) in triggered mode
- Low HF production rate (rare events)
- Very high pp rate at RHIC ~10 MHz
  - **Charm production rate: ~ 100 kHz**
    - 0.5 mb/42 mb ~ 1%
  - **Beauty production rate: ~ 500 Hz**
    - 2 ub/42mb ~ 0.005%
- No effective trigger to select low p_t HF events
  - **Lost most of the HF event at low p_t**
- Plan: Use tracker SRO to recover some heavy flavor physics potential
  - **Huge data volume, DAQ/tape cost**

Year	Species	$\sqrt{s_{NN}}$ [GeV]	Cryo Weeks	Physics Weeks	Rec. Lum. $\lvert z\rvert$ <10 cm	Samp. Lum. $\lvert z\rvert$ <10 cm
2023	Au+Au	200	24 (28)	9 (13)	3.7 (5.7) nb^{-1}	4.5 (6.9) nb^{-1}
2024	$p^{\uparrow}p^{\uparrow}$	200	24 (28)	12 (16)	0.3 (0.4) pb^{-1} [5 kHz] 4.5 (6.2) pb^{-1} [10%-*str*]	45 (62) pb^{-1}
2024	$p^{\uparrow}$+Au	200	–	5	0.003 pb^{-1} [5 kHz] 0.01 pb^{-1} [10%-*str*]	0.11 pb^{-1}
2025	Au+Au	200	24 (28)	20.5 (24.5)	13 (15) nb^{-1}	21 (25) nb^{-1}

- sPHENIX beam-use proposal. 5 kHz refers to final rate with triggered readout
- 10%-str refers to 10% streaming readout

The DOE FOA Call in 2021

- Proposals called on 3/16, 2021
  - **Short deadline, 4/30/2021**
  - **Very intense work**


Initial team of NP, HEP and CS

- **LANL, MIT, FNAL and NJIT**
  - **ORNL, CCNU and UNT joined later**

DEPARTMENT OF ENERGY
OFFICE OF SCIENCE
NUCLEAR PHYSICS

DATA ANALYTICS FOR AUTONOMOUS OPTIMIZATION AND CONTROL OF ACCELERATORS AND DETECTORS

FUNDING OPPORTUNITY ANNOUNCEMENT (FOA) NUMBER:
DE-FOA-0002490

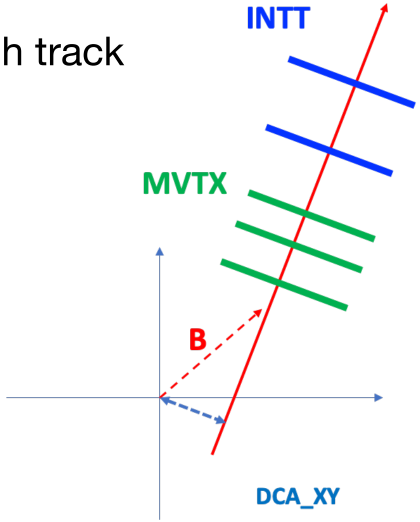
ANNOUNCEMENT TYPE: INITIAL
CFDA NUMBER: 81.049

The proposal

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

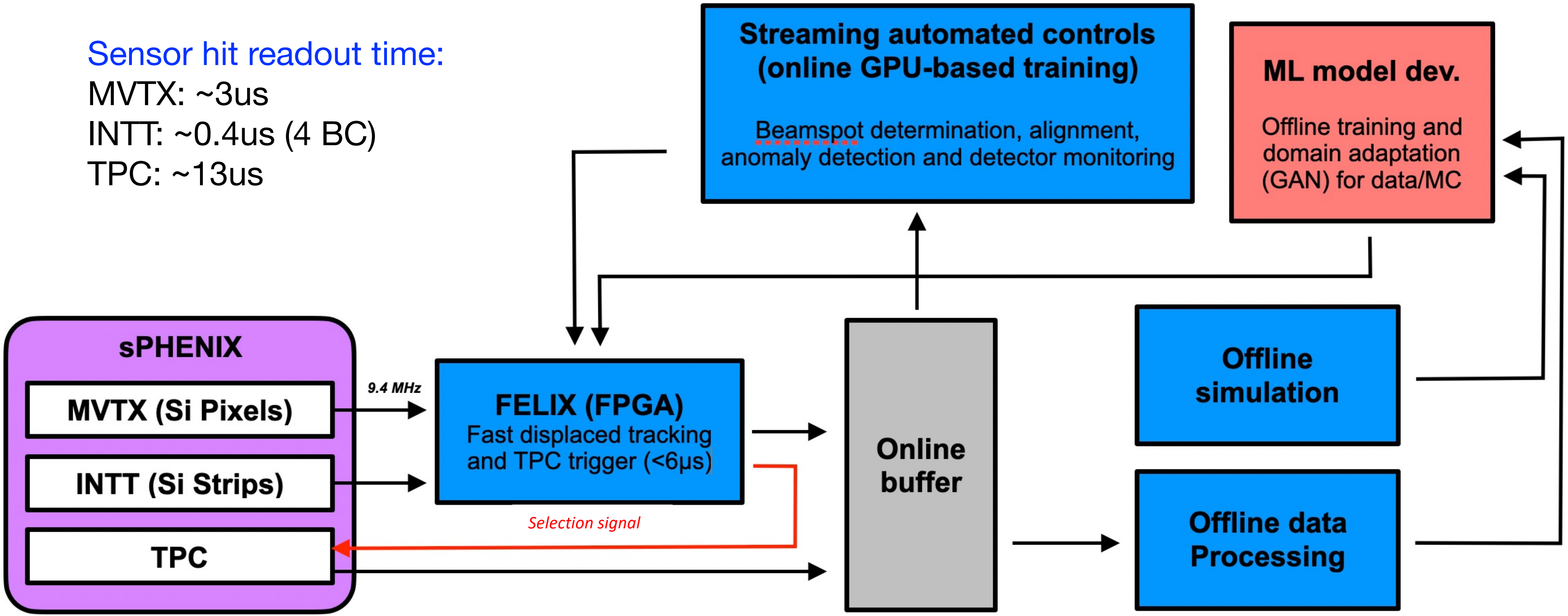
A proposal submitted to the DOE Office of Science
April 30, 2021

- Embed AI/ML algorithms on FPGA-based trigger system
  - **Low trigger decision latency**
- Streaming readout key inner trackers to FPGAs to identify HF events through track topology
  - **High efficiency in HF tagging with AI/ML**
  - **HLS4ML package developed by HEP**
- Monitor and update beam-spot and detector alignment in real time
  - **Update geometry in real time**
- Send HF-trigger signal to the rest of other detectors
  - **Initiate readout if not already in the data stream**



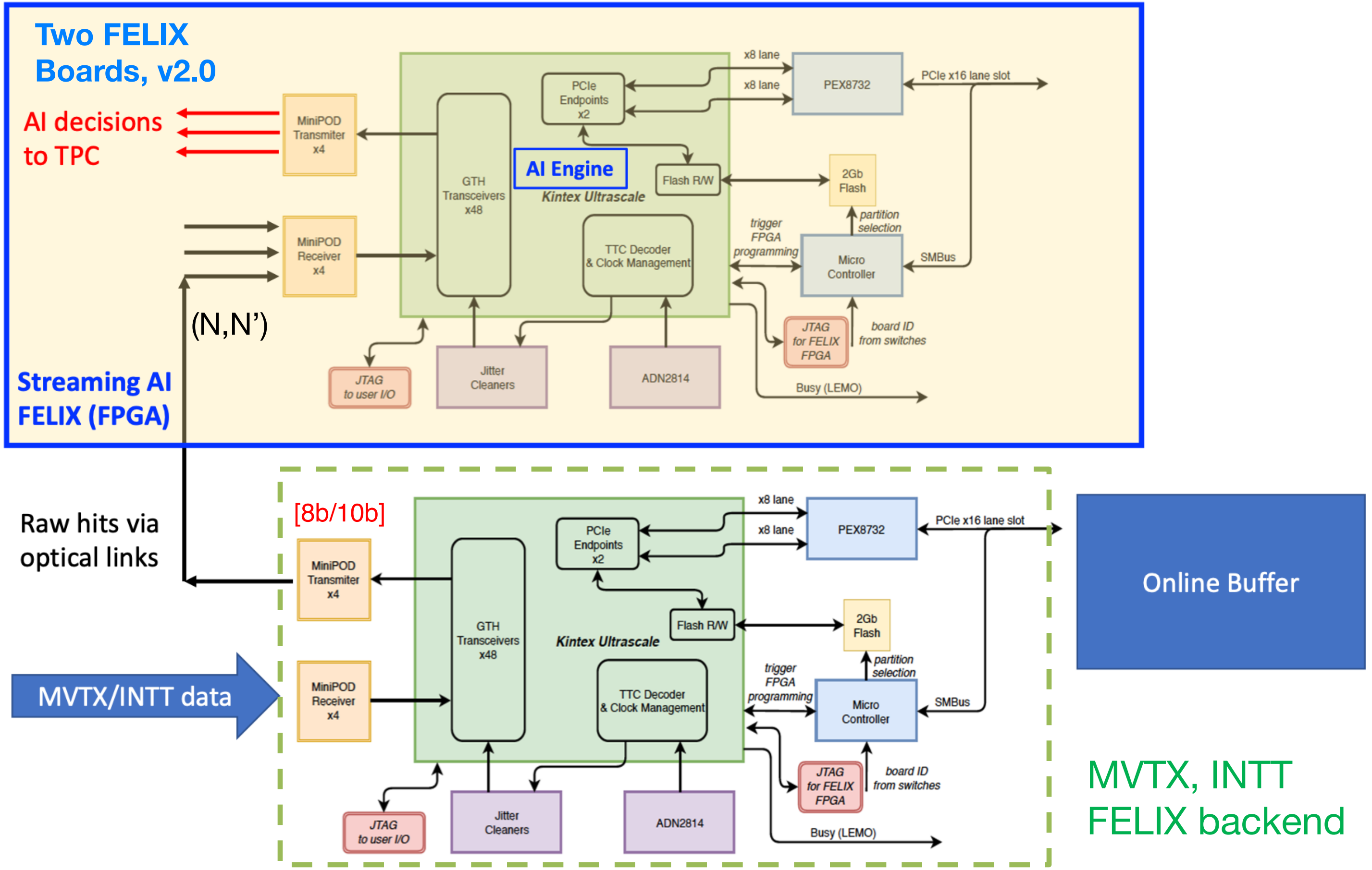
HF AI Trigger: sPHENIX as a Test Ground

Sensor hit readout time:
MVTX: ~3us
INTT: ~0.4us (4 BC)
TPC: ~13us



sPHENIX DAQ & Trigger integration challenge

HF AI Trigger: sPHENIX as a Test Ground



The first steps

- Success stories since proposal outcome

1. Full Geant4 simulations of MVTX and INTT
2. Convert simulation output to equivalent bit pattern
3. Tracking GNN algorithms are being developed at NJIT
4. Prototype hardware set up at LANL with host-to-client transfers running
5. Second lab being set up at MIT
6. HLS4ML development at Fermilab and MIT
7. FELIX FW development at ORNL and LANL

The next steps

- Aims for the next several months:

1. Improve initial tracking and selection algorithms

- With this we can:

1. Convert algorithms to HLS code to go on FPGA
2. Pass simulated data to FPGA as if it were real data

- Aim to install device in sPHENIX before 2024 (RHIC *pp* run)

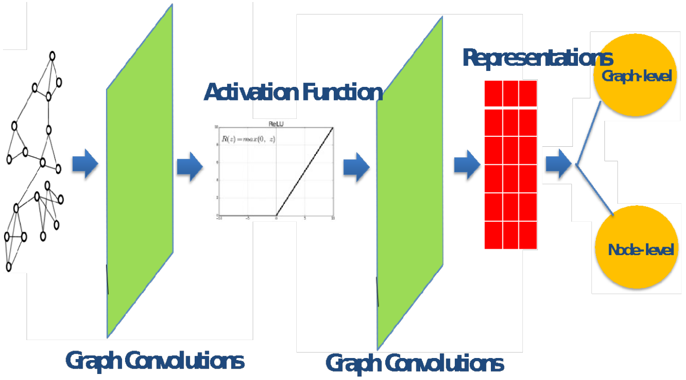
Simulation Hits to Raw Data Bit stream

- sPHENIX physics simulation MVTX and INTT
  - **cc, bb and MB samples**
- Convert MC hits into “Raw data stream” to AI-Engine
  - **MVTX: ~RU output bit-stream**
  - **INTT: ~ROC output bit-stream**
- Feasibility study of high-speed data transfer from FELIX (MVTX, INTT) to AI-Engine at ORNL
  - **FELIX loopback tested at 12.8 Gbps (MVTX 3.2Gbps/Link payload), to reduce # of g-links to AI-Engine**

Trigger AI Algorithm R&D – NJIT/LANL

- Implemented several models to solve the trigger detection problem:
  - **Directly applied GNN model to trigger detection problem (GNN)**
  - Added a global vector to the GNN model to represent some global feature (VPGNN)
  - DiffPool model (DiffPool)
  - VpGNN + DiffPool (GNNDiffPool)
  - ParticleNet , Giorgian
- Another model we tried: Set2Graph (Affinity Matrix Prediction)

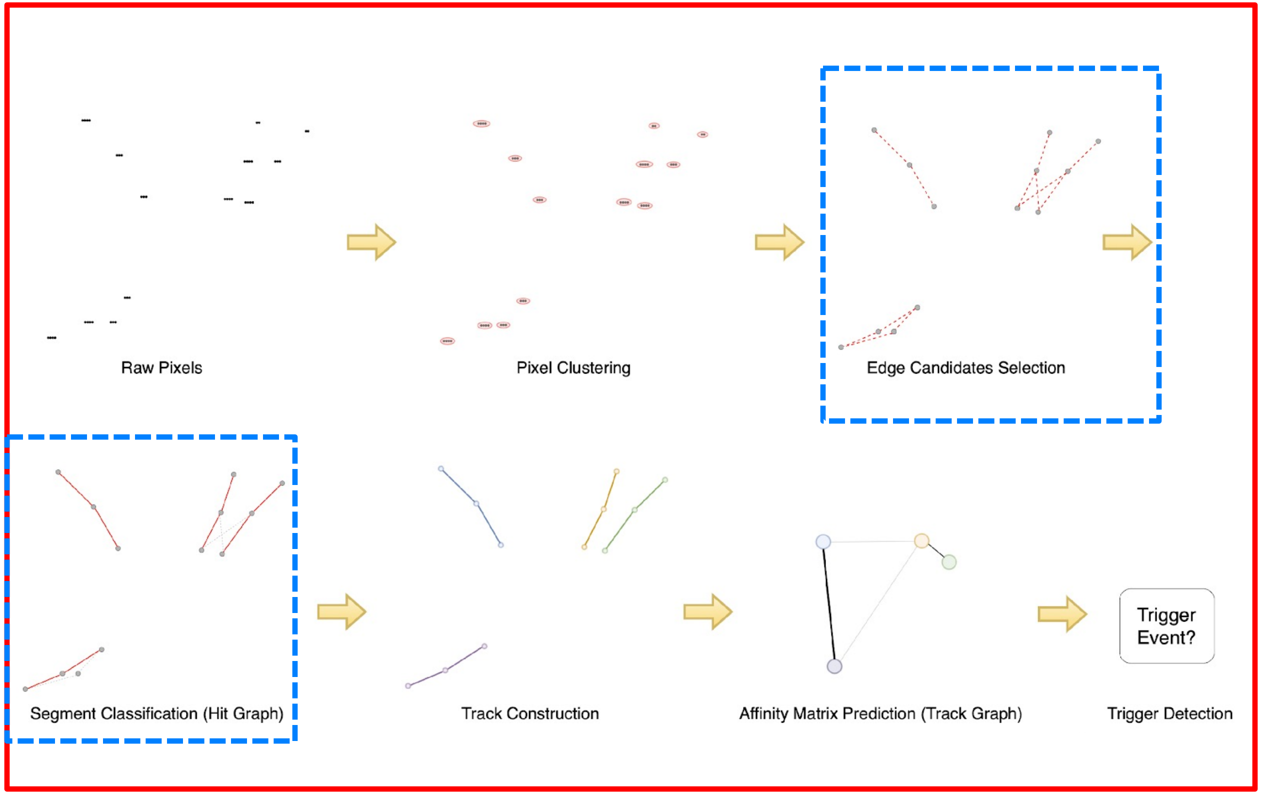
Inputs:
- raw hits
- tracklets



GNN Fast Tracking at NJIT

1. MVTX 3-layer hits
2. Clustering
3. Track-seeding
4. Track-finding
5. HF trigger

True associated hits used for initial study;
- Better tracking with MVTX+INTT



Performance - work in progress

Signal:
D -> K + pi

Background:
MB QCD

True_tracklets:
1. 90% BG_rej. Sig_eff = 89%
2. 99% BG_rej. Sig_eff= 41%

Reco'd tracklets:
1. 90% BG rej. Sig_eff=61%
2. 99% BG rej. Sig_eff = 12%

- for gt_track with calculated radius

```
ic| model_mode: 'gt_track'
ic| data_mode: 'gt_track'
/home1/tingtingxuan/physics-trigger-graph-level-prediction/train_results/garnet/experiment_2022-03-22_10:43:39/checkpoints/model_chec
_100.pth.tar
Successfully reloaded!
Loaded 500000 inference samples
{'prec': 0.08458174374243975, 'recall': 0.9139352503519003, 'acc': 0.8902208062303641, 'F1': 0.15483409416093208, 'auroc':
0.9668233993365721}
Trigger: 4973 Non-Trigger: 447007
Input 1.0% Trigger Events      drop_rate: 90.0%      efficiency: 89.66%      purity: 9.87%
Input 1.0% Trigger Events      drop_rate: 95.0%      efficiency: 79.99%      purity: 17.6%
Input 1.0% Trigger Events      drop_rate: 99.0%      efficiency: 41.06%      purity: 45.18%
Input 1.0% Trigger Events      drop_rate: 99.33%     efficiency: 31.39%      purity: 51.79%
```

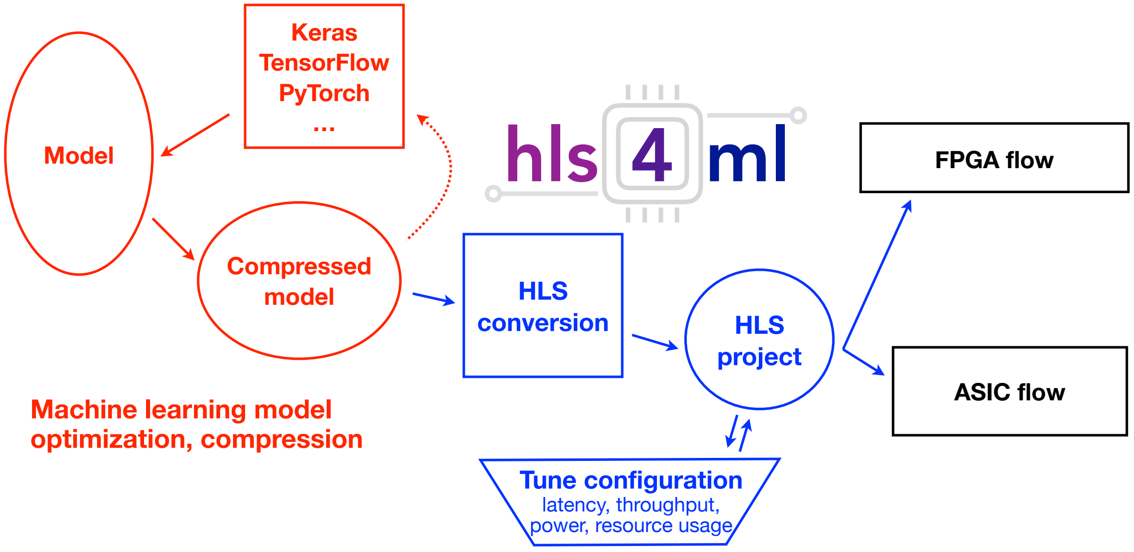
- for predicted_track with calculated radius:

```
ic| model_mode: 'predicted_trk'
ic| data_mode: 'predicted_trk'
/home1/tingtingxuan/physics-trigger-graph-level-prediction/train_results/garnet/experiment_2022-03-21_21:17:13/checkpoints/model_chec
_100.pth.tar
Successfully reloaded!
Loaded 500000 inference samples
{'prec': 0.0388423850859006, 'recall': 0.9033613445378151, 'acc': 0.7592362127343653, 'F1': 0.07448221252587905, 'auroc':
0.8983579838990827}
Trigger: 4998 Non-Trigger: 461048
Input 1.0% Trigger Events   drop_rate: 90.0%    efficiency: 60.82%   purity: 6.52%
Input 1.0% Trigger Events   drop_rate: 95.0%    efficiency: 39.36%   purity: 8.44%
Input 1.0% Trigger Events   drop_rate: 99.0%    efficiency: 12.24%   purity: 13.13%
Input 1.0% Trigger Events   drop_rate: 99.33%   efficiency: 8.86%    purity: 14.26%
```

Translating Models into FPGA Firmware, FNAL/MIT

- Algorithms must have low latency and resource use
  - **5us latency to decide whether acquire event data in TPC**
- **hls4ml** translates NN algorithms into high level synthesis

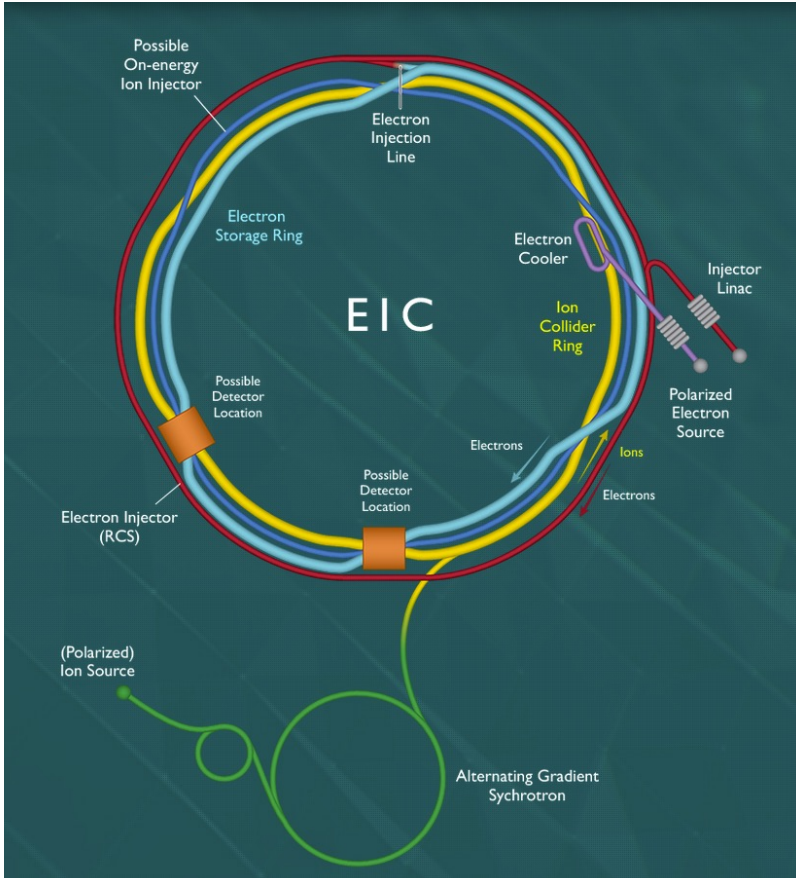
  https://hls-fpga-machine-learning.github.io/hls4ml/
- Also generates IP cores for easy implementation



Red – typical ML algorithm development stages
Blue – HLS conversion to IP
Black – typical implementation onto chips

The Electron-Ion Collider

- Next generation accelerator
  - **To be operating at BNL from the early 2030s**
  - **the future of nucleon structure probes and many other studies**
- Three collaborations have submitted detector proposals:
  1. **ATHENA**
  2. **CORE**
  3. **ECCE (selected)**



[https://www.bnl.gov/eic/]

How this fits in

- sPHENIX takes data from 2023
  - Can be used as a proof-of-principle (as well as a real use case)
- EIC has lower average multiplicity, should be relatively easier to select
- Current thoughts are to use similar tracker technology to MVTX (ITS-2 vs ITS-3)
- Large overlap of team between sPHENIX and EIC, knowledge preservation
- They currently share a simulation framework
  - Work can commence immediately
  - Framework may change in future

Predicted timeline

sPHENIX

2021	2022	2023	2024	2030+
• Project started • Initial simulations constructed • First data for algorithm training	• MVTX & INTT SRO • Fast tracking algorithms in place • GPU feedback machine R&D • Initial FPGA bitstream	• Refine interface between system and detectors • Improve algorithms with latest data stream • Pre-commissioning	• Deploy device at sPHENIX • pp/pA run • Design updated system for EIC • Take advantage of new technology if required	• Deploy device at EIC

Conclusions

- We have successfully received funding for FY 22 and 23
- Project will significantly improve sPHENIX HF capabilities
- Project relies on inner tracker SRO
- After successful deployment at sPHENIX, focus shifts to future EIC detectors
- Great progress has already been achieved

Thank you for your attention

BACKUP SLIDES

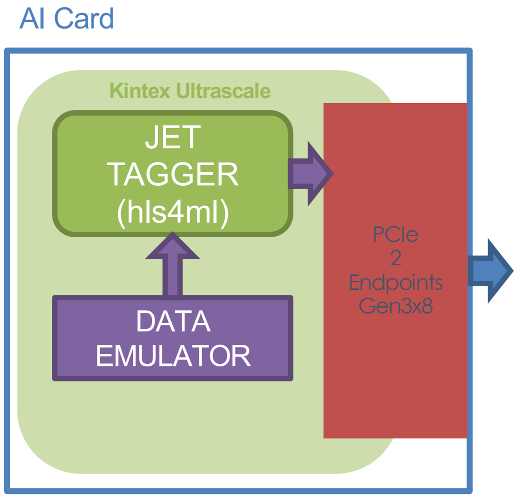
The Team

- LANL (NP)
  - **Yasser Corrales, Cameron Dean, Zhaozhong Shi, Noah Wuerfel, Kun Liu, Cesar da Silva, Hugo Pereira da Costa, Ming Liu … new PDs**
- MIT (NP, HEP)
  - **Gunther Roland, Philip Harris (HLS4ML), Yen-Jie Lee, Or Hen, Cristiano Fanelli et al**
- FNAL(HEP)
  - **Nhan Tran(HLS4ML), Micol Rigatti/Engineer, Yu-Dai Tsai (Theorist, ML) et al**
- NJIT(CS)
  - **Dantong Yu, students**
- ORNL(NP)
  - **Jo Schambach**
- CCNU(EE, NP)
  - **Kai Chen(FELIX), Yaping Wang, students et al**
- UNT (CS)
  - **Fu Song, students + PDs**

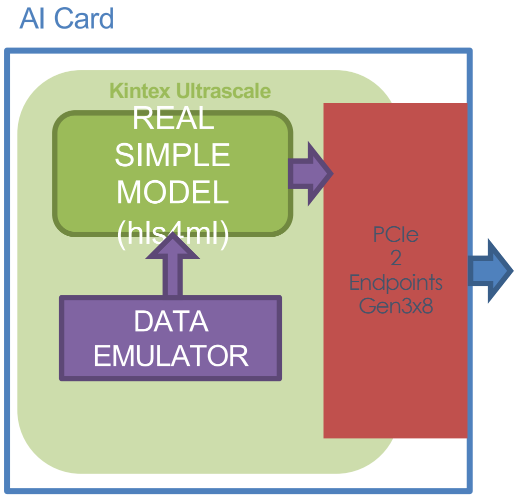
In collaboration with experts from BNL - Jin Huang, Martin Purschke, John Haggerty et al

BACKUP SLIDES

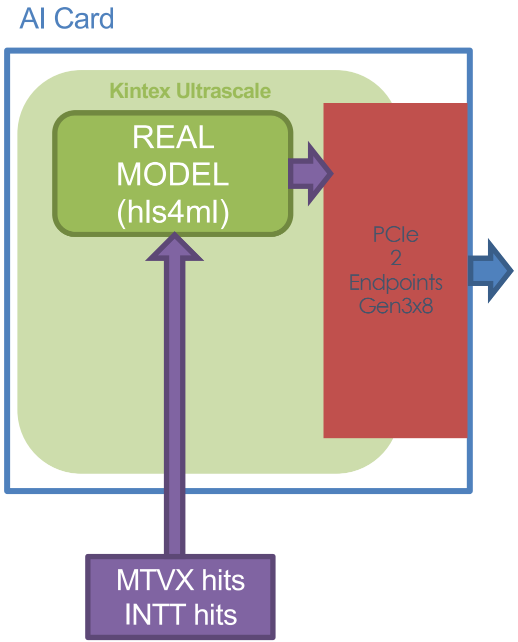
1. Jet Tagger



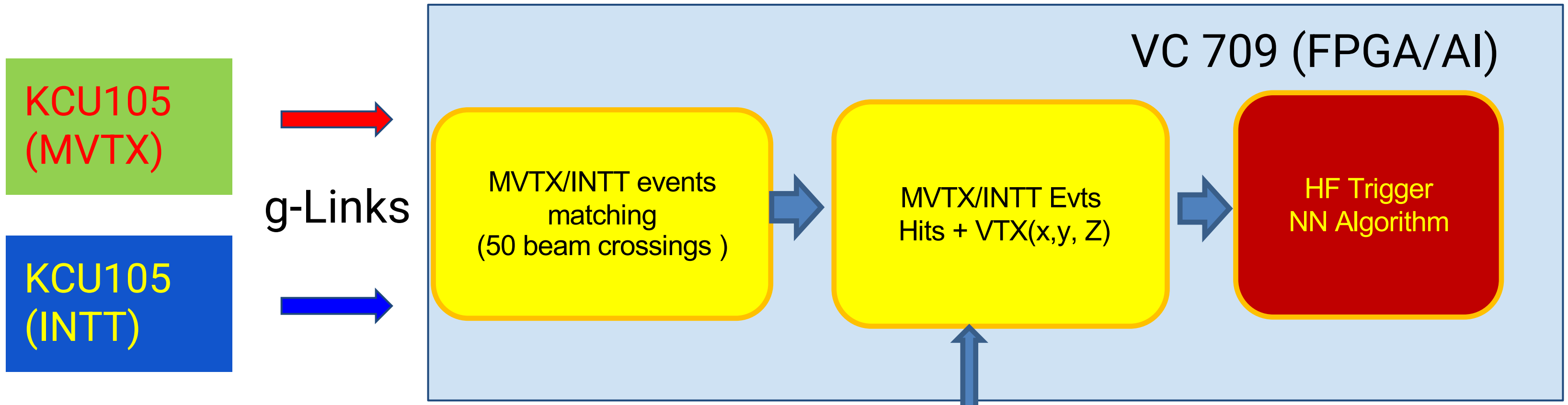
2. Real Simple Model

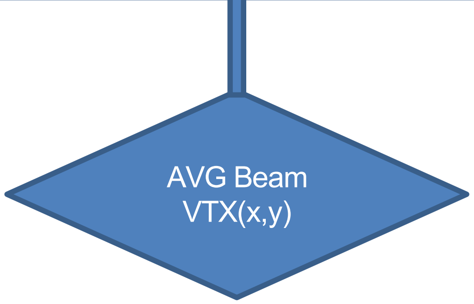


3. Real Model



Feed MVTX/INTT MC Hits to AI Engine

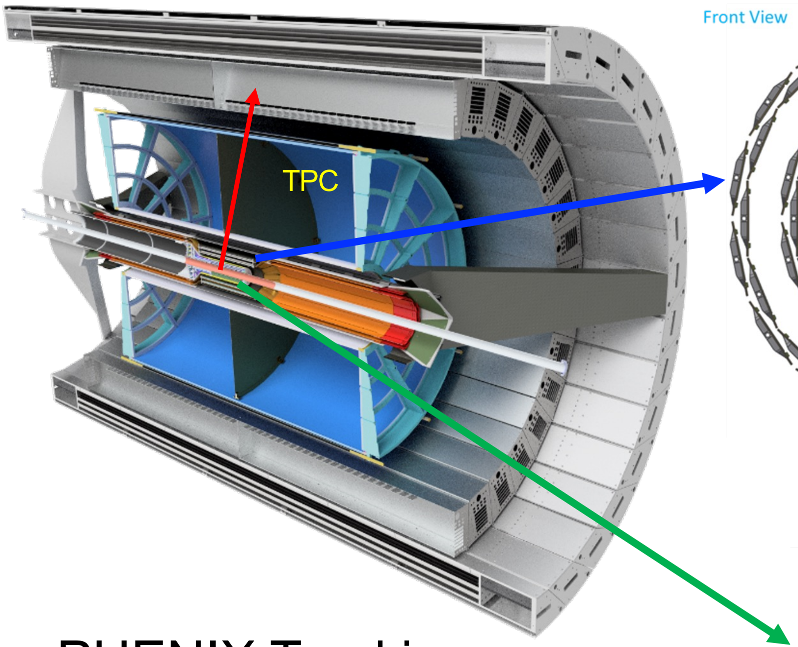




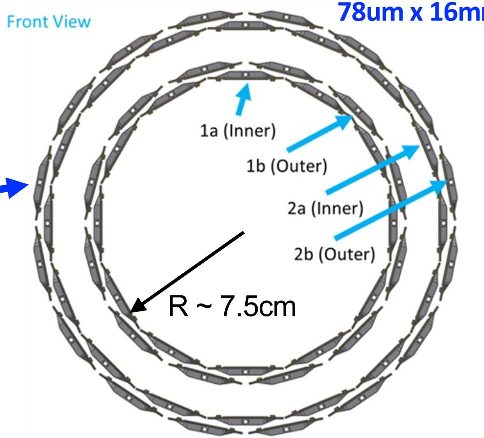
Glinks: 1 MVTX and 50 INTT data packets (~5us worth of data)

Raw MVTX and INTT data packets:
~20 pp collisions (MB events) @4MHz pp
- 1 MVTX packet @5uS strobe:
- 50 INTT packets (RF)
- MIT students working on event-alignment

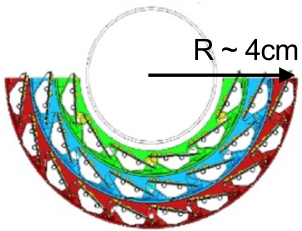
Tag Beauty Events in sPHENIX with MVTX + INTT: 3 + 2 layers

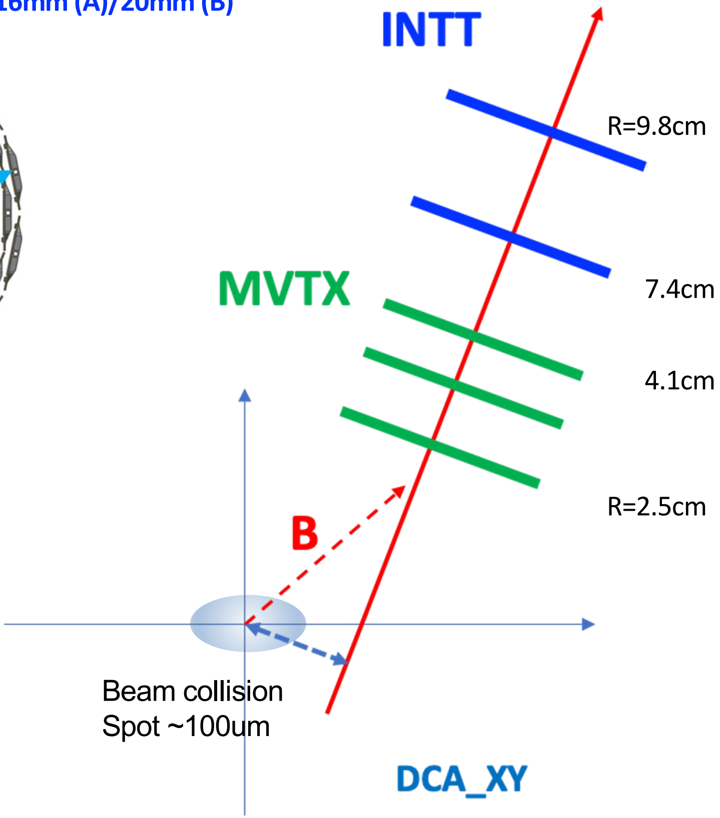


INTT Silicon Strips:
78um x 16mm (A)/20mm (B)



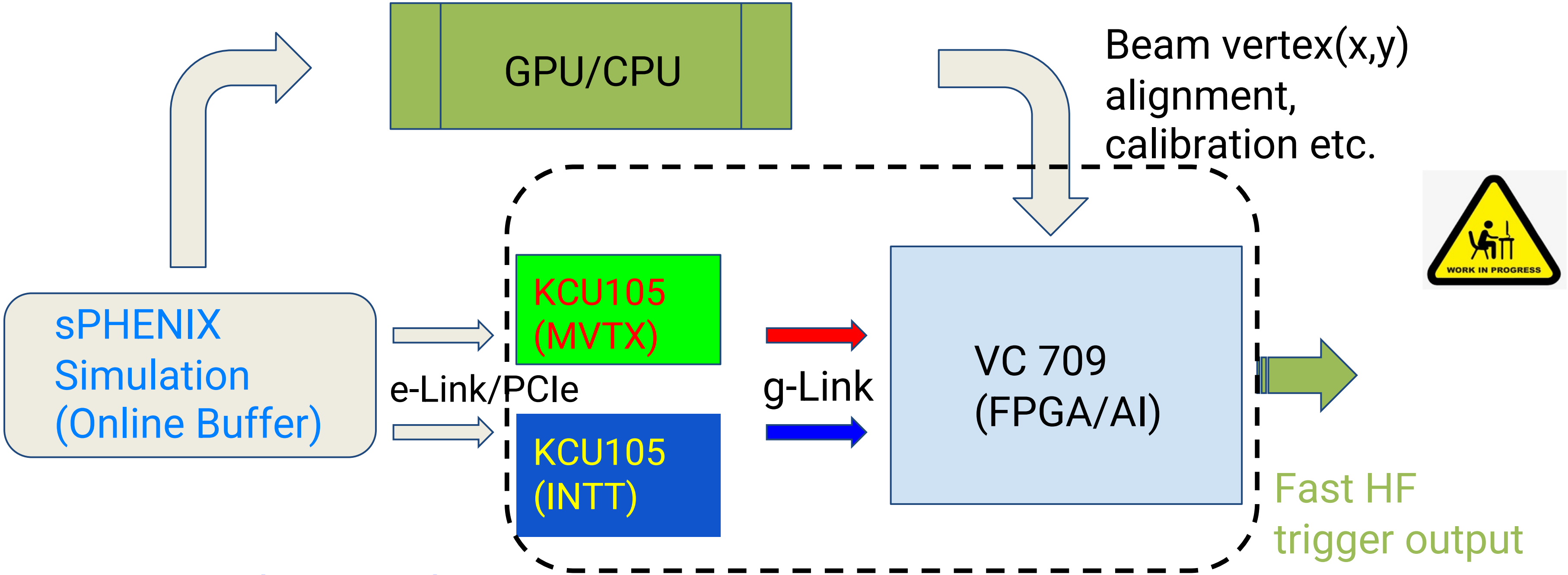
MVTX Silicon pixels:
27um x 29um





sPHENIX Tracking:
MVTX + INTT + TPC(slow)

A Toy Model – Hardware Implementation



Streaming readout sim data:

8b/10b MVTX/INTT data (KCU105) to FPGA/AI Engine (VC709)

Realizing Toy Model in Hardware/Firmware

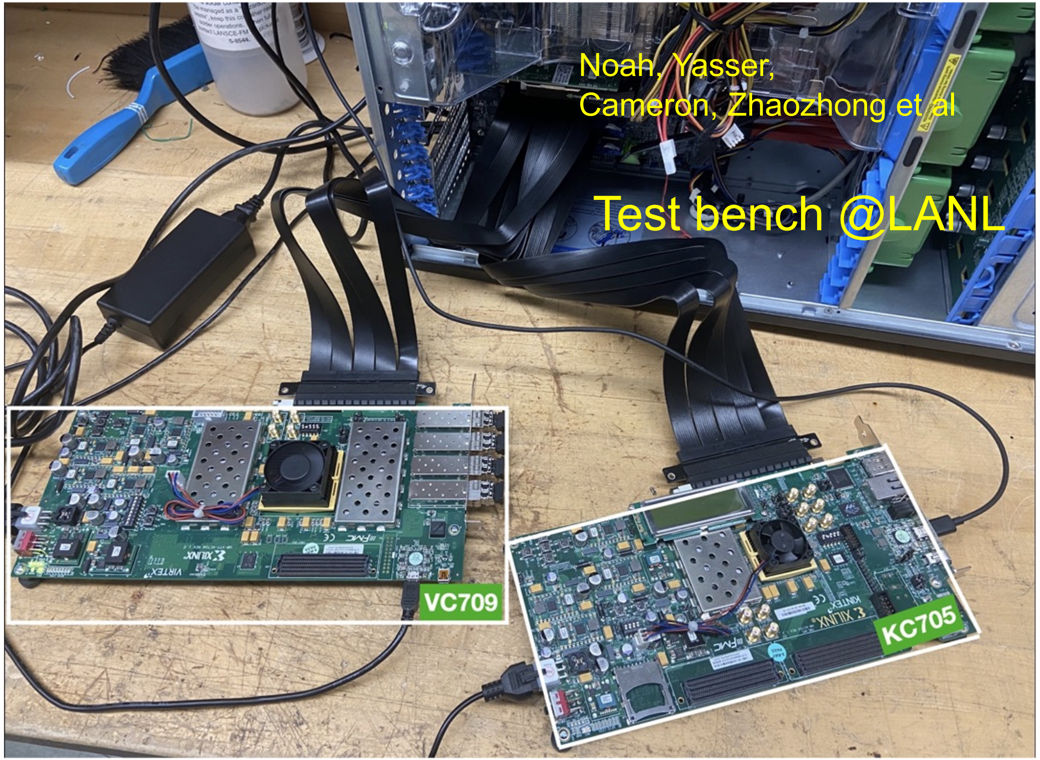
sPHENIX

- VC709 card shares similar FPGA as FELIX, ideal testing ground

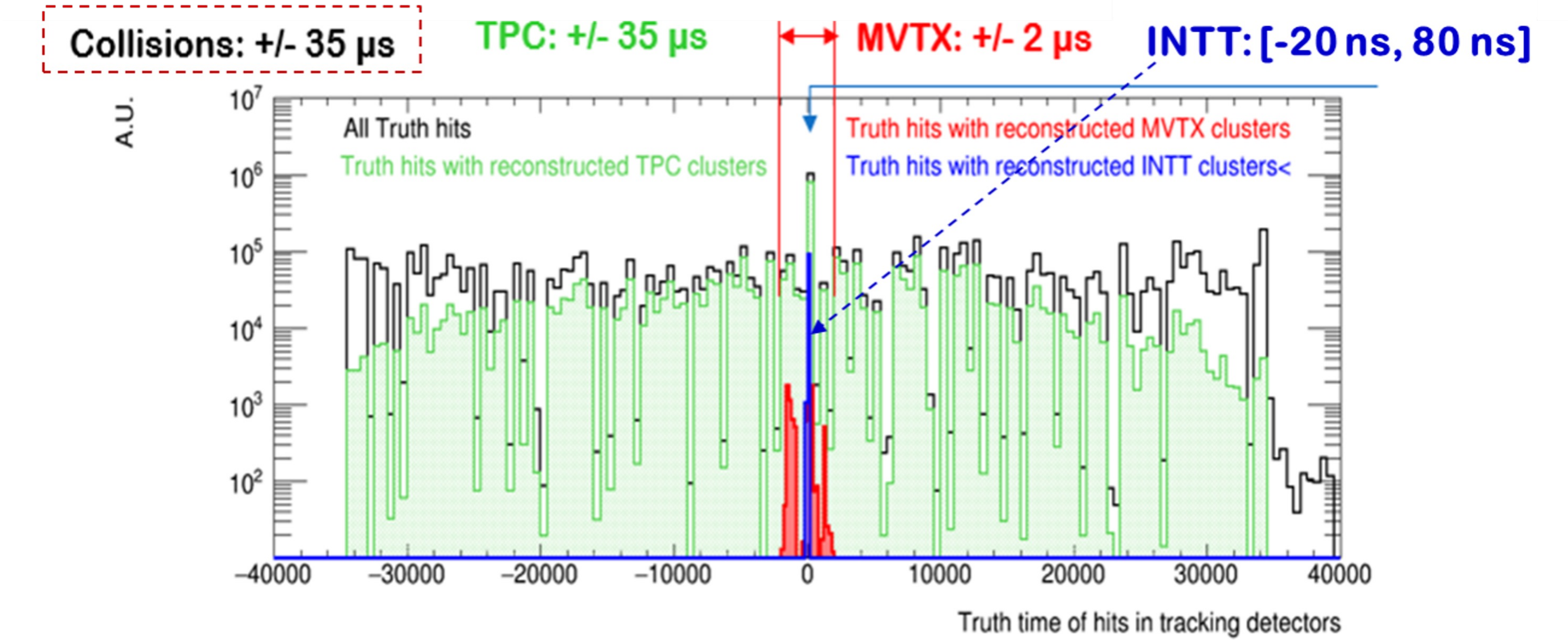
- KC705 represents our MVTX+INTT Data Aggregation Module
  - **Replaced with more powerful KCU105**

- Successfully transmit data from host PC to DMA/FPGA
  - **Convert MVTX sim data to real-data-like bit-stream**
  - **INTT later**
- Next:
  - **Transmit MVTX/INTT sim data to VC709(AI-Engine) through G-Links**

2nd FELIX TB setup at MIT in progress



MVTX and INTT Event Alignment (I)

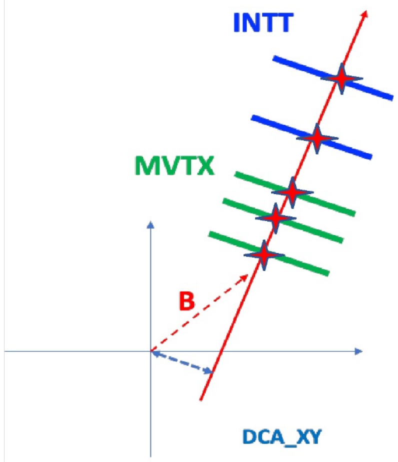


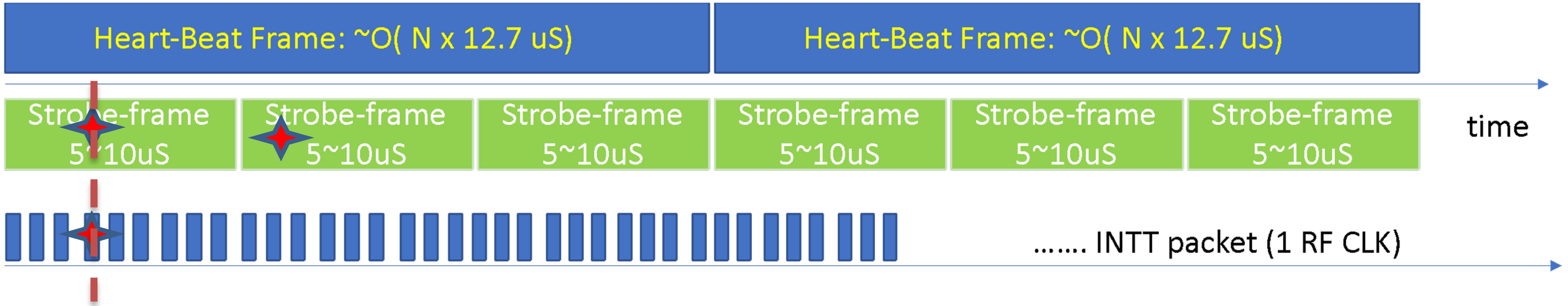
MVTX + INTT Event Alignment (II)

INTT 1-RF packet ⍰ select MVTX hits (track fitting)
To form (MVTX_Hits + INTT_Hits) events for AI-Engine

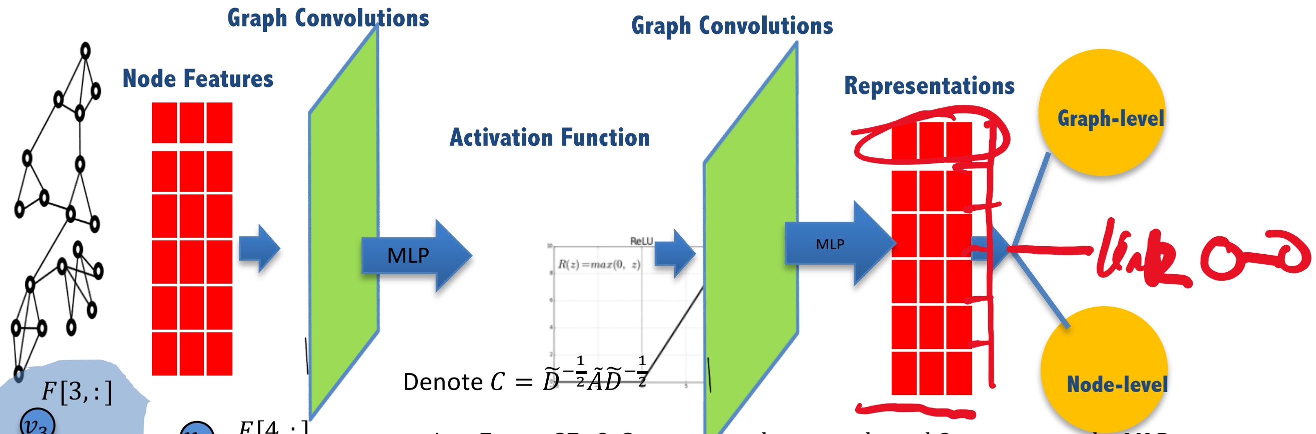
sPHENIX global RF timing (RF counters) used to sync MVTX and INTT packets
- MVTX: @beginning of strobe signal
- INTT: @ a given RF CLK

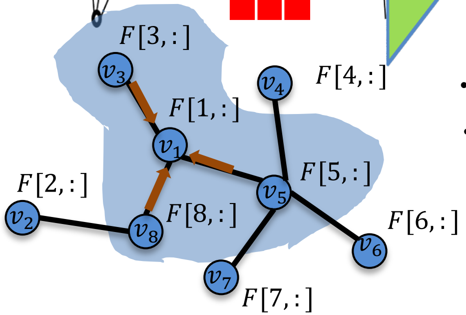
RHIC cycle: 120 RF budgets: 12.7 uS





Graph Neural Networks





Denote $C = \widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}$

- Then $F_{out} = CF_{in}\Theta$, C represents the networks and Θ *represents the MLP*
- For node v_i, $F_{out}[i,:] = \sum_j C[i,j]F_{in}[j,:]\Theta$

Observe that:

$$\mathbf{F}_{out}[i,:] = \underbrace{\sum_{v_j \in \mathcal{N}(v_i)\cup\{v_i\}}}_{\text{Aggregation}} \underbrace{C[i,j]\mathbf{F}_{in}[j,:]\Theta}_{\text{Feature transformation}}$$

- $C[i,j] = 0$ for $v_j \notin N(v_i) \cup \{v_i\}$

Data Pre-Processing-Clustering

- Particles leave a blob of hits in detectors
- The goal of the clustering algorithm is to reduce the amount of data being processed in the machine learning models, in order to improve inference times
- Connected Components.
- Simple Algorithms, challenging to have fast parallel algorithms on FPGA
  - **Grow Clusters by traversing neighbors.**
  - **Zero-Skipping Architecture to find a neighbors**
  - **Achieved a latency of 15,000 cycles on average, or 150 us (@ 100MHz clock)**
  - **Resource utilization of 15%**
  - **Reduction of data by 85%**

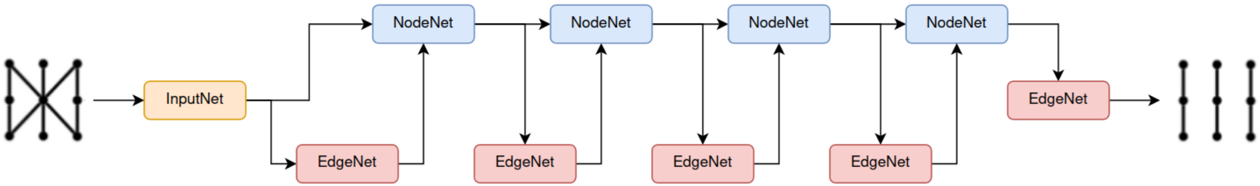
0	0	0	0	0	0	0
1	0	0	0	1	1	1
1	1	1	0	0	1	0
1	1	0	0	0	1	0
0	0	0	1	0	1	0
0	0	1	1	0	0	0
0	0	0	0	0	1	1

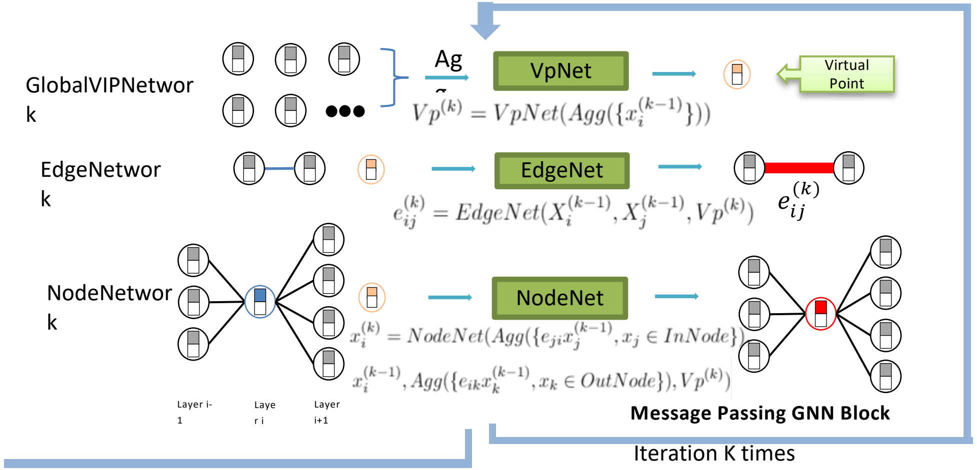
(a)

0	0	0	0	0	0	0
1	0	0	0	2	2	2
1	1	1	0	0	2	0
1	1	0	0	0	2	0
0	0	0	3	0	2	0
0	0	3	3	0	0	0
0	0	0	0	0	4	4

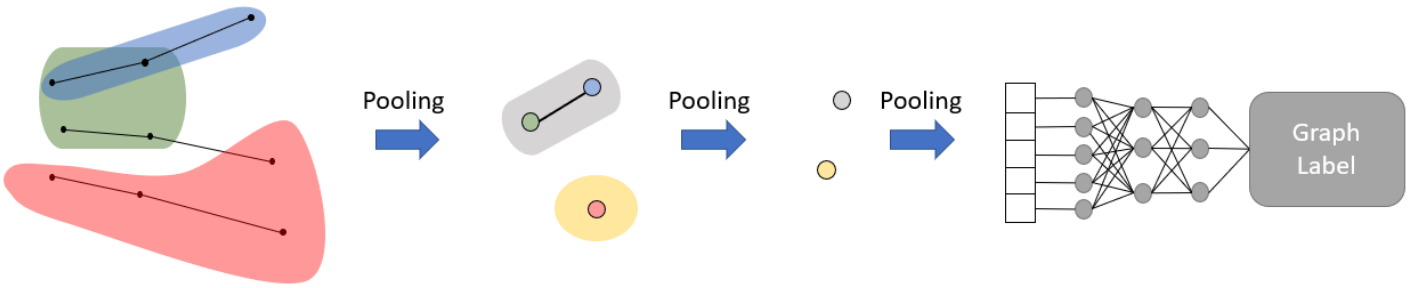
(b)

Tracking Algorithm





Trigger Detection



- At each layer l, the soft cluster assignment matrix $S^l \in R^{n_{l-1} \times n_l}$
- $GNN^l(A, X) = Relu(\widehat{D}^{-\frac{1}{2}} \hat{A} \widehat{D}^{-\frac{1}{2}} X\ W)$
- $S^l = Softmax(GNN_{pool}^l(A^{l-1}, X^{l-1})$
- $X^l = {S^l}^{\top}\ (GNN_{diffuse}^l(A^{l-1}, X^{l-1})$
- $A^l = {S^l}^{\top}\ A^{l-1} S^l$
- Trigger = MLP (Readout(X^L))