Intelligent Experiment Through Real-Time AI: Fast Data Processing and Autonomous Detector Control for sPHENIX and Future EIC Detectors

Ming Liu
Los Alamos National Lab
For the Fast-ML Team

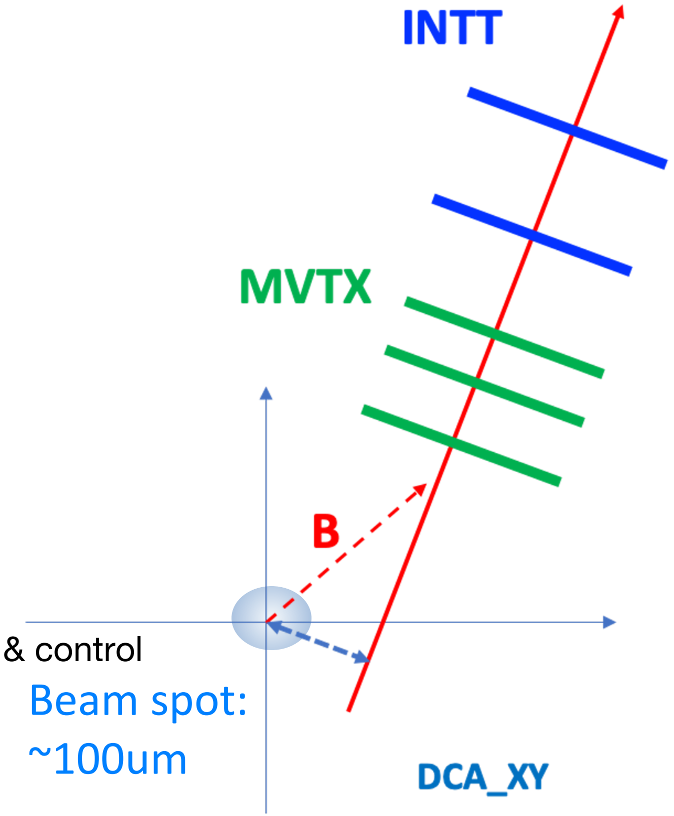
sPHENIX Collaboration Meeting
Jan. 5-7, 2022

The Problem
– Rare events hard to trigger by conventional methods

sPHENIX challenge p+p@200GeV:
- Very high p+p collision rate: ~10MHz
  - **Charm production rate: ~100kHz**
    - **0.5mb/42mb ~1%**
  - **Beauty production rate: ~ 500Hz**
    - **2 ub/42mb ~ 0.005%**
- *No effective trigger to select low pT HF events*
  - **Triggered MB rate ~1kHz << 10MHz**
    - **Lost most of the HF events at low pT**
  - **High pT jet trigger, pT > 10GeV**
  - **Streaming readout -> huge data volume, DAQ/tape cost**

EIC challenge:
- high rate e+p and e+A collisions

Our approach:
Develop effective HF (or other rare events) triggers for p+p and e+p
- Streaming readout key detectors for high efficiency
- AI-based beam/detector monitoring and autonomous feedback & control
- ML-trained algorithm for HF tagging
  - **Rare physics don't require "slow detectors" for measurements**



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 NTU 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

Our 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 fast FPGA-based trigger system
  - **Low trigger decision latency ~10us**
- 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**

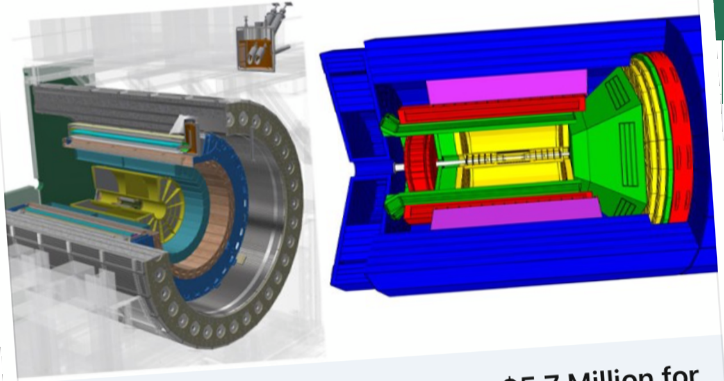
DOE Awards Announced 12/2/2021

sPHENIX

$1.5M for our proposal
FY22-23

Brookhaven National Laboratory
26,734 followers
3w •

New funding from the U.S. Department of Energy (DOE) will support artificial intelligence advancements at #RHIC and the #ElectronIonCollider.



Department of Energy Announces $5.7 Million for Research on Artificial Intelligence and Machine...
bnl.gov • 2 min read

Office of Science

Department of Energy Announces $5.7 Million for Research on Artificial Intelligence and Machine Learning (AI/ML) for Nuclear Physics Accelerators and Detectors

DECEMBER 2, 2021

Office of Science »
Department of Energy Announces $5.7 Million for Research on Artificial Intelligence and Machine Learning (AI/ML) for Nuclear Physics Accelerators and Detectors

Projects will advance understanding of atomic structure and the nature of matter and antimatter

WASHINGTON, D.C. - Today, the **U.S. Department of Energy (DOE)** announced $5.7 million for six projects that will implement artificial intelligence methods to accelerate scientific discovery in nuclear physics research. The projects aim to optimize the overall performance of complex accelerator and detector systems for nuclear physics using advanced computational methods.

"Artificial intelligence has the potential to shorten the timeline for experimental discovery in nuclear physics," said **Timothy Hallman, DOE Associate Director of Science for Nuclear Physics**. "Particle accelerator facilities and nuclear physics instrumentation face a variety of technical challenges in simulations, control, data acquisition, and analysis that artificial intelligence holds promise to address."

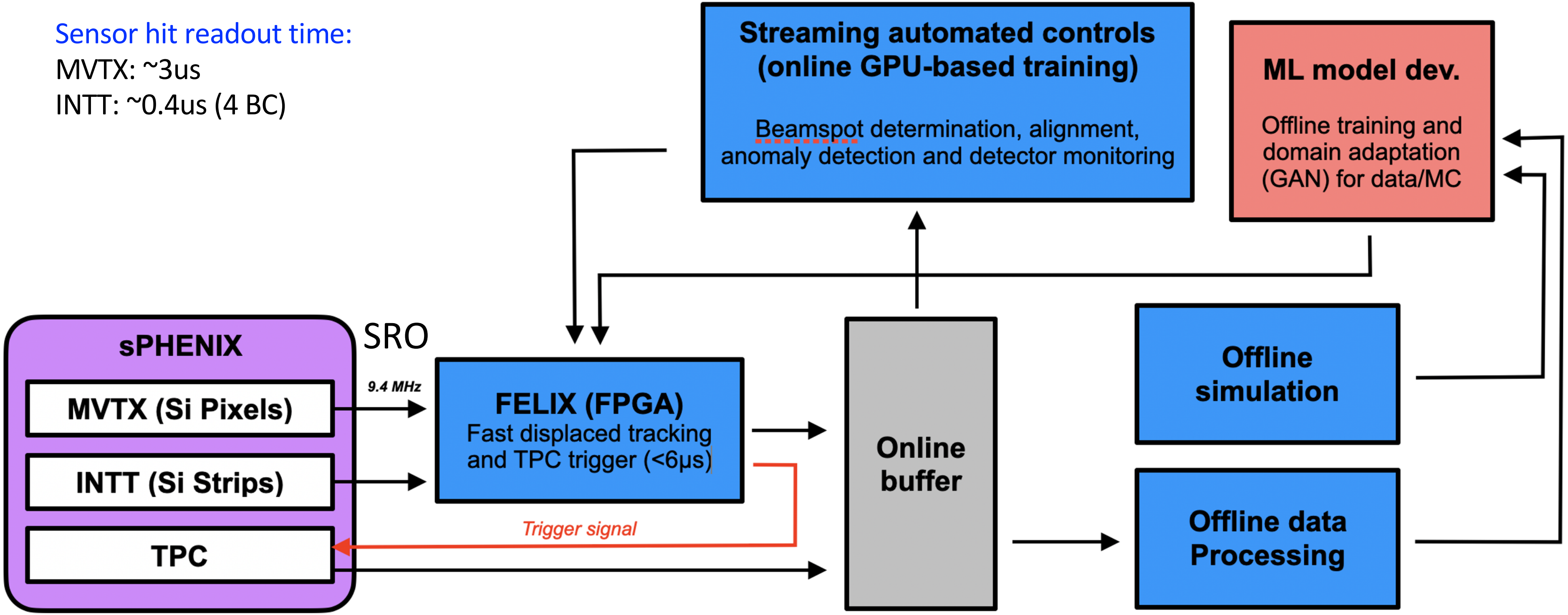
The Team

- 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), Engineer, Yu-Dai Tsai (Theorist, ML) et al**
- NJIT(CS)
  - **Dantong Yu, students + PDs**
- ORNL(NP)
  - **Jo Schambach**
- CCNU(EE, NP)
  - **Kai Chen(FELIX), Yaping Wang et al**
- NTU (CS)
  - **Fu Song, students + PDs**

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

HF AI Trigger: sPHENIX as a Test Ground

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



sPHENIX DAQ & Trigger integration challenge

Timeline

2021
- Project started
- Initial simulations constructed
- First data for algorithm training

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

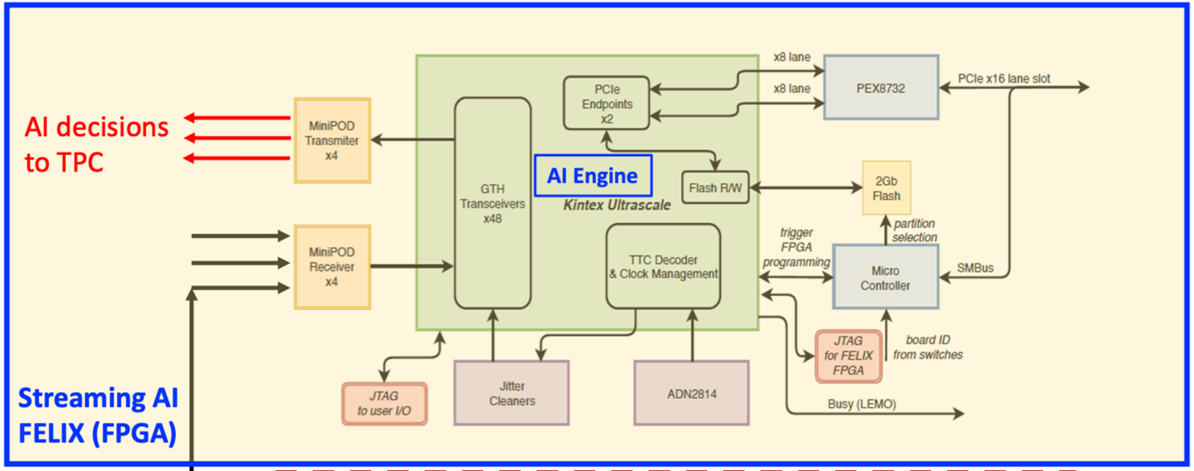
2023
- Refine interface between system and detectors
- Improve algorithms with latest data stream
- Pre-commissioning

2024
- Deploy device at sPHENIX
- pp/pA run

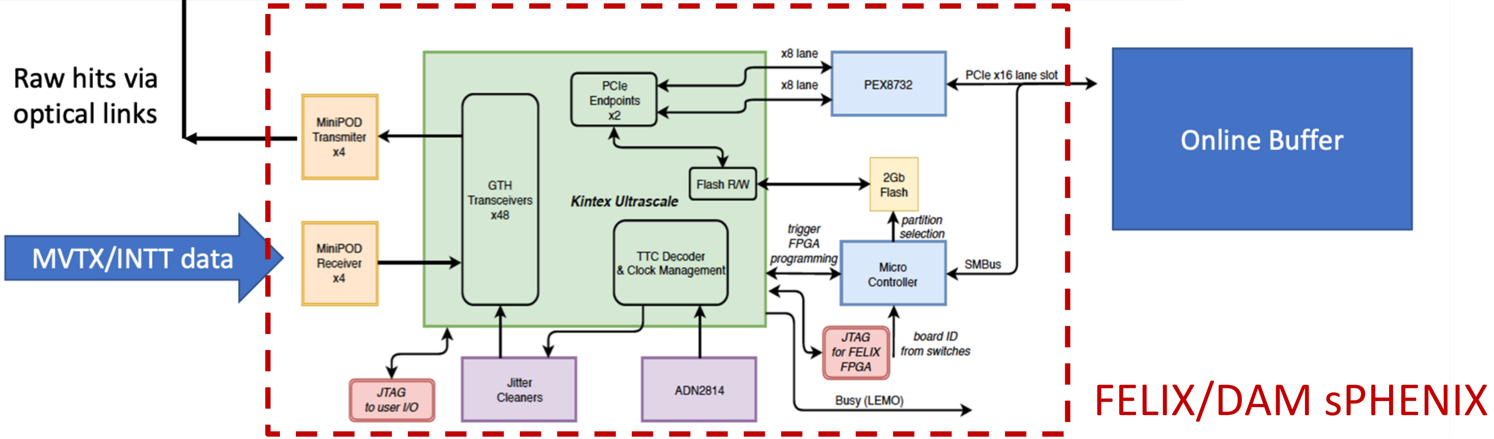
- Design updated system for EIC
- Take advantage of new technology if required

2030+
- Deploy device at EIC

A Prototype Implementation Proposal



Trigger boards, TBD
- FELIX
- CMS trigger boards
- Other options



Physics and Detector Simulations

(Pythia + GEANT) → MVTX/INTT hit maps -> raw data (pixel hits) in JSON

1) trigger AI-ML training
2) generate raw real data like electronics bit stream for hardware simulations

Cameron, Zhaozhong, Jin,
Yasser, Noah et al

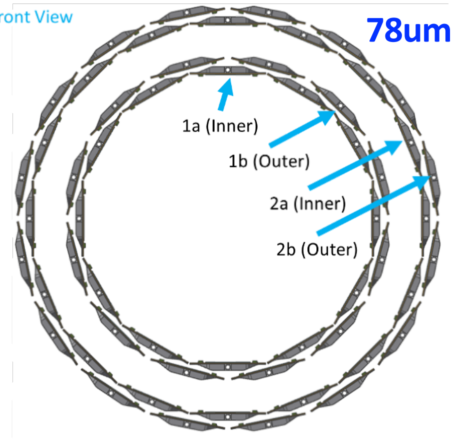
MVTX + INTT: 3 + 2 layers

INTT

Barrel	Center of Sensor Tangent Radius (mm)
1	-
1a (Inner)	71.88
1b (Outer)	77.32
2	-
2a (Inner)	96.80
2b (Outer)	102.62

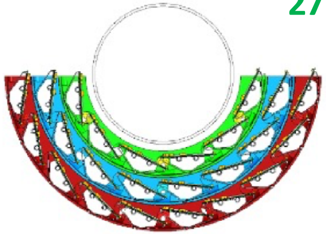
MVTX

R (mm)	min	mid	max
Layer 0	24.61	25.23	27.93
Layer 1	31.98	33.35	36.25
Layer 2	39.93	41.48	44.26

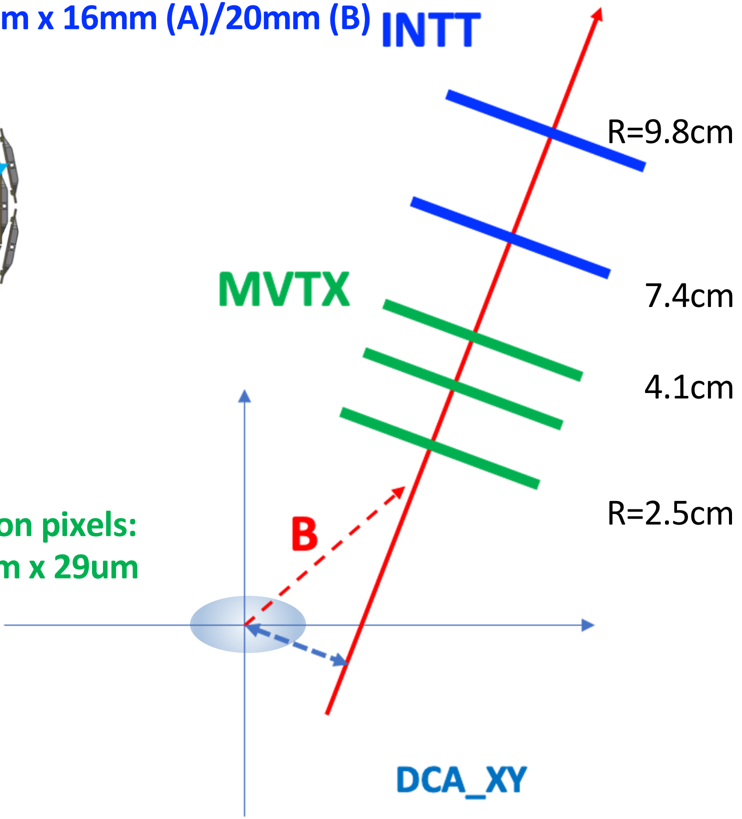


Silicon Strips:
78um x 16mm (A)/20mm (B)

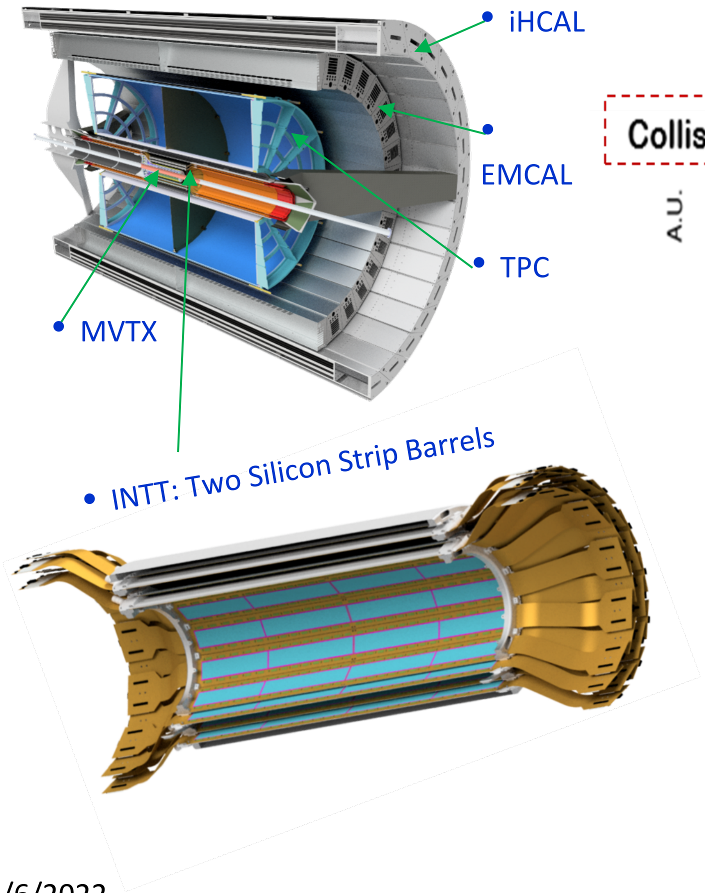
3-layer sensor barrel
- 48 staves, 432 chips

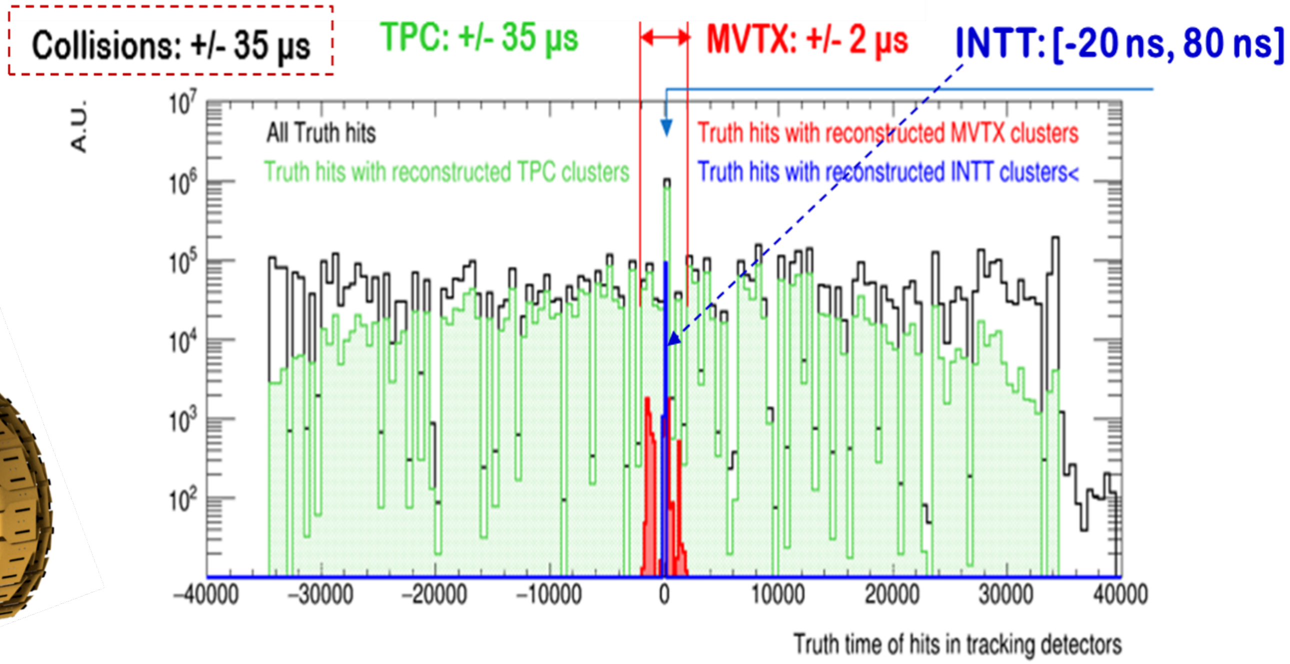


Silicon pixels:
27um x 29um



Event Timing: Reject out of time MVTX hits with INTT

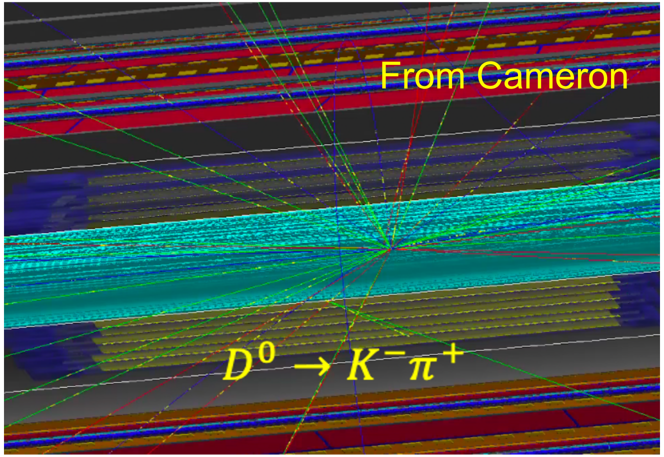


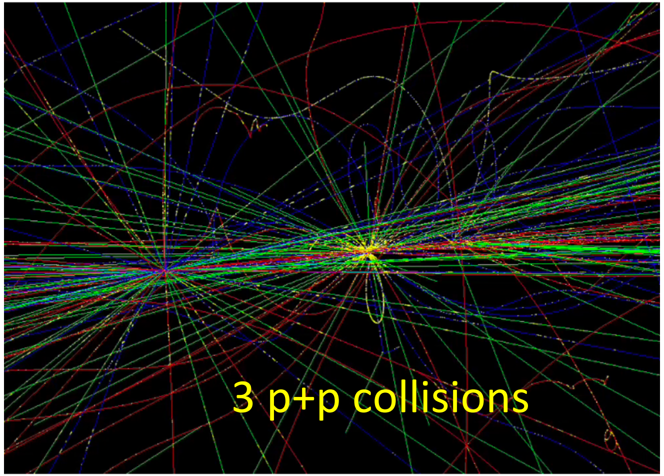


Detector/Physics Simulations (sPHENIX)

- Can simulate any number of signal and background events with full digitization

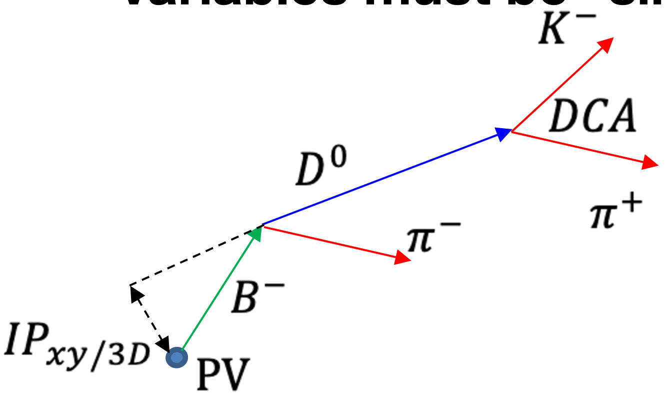
- Package developed to extract raw hit information, used for
  - **algorithm training (JSON output)**
  - **sim data to raw data bit pattern**

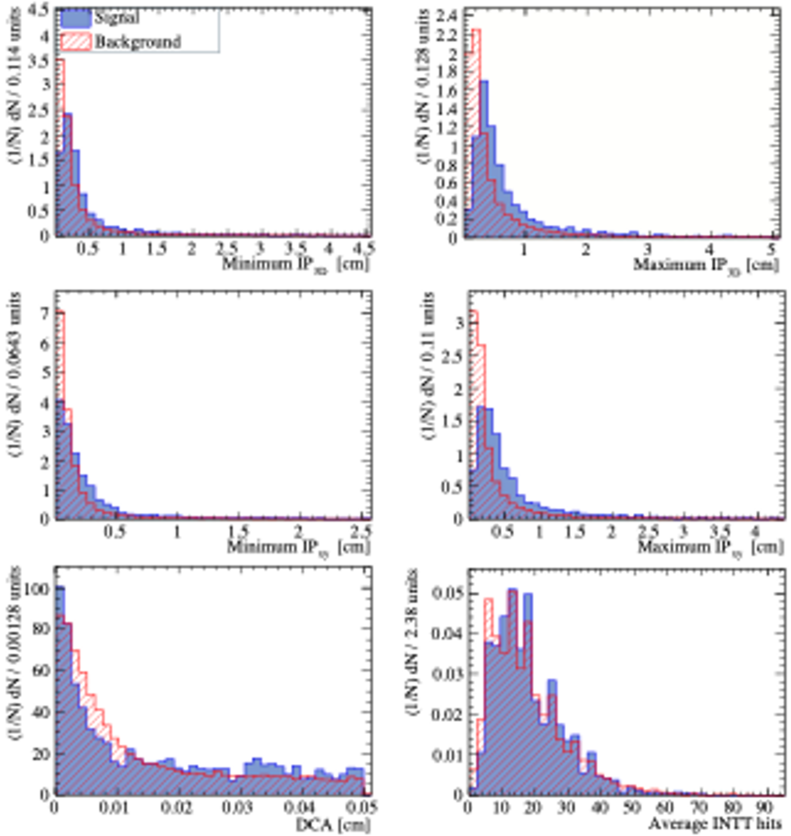




Case study: AI HF selections

- Is ML better for selecting HF decays over conventional selections?
- Challenge:
  - **Must run online, in FPGA. Hence variables must be "simple"**

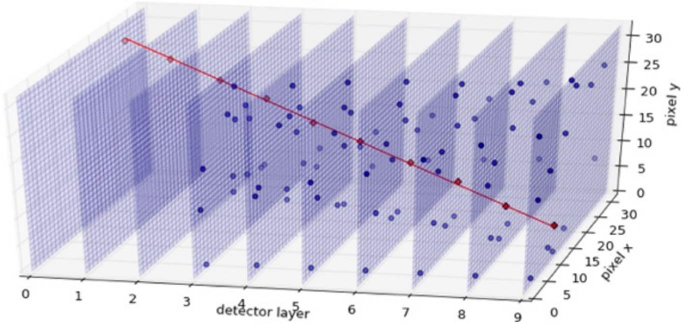


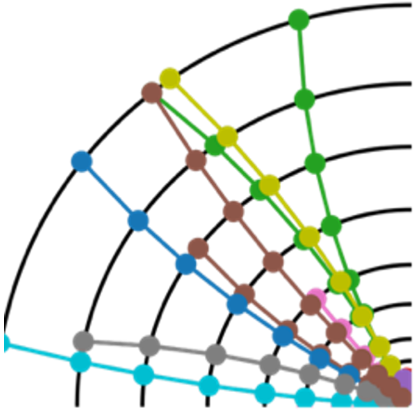


Event Data Descriptions – NJIT/LANL

Moving from images to points

- **Image-based methods face challenges scaling up to realistic HL-LHC conditions**
  - **High dimensionality ($1\text{k} * 0.5K * 9$ per MVTX stave alone) and sparsity**
  - **Irregular detector geometry**
- **Instead of forcing the data into an image, use the space point representation**
  - **Harder to design models (variable-sized inputs/outputs, MVTX + INTT)**
  - **But now we can exploit the structure of the data with full precision**
- What ML models are appropriate for the event
  - **Recurrent Neural Networks and Graph Neural Networks**





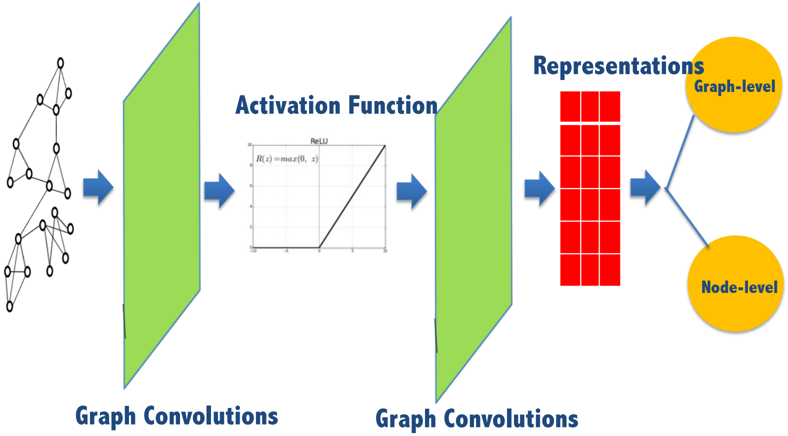
Trigger Algorithm R&D – NJIT/LANL

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:
-hits
-tracklets



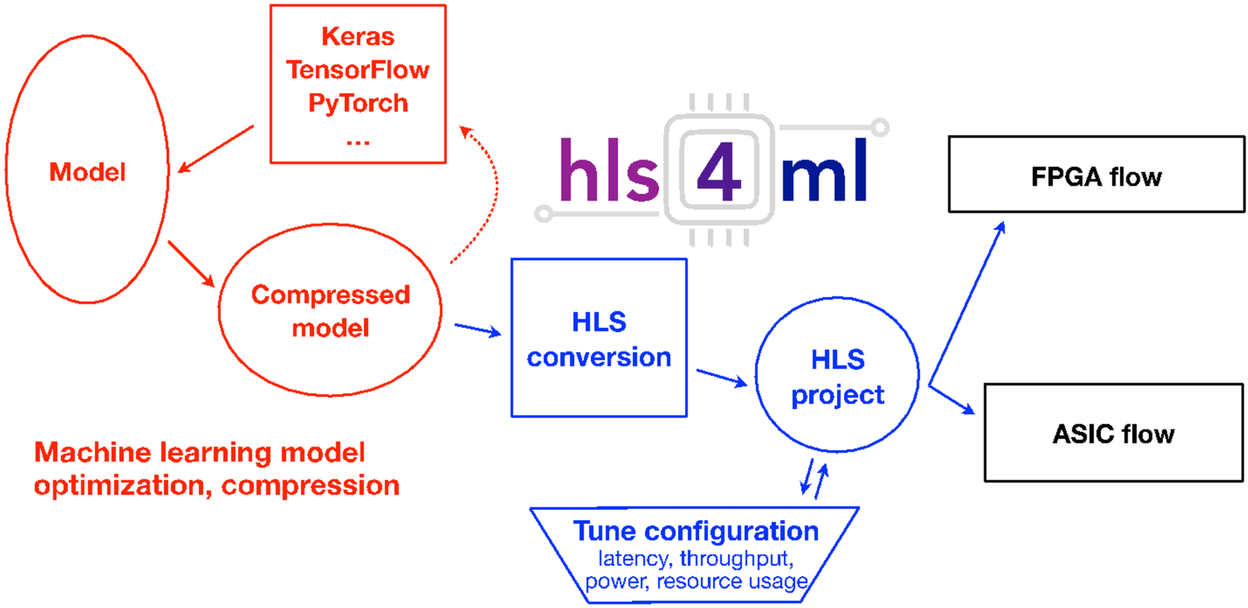
Initial Study from NJIT: Efficiency and Purity

Proper mix of simulated data: 100(BG) + 1 HF(charm) events

	Current performance		Goal 1	Goal 2
Efficiency	50%	25%	20-50%	90%
Purity	5%	5%	5%	5%
Background Rejection	90%	95%	99%	95%

Translating Models to FPGA Firmware, FNAL/MIT

- Algorithms must have low latency and resource usage
- **hls4ml** translates NN algorithms into high level synthesis
- Also generates IP cores for easy implementation



Red – typical ML algorithm development stages, Python/C++

Blue – HLS conversion to FPGA IP

Black – typical implementation onto chips

AI-Trigger SW/FW Pipeline - Status

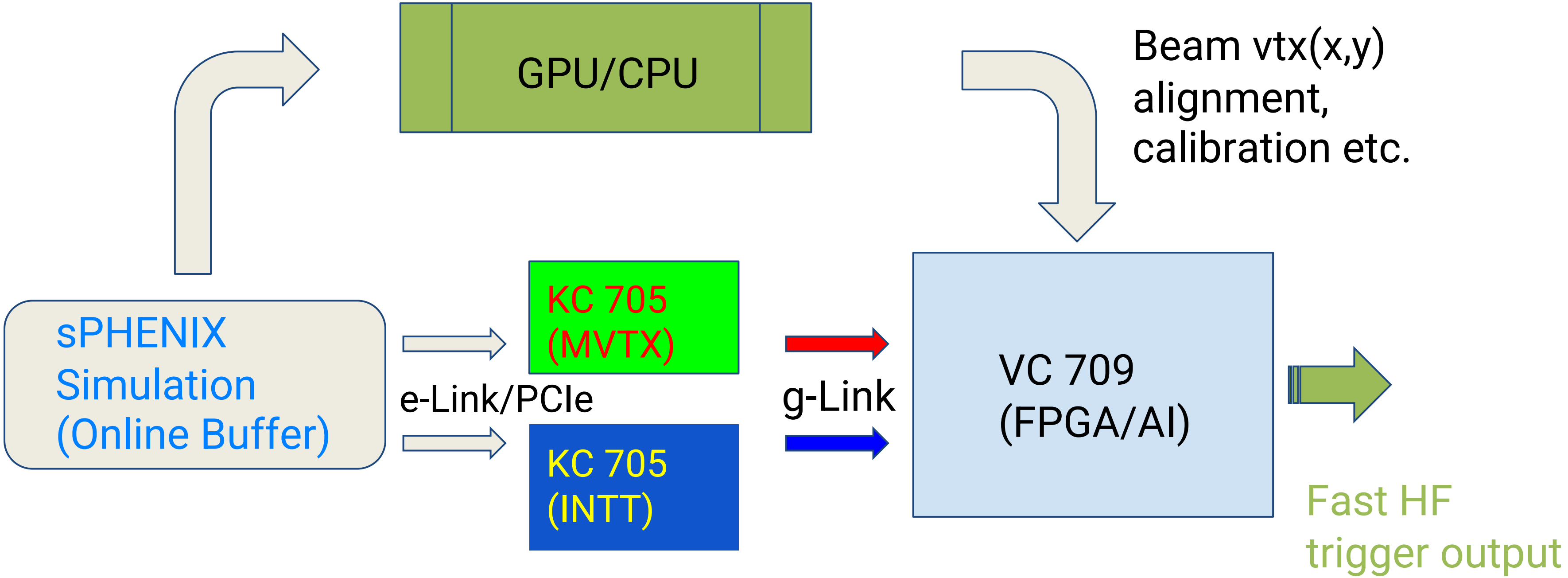
1. Fetch events from event buffer (Work in Progress, sim to raw data)
2. Data Pre-processing Clustering (Work in Progress on FPGA implements)
3. Tracking + Outlier hits Removal (Done in FPGA)
4. Triggering (Done in FPGA, need performance tuning)
5. Triggers on TPC (Interface and integration with sPhenix Detector, last step)

Hardware and Computing

Hardware implementation

1. data stream processing
2. Test AI algorithms on FPGA

A Toy Model – Hardware Implementation (sPHENIX)

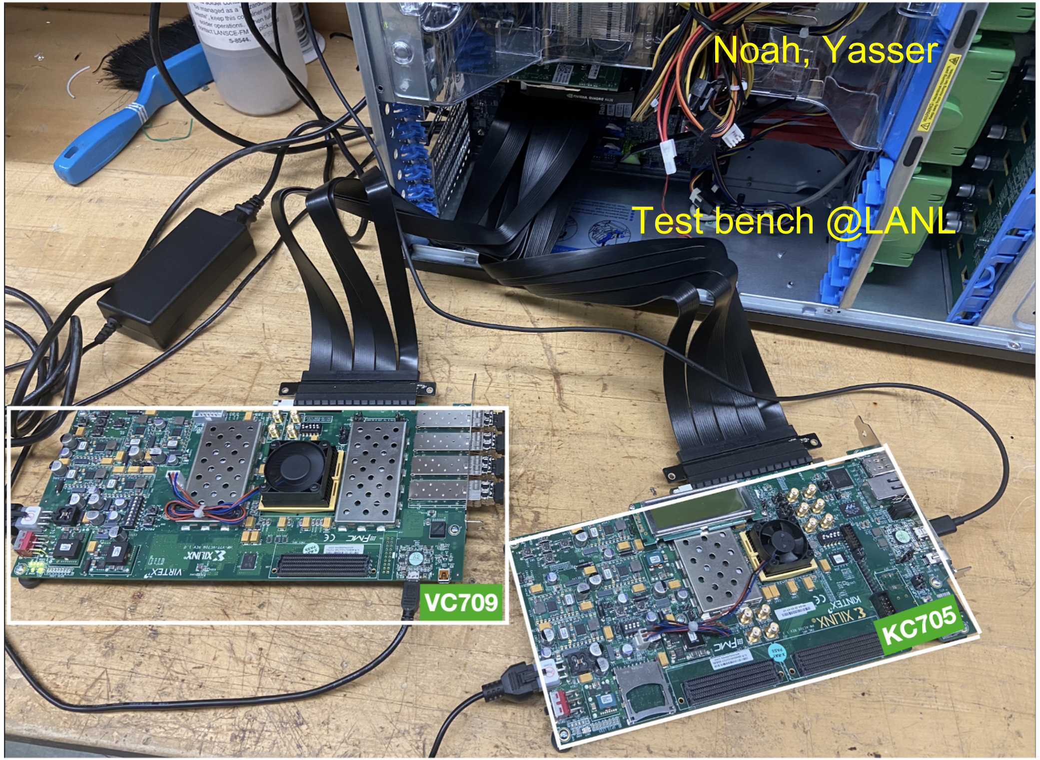


Streaming readout sim data:

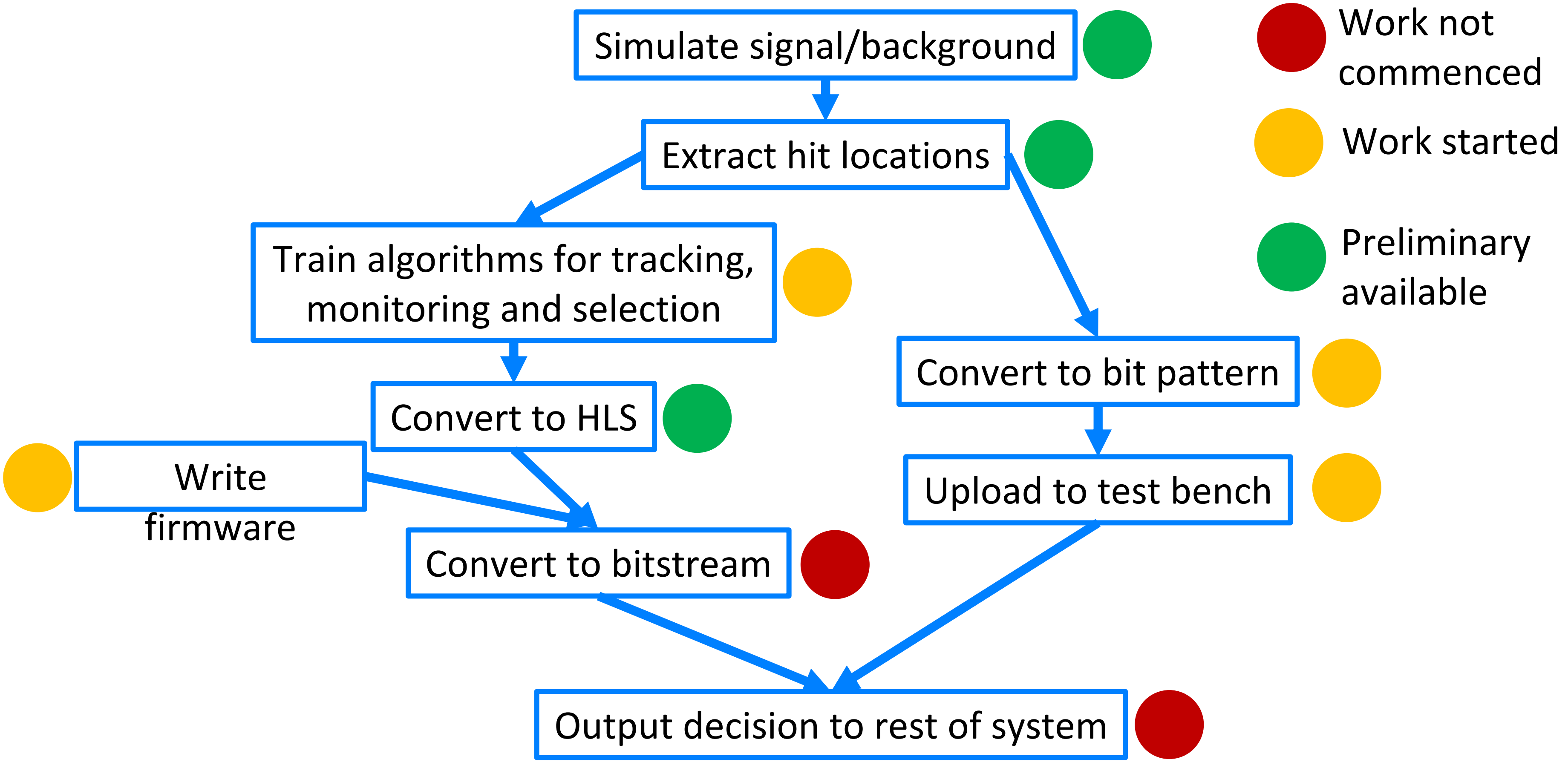
8b/10b MVTX/INTT data (KC705) to FPGA/AI Engine (VC709)

Realizing in Firmware

- FELIX card shares same FPGA as Xilinx VC709, ideal testing ground
- KC705 represents our MVTX+INTT Data Aggregation Module
- Successfully transmit data from host PC to DMA
  - **Convert MVTX sim data to real-data-like bit-stream in progress**
- Next:
  - **Transmit MVTX/INTT sim data to VC709(AI-Engine) through G-Links**



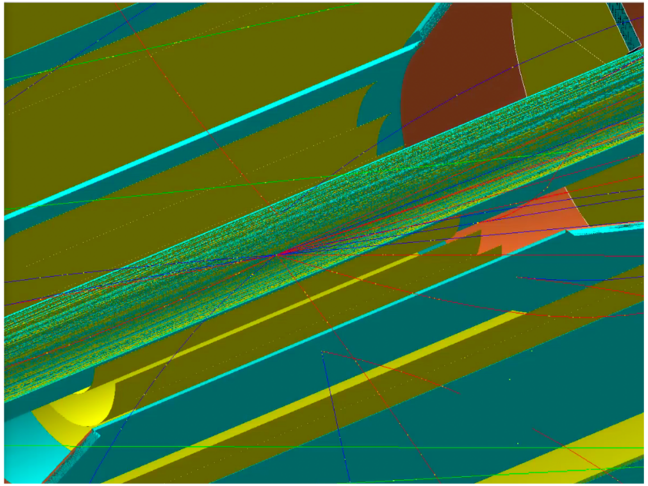
Project Workflow Status

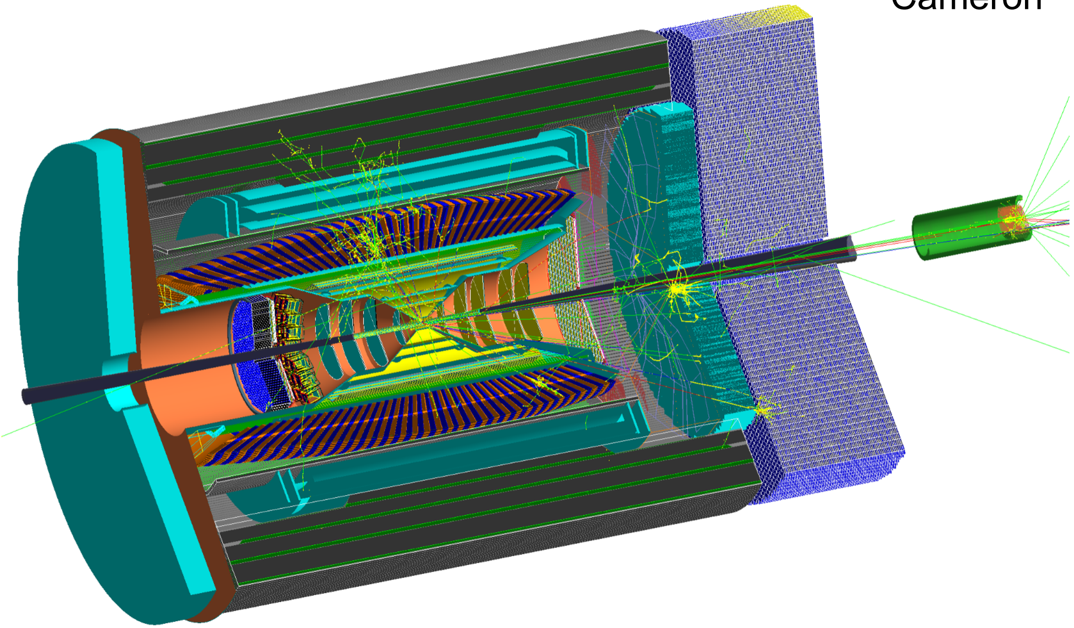


EIC: Simulating events (ECCE)

- EIC physics simulations progressed rapidly in 2021
- Large volume of data already at hand (>800M events)
- No digitization yet but we can use smeared hits to understand potential

Cameron





Summary of Progress

Success stories since proposal approved

1. Full Geant4 simulations of MVTX and INTT plus Geant4 simulation of EIC detectors
2. Tracking GNN algorithms are being developed at NJIT
3. Prototype hardware set up at LANL with host-to-client transfers running
4. Second lab (FELIX) being set up at MIT
5. HLS4ML development at Fermilab, MIT and NJIT
6. FELIX FW development at ORNL and LANL

Outlook

- Next several months:
  1. **Convert simulation output to equivalent bit pattern through G-Link**
  2. **Develop initial tracking and selection algorithms**
  3. **Convert algorithms to HLS code to run on FPGA**
  4. **Pass simulated data to FPGA as if it were real data**

- Goals:
  1. **Build a full prototype and benchmark performance with simulations by 2023;**
  2. **Install device in sPHENIX before 2024 (RHIC *pp* run)**

- Project could significantly improve sPHENIX HF capabilities
  - **Project relies on inner tracker MVTX and INTT SRO**
- After successful deployment at sPHENIX, focus shifts to future EIC detectors

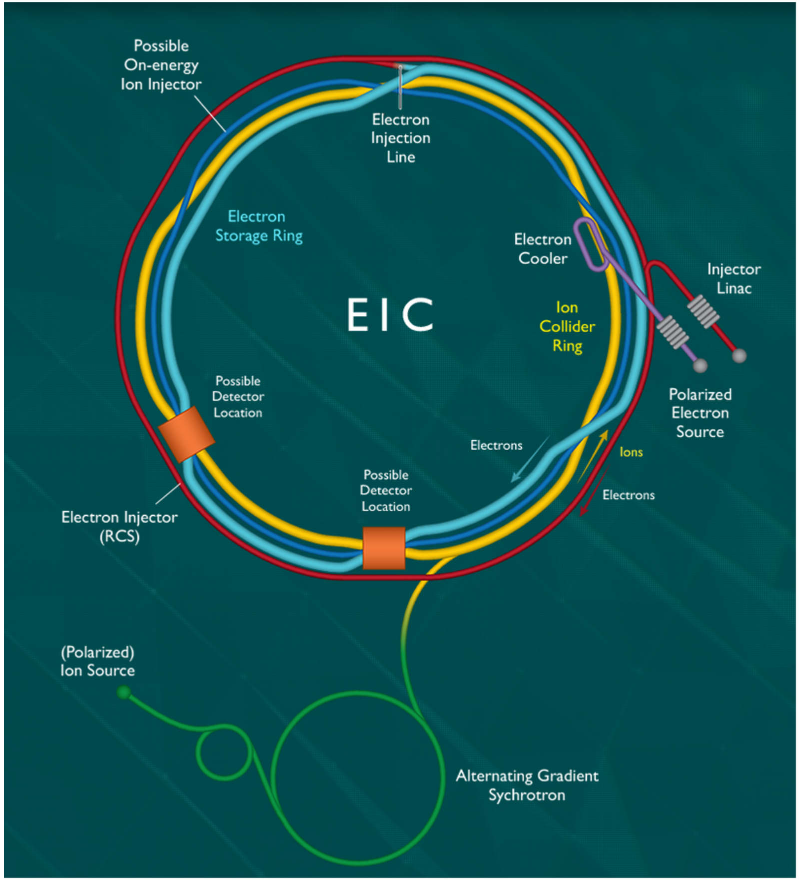
Backup

from sPHENIX to EIC

- 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
  - **relatively easier to select**
  - **likely to use similar tracker technology to MVTX (ITS-2 vs ITS-3)**
- Large overlap of team between sPHENIX and EIC/ECCE
  - **knowledge preservation**
  - **share a simulation framework**

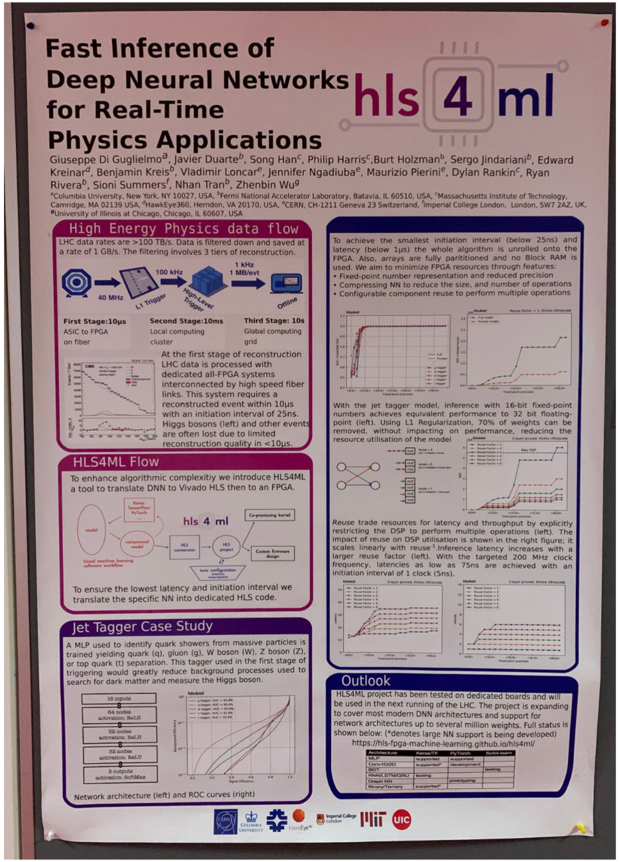
The Electron-Ion Collider

- 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**

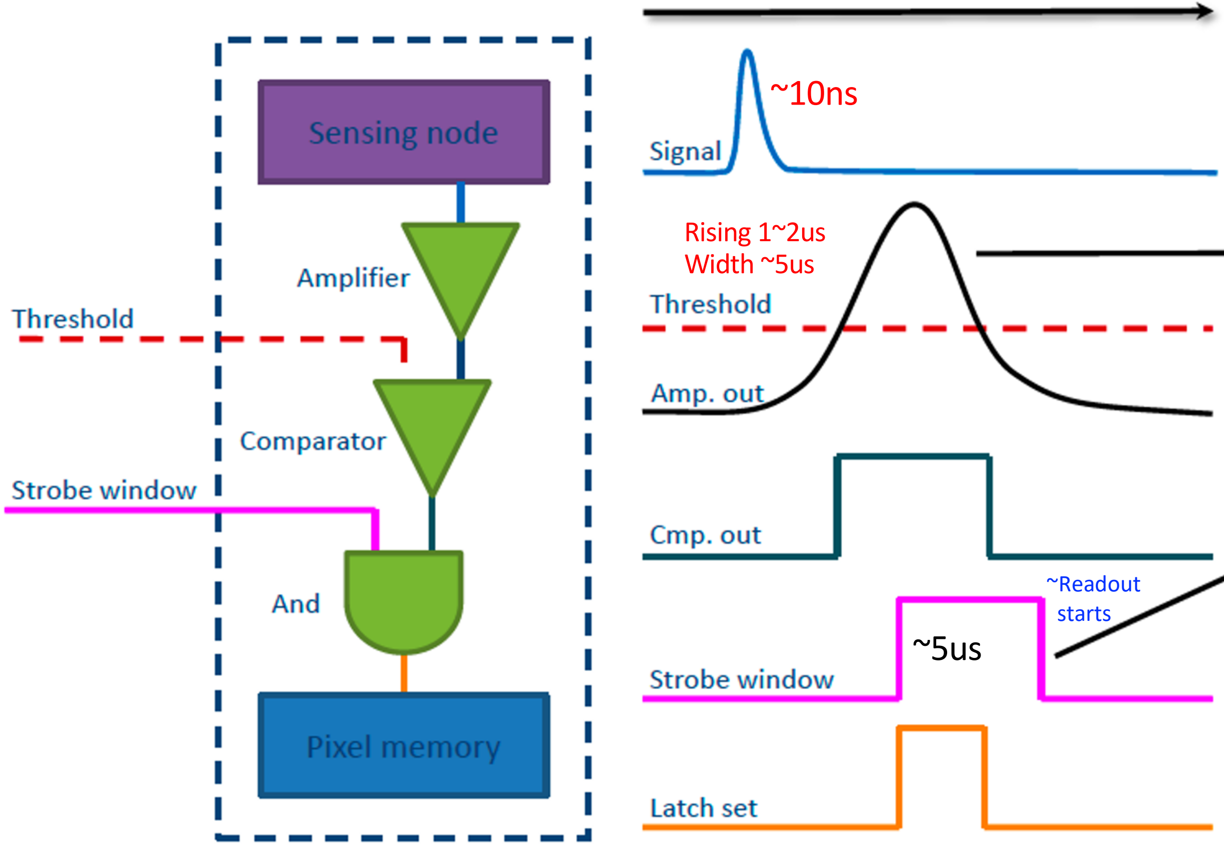


[https://www.bnl.gov/eic/]

Discussions at MIT, 12/14/2021



ALPIDE Timing



Analog "slow" shaping time acts as a memory (µs order) to buffer the signal until the trigger arrives.
Called "pulse length"

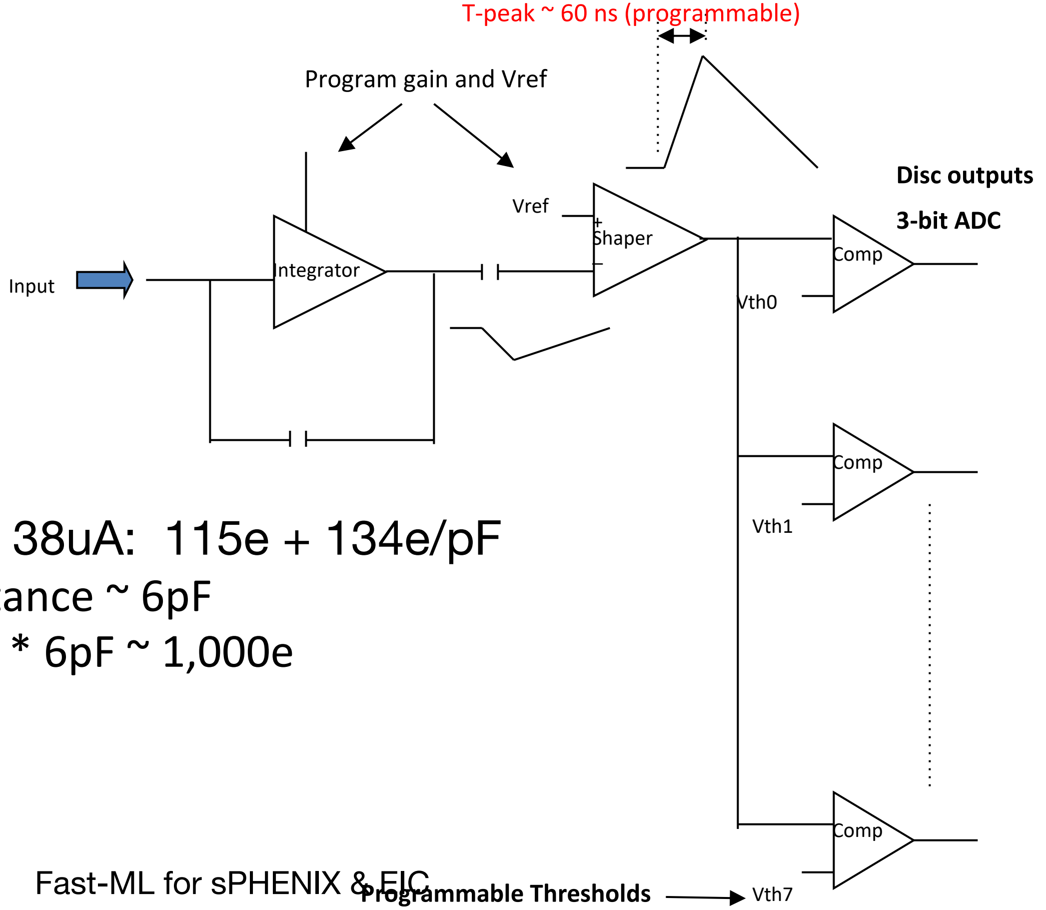
Only pixel with signal higher than threshold DURING the strobe window will get latched in memory

Confirmed at LANL: pulse length can be extended to meet sPHENIX trigger latency requirement

FPHX Chip (Analog Section)

Tracking System: MVTX+INTT+TPC

- 128 channel
- 46, 50, 60, 67, 85, 100, 150, or 200 mV/fC
- 60 ns peak time (rt- pgmbl)
- 3—bit ADC (th-pgmbl)
- Optimized to 1 to 2.5 pf input
- 115e + 134e/pf
- ~ 70 to 140 μW/ch (dep. gain)



At max. input transistor bias of 38uA: 115e + 134e/pF

Silicon strip + readout line capacitance ~ 6pF

Predicted noise ~ 115e + 134e/pF * 6pF ~ 1,000e

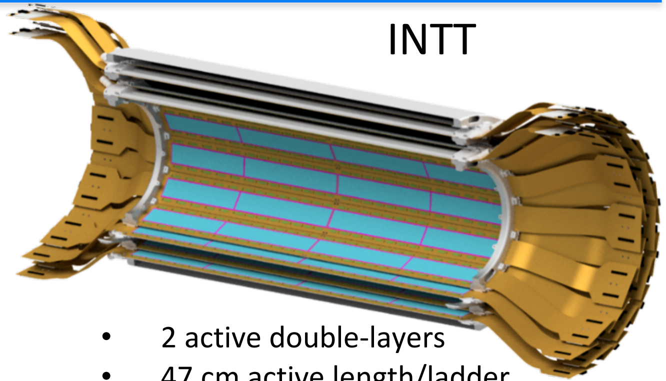
Signal ~ 20,000e

Expected S/N ~ 20/1

Tracking at sPHENIX

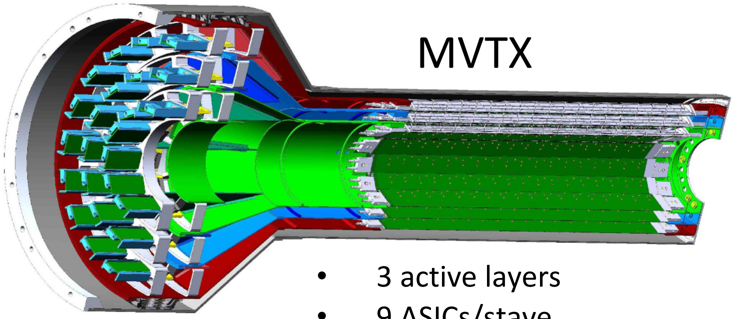
- Tracking consists of 3 sub-detectors:
  - Pixel Vertex Detector (MVTX)
  - Intermediate Silicon Tracker (INTT)
  - Time Projection Chamber (TPC)
- MVTX and INTT are both capable of streaming readout
- Combined tracking to r = 10.3 cm

INTT



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

MVTX



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

sPHENIX HF constraints

- sPHENIX has great tracking and calorimetry
- However, limited by calorimetry backend readout rate (15kHz) in triggered mode
- RHIC pp rate is ~10 MHz
- Plan: Use tracker SRO to recover some heavy flavor physics potential

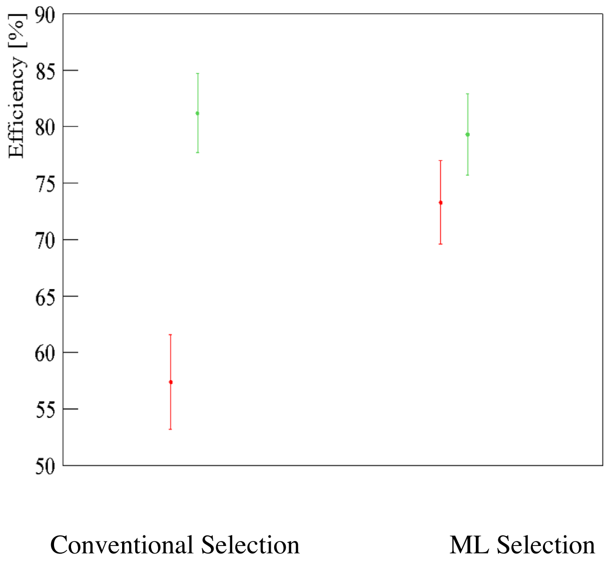
Year	Species	$\sqrt{s_{NN}}$ [GeV]	Cryo Weeks	Physics Weeks	Rec. Lum. $\|z\|$ <10 cm	Samp. Lum. $\|z\|$ <10 cm
2023	Au+Au	200	24 (28)	9 (13)	3.7 (5.7) nb^{-1}	4.5 (6.9) nb^{-1}
2024	$p^{\uparrow}p^{\uparrow}$	200	24 (28)	12 (16)	0.3 (0.4) pb^{-1} [5 kHz] 4.5 (6.2) pb^{-1} [10%-*str*]	45 (62) pb^{-1}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

Case study: AI HF selections

sPHENIX

- Several algorithms trained using TMVA
  - Fast turnaround due to proposal time constraints
  - Algorithms used "out-of-the-box", no optimizations
- Trained using samples with no HF signal and with $D^0 \rightarrow K^- \pi^+$ signal
- Selection tuned for approx. equal signal efficiency



Green – The signal selection efficiency
Red – The background rejection efficiency

How are FPGAs programmed?

Hardware Description Languages

HDLs are programming languages which describe electronic circuits

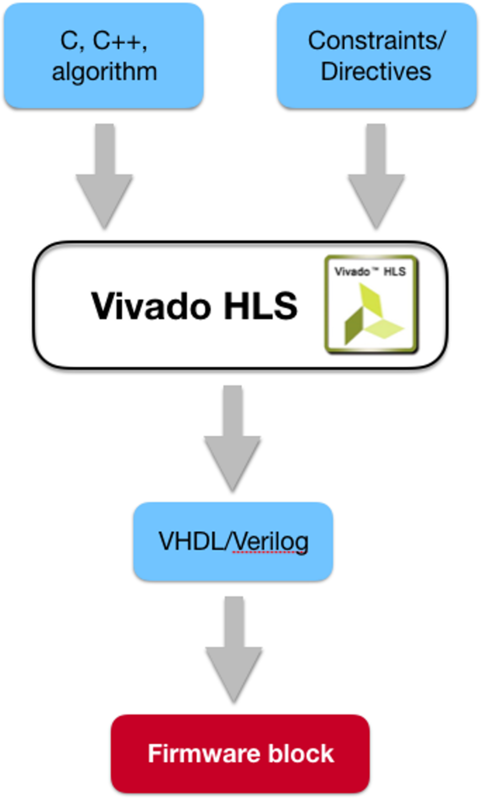
High Level Synthesis

Compile from C/C++ to VHDL

Pre-processor directives and constraints used to optimize the design

Drastic decrease in firmware development time!

Today we'll use Xilinx Vivado HLS [*]



[*] https://www.xilinx.com/support/documentation/sw_manuals/xilinx2020_1/ug902-vivado-high-level-synthesis.pdf

Constructing ML algorithms

- Aim to develop algorithms as Graph Neural Networks (GNN)
- Advantageous over Convolutional Neural Networks (CNN) by adding edge information
- Detector and physics knowledge will improve predictions
- Algorithms deployed at several points:
  1. **Fast tracking on FPGA**
  2. **Topological separation of HF signals on FPGA**
  3. **Beam-spot and anomaly detection on GPU**
  - Part of feedback system to improve 1 & 2 plus inform detector operators

Constructing ML algorithms

- Anomaly detection is important in all experiments
- RHIC experiments cannot be accessed during beam to fix issues
- Aim to use variational autoencoders
- Incoming data can be compared for:
    1. Noisy pixels
    2. Dead strips
    3. Change in beam spot or alignment
- Pass info. back to selection system to improve yield
- Pass back to control room

What we actually used in Json Data Files?

‘RawHit’ contains two part: [u'MVTXHits', u'Description']

- ‘MVTXHits’ contains all the hits information.

  Each hit contains: [u'Coordinate', u'ID']

- ‘ID’ contains: [u'Layer', u'PixelZIndex', u'Chip', u'Stave',

  u'PixelPhiIndexInLayer', u'HitSequenceInEvent',

  u'PixelPhiIndexInHalfLayer',u'PixelHalfLayerIndex', u'Pixel',

  u'HalfLayer']

Training Dashboard