High Performance FPGA-Based Embedded System for Decision Making in Scientific Environment

Supported by DOE grant DE-SC0019518
Phase I: 2019 - 2020, Phase II: 2020 - 2023

Yu Sun, PI
CEO of Sunrise Technology Inc.
yu.sun@sunriseaitech.com

DOE SBIR NP Exchange Meeting, August/23/2022

Outline

- Company Introduction
- Description of the Phase II project
  - Objectives
  - Nuclear Physics Background
  - Schedule and Deliverables
- Project Highlights
  - Trigger Algorithm Description
  - Performance Report
  - Highlights of the final product
- Conclusions, Future Plans and Milestones

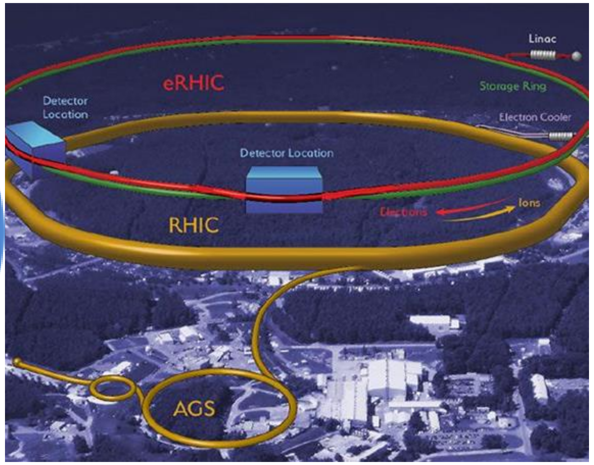
About Sunrise Technology Inc.

- Founded in 2017
- Located in an incubator at Stony Brook University, Brookhaven, NY
- Team: three full-time employees, a part-time consulting scientist, and several graduate interns.
- Developing advanced machine learning/deep learning and AI technology for autonomous systems, such as scientific experiments decision-making engines and education platforms.
- Product and Service areas

1) FPGA-enabled GNN Solutions
2) Embedded system for modeling training
3) Deep Reinforcement Learning for large facility control

Yu Sun	CEO and Founder
Giorgian Borca-Tasciuc	employee, Machine Learning Research, FPGA implementation
Kevin Mahon	employee, GNN Development, embedded system
Mingxiong Liu	Collaborator, Nuclear Physics Science Lead, LANL
Cameron Dean	Collaborator, Physics Detector simulation
Zhaozhong Shi	Collaborator, simulation data
Tingting Xuan	Ph.D. candidate, Deep Learning Research and Algorithm development
Yimin Zhu	Ph.D. candidate, Deep Graph Neural Network Research and Parallel Algorithm

Core Competencies
&
Potential Markets



- Machine Learning and Deep Learning Algorithm Design
- AI/ML for Science Facilities
- Data Science for Physics Analysis
- Deep Reinforcement Learning for Orbit Control
- AI-Enabled Heterogeneous Embedded systems (with CPU, GPU, and FPGA) for Science Facilities Automation (particularly accelerators detectors)
- Edge Systems Software Stack

SBIR Phase II Objectives

- Ultimate Goal
  - Design an AI-enabled DAQ trigger system
  - Integrated into sPhenix experiment and reaches the target of 15Khz data acquisition rate.
- Phase II Technical Objectives
  - Designing Graph Neural Networks for High-Speed Physics Event Triggers.
  - Implementing High-Speed Triggers for Nuclear Physics Experiments.
  - Creating a flexible embedded hybrid system to support training and inference.
- Phase II Commercialization Objective
  - Manufacture smart embedded hybrid system to facilitate real-time data collection in large scale experiments and facility control

SBIR Phase II Project Periods

Project Period April/2020-April/2021: Basic Algorithmic Development

- Identify and Train Machine Learning Models with Inner-most MVTX detector simulation data
- Verify the baseline Performance Accuracy and Throughput
- Preliminary Hardware Development on FPGA

Project Period May/2022-April/2023:

Advanced Algorithmic and Field Validate in sPhenix

- Advanced ML-Algorithm Development with MVTX + INTT detector simulation data
- Improve model performance
- Reduce model inference latency
- Objective: 200K-500K events/second/FPGA card

May 2021 – Apr. 2022

May 2022 – Apr. 2023

Apr. 2020 – Apr. 2021

Project Period May/2021-April/2022:

Advanced Algorithmic and FPGA Development

- Advanced ML-Algorithm Development with MVTX + INTT detector simulation data
- Improve model performance
- Reduce model inference latency
- FPGA HLS development and deployment
- Objective: 100K-200K events/second

Relevance to DOE NP SBIR Program

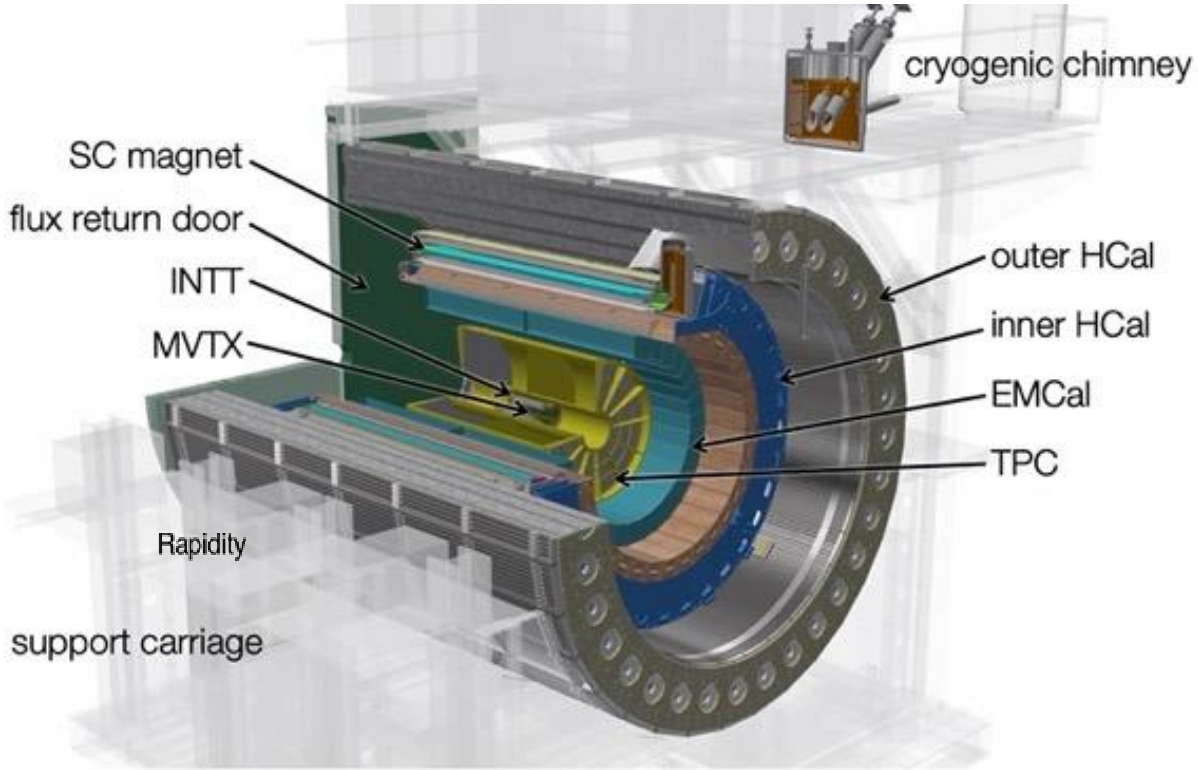
- Project Focus:
  - Real-time AI technologies will be applied to the very high-rate data streams from detectors.
  - Accelerate GNN on FPGA, one of the first work that attempts to accelerate GNN prediction.
  - Play the central role in sPhenix and Future EIC detectors running under trigger systems and in-situ streaming analysis for event selections.
- Project Impacts:
  - **ASCR C55-01 (ACCELERATING THE DEPLOYMENT OF ADVANCED SOFTWARE TECHNOLOGIES), Subtopic a):** Deployment of ASCR-Funded Software
  - **NP C55-21: Nuclear Physics Software and Data Management and subtopic**
    - **b. Applications of AI/ML to Nuclear Physics**
    - **c. Heterogeneous Concurrent Computing**.
- Subcontractor/Collaborators
  - Dr. Ming Xiong Liu, Dr. Cameron Dean, LANL
  - Dr. Jin Huang, Dr. Zhaozhong Shi, BNL

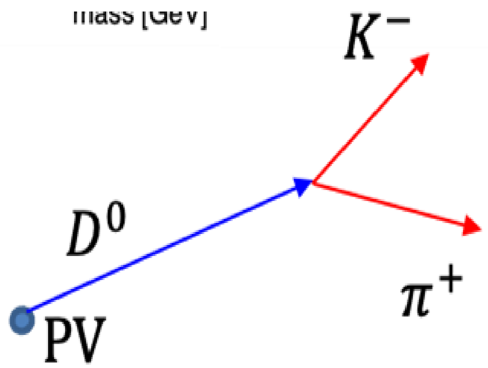
The Readout Challenge for High Luminosity Physics

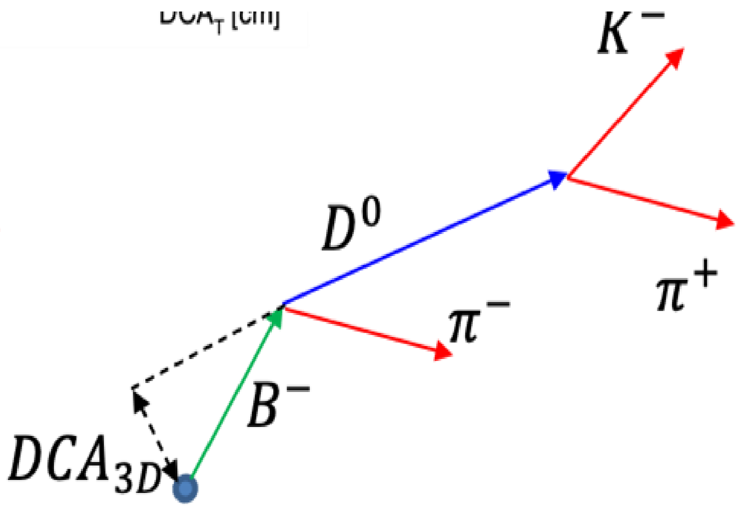
- The readout challenge
  - Raw data volume >> hardware bandwidth/storage
  - Only a small fraction of data will be recorded to tape

- **sPHENIX: DAQ trigger rate, 15kHz**
  - AuAu collisions
    - Max collision rate ~50kHz
    - Can collect all central collisions, OK
  - p+p and p+Au
    - Collisions on each beam crossing, ~9.4MHz
    - Okey for high energy jet program with triggers
    - Lose most of the low pT physics events
- AI-based Triggering: filter events to reduce data rates for data archive and offline processing
- sPhenix Trigger TPC (Time Projection Chamber) Data Acquisition
- SBIR project focuses on designing, building, simulating, and benchmarking a prototype event readout system with AI-based fast online data processing and autonomous detector control system that meets the physics and engineering requirements.

sPHENIX experiment under construction at RHIC:
- Day-1 physics in 2023



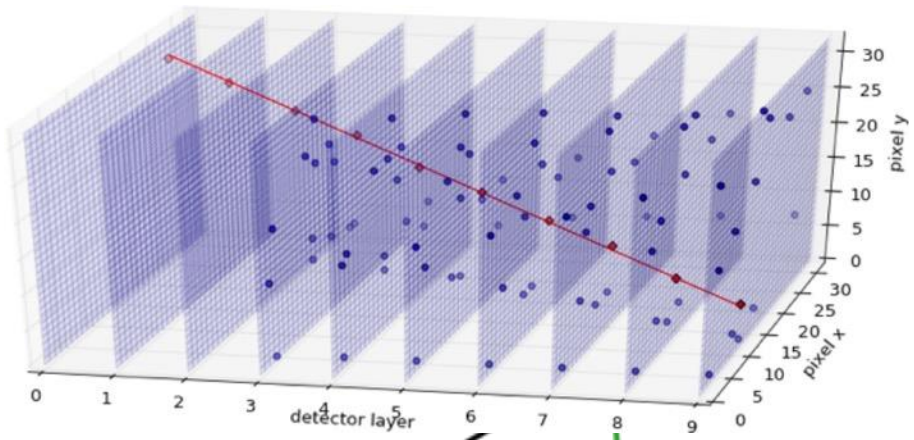


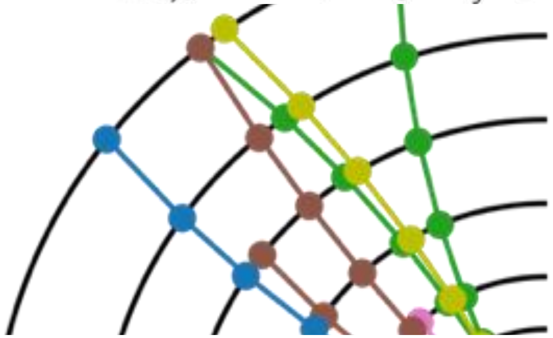


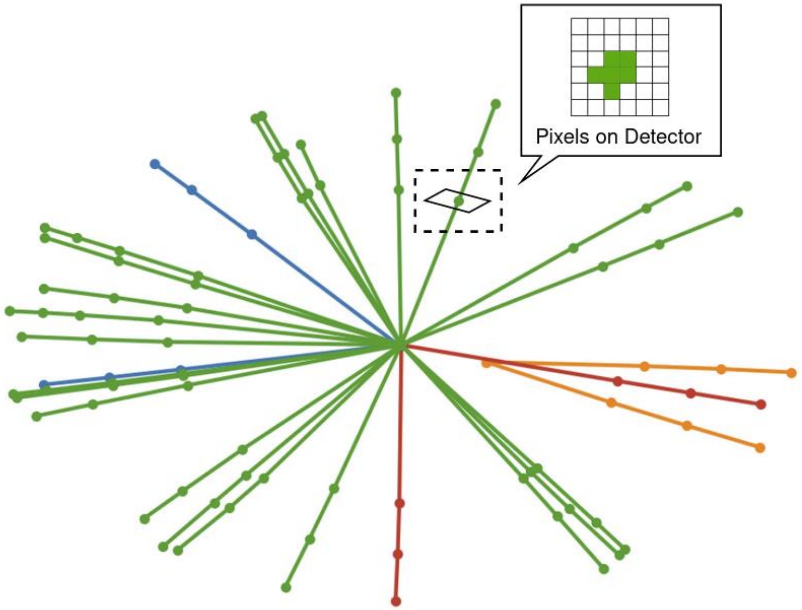
Event Data Descriptions

Moving from images to points

- **Image-based methods face challenges scaling up to realistic HL-LHC conditions.**
  - High dimensionality ($9K \times 9K \times 3$) 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)
  - But now we can exploit the structure of the data with full precision
- What ML models are appropriate for the event on right
  - – Graph neural networks







Trigger Software Pipeline

1. Fetch events from Detector Readout (Use Simulation Data)

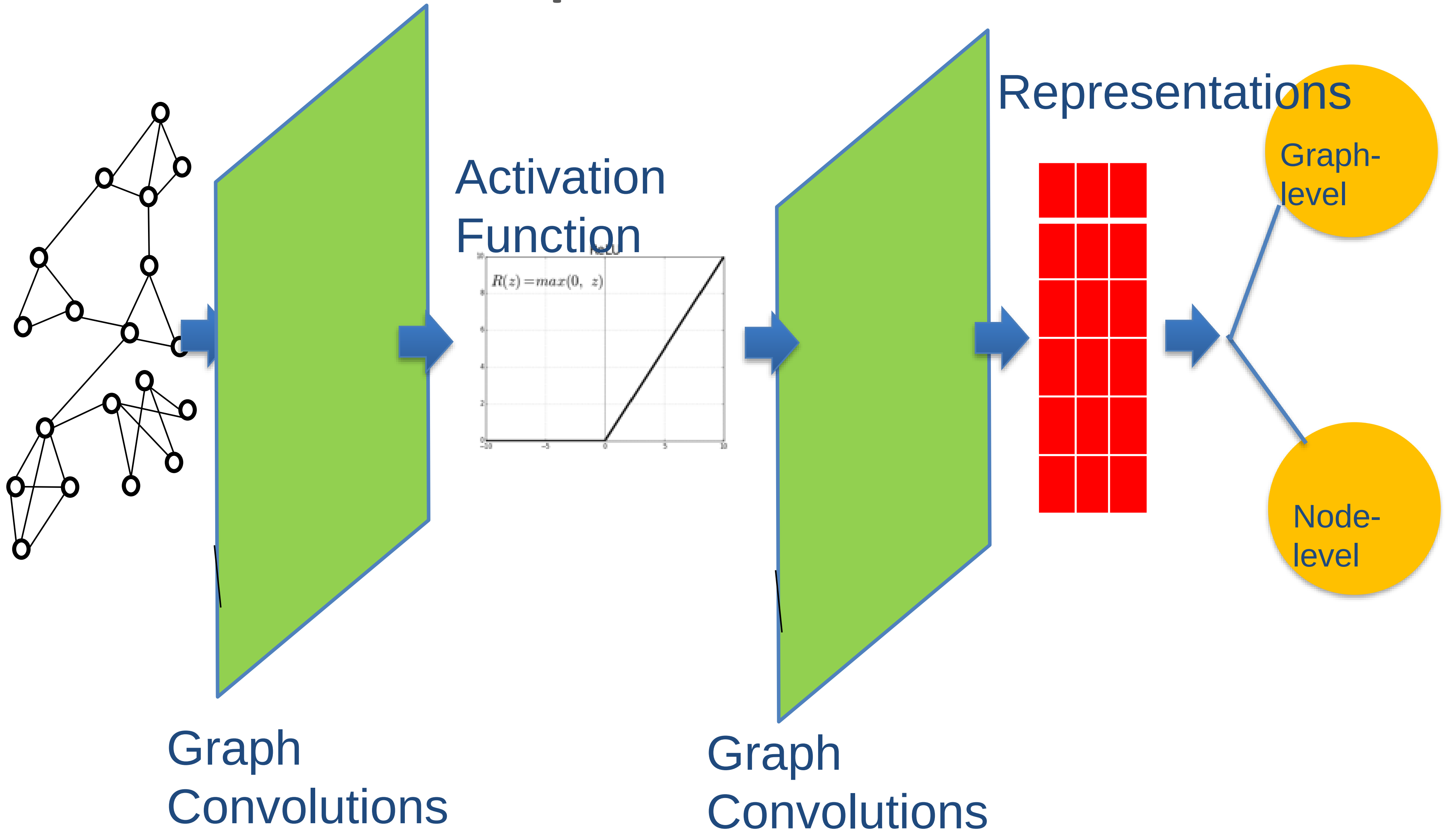
2. Data Pre-processing Clustering

3. Tracking + Outlier hits Removal

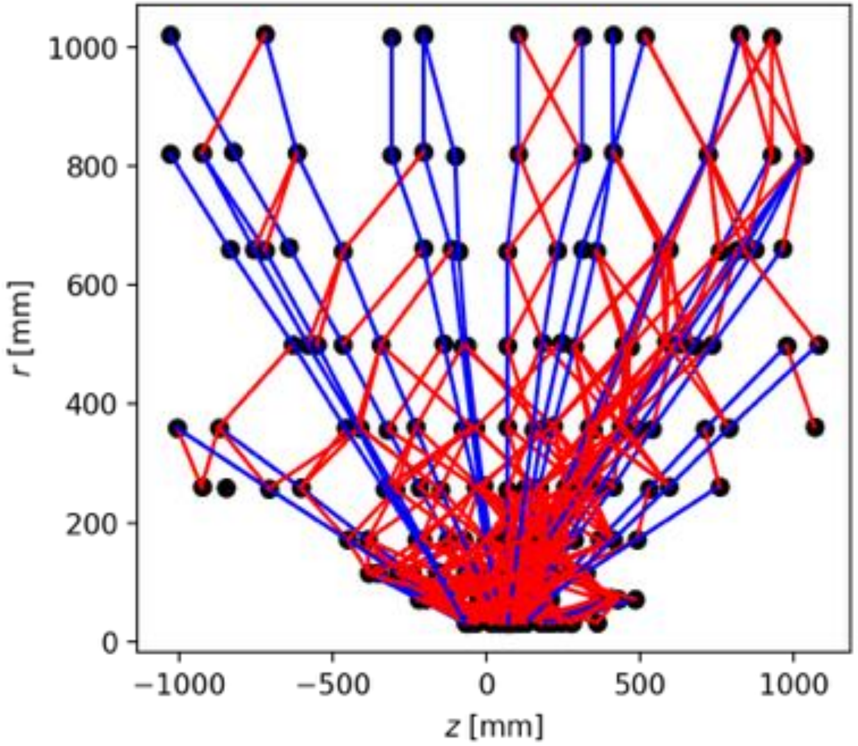
4. Triggering Decision

5. Triggers on TPC (Interface and integration with sPhenix Detector)

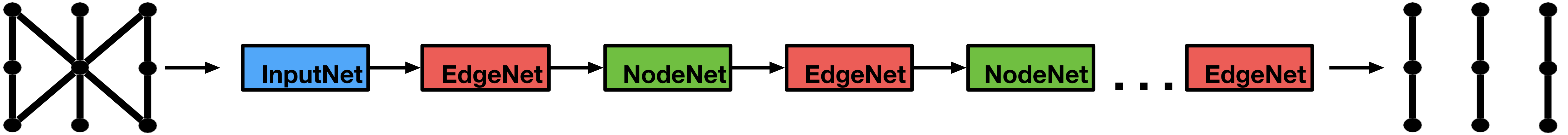
Graph Neural Networks



Graph Tracking and Outlier Removal



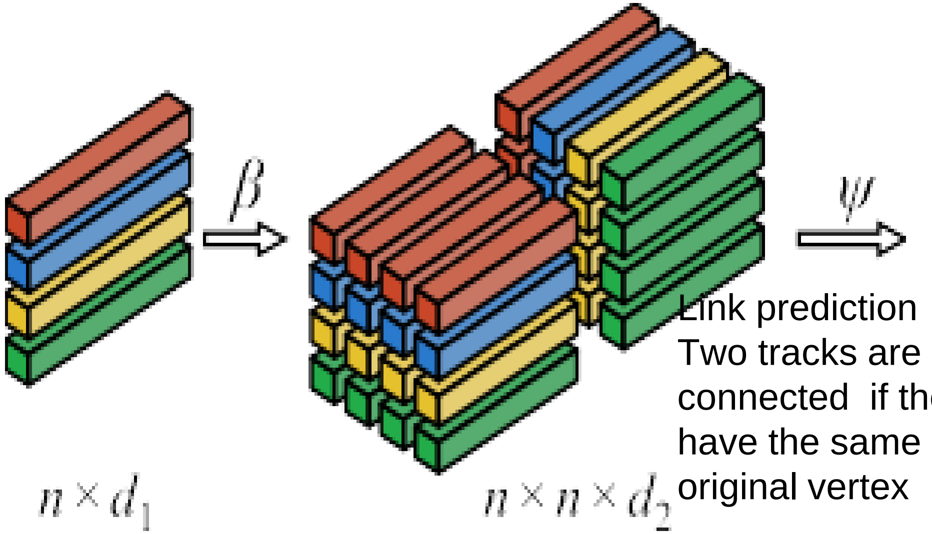
- **What if we structure our data as a *graph* of connected hits?**
  - Connect plausibly-related hits using geometric constraints
- **What kinds of models can we apply to this representation?**
  - Traditional architectures clearly don't work
  - but there's a growing sub-field of ML called *Geometric Deep Learning*
- Connect hits on adjacent layers using crude geometric constraints, i.e., $\delta(\phi) \leq \frac{\pi}{4}$ and $\delta(z) \leq 300mm$



With each iteration, the model propagates information through the graph, strengthens important connections, and weakens useless ones.

Trigger Detection Algorithm

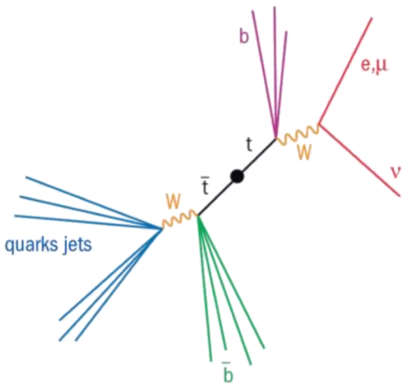
- A GNN-based trigger system to decide whether to record the events or not, with the processed track information retrieved from the captured 3D sparse images by the sPHENIX detectors.
  - Hits based algorithm: each graph node is a hit on detector and events are represented by a collection of hit clouds. Graph Neural Network is simple to implement and has fast computation time.
  - Track based algorithm: each graph node is a track that represents a particle generated from collision. Event consists of the tracks for particles. Graph Neural Network is hard to implement but we can learn high-level physics knowledge to improve the trigger accuracy.



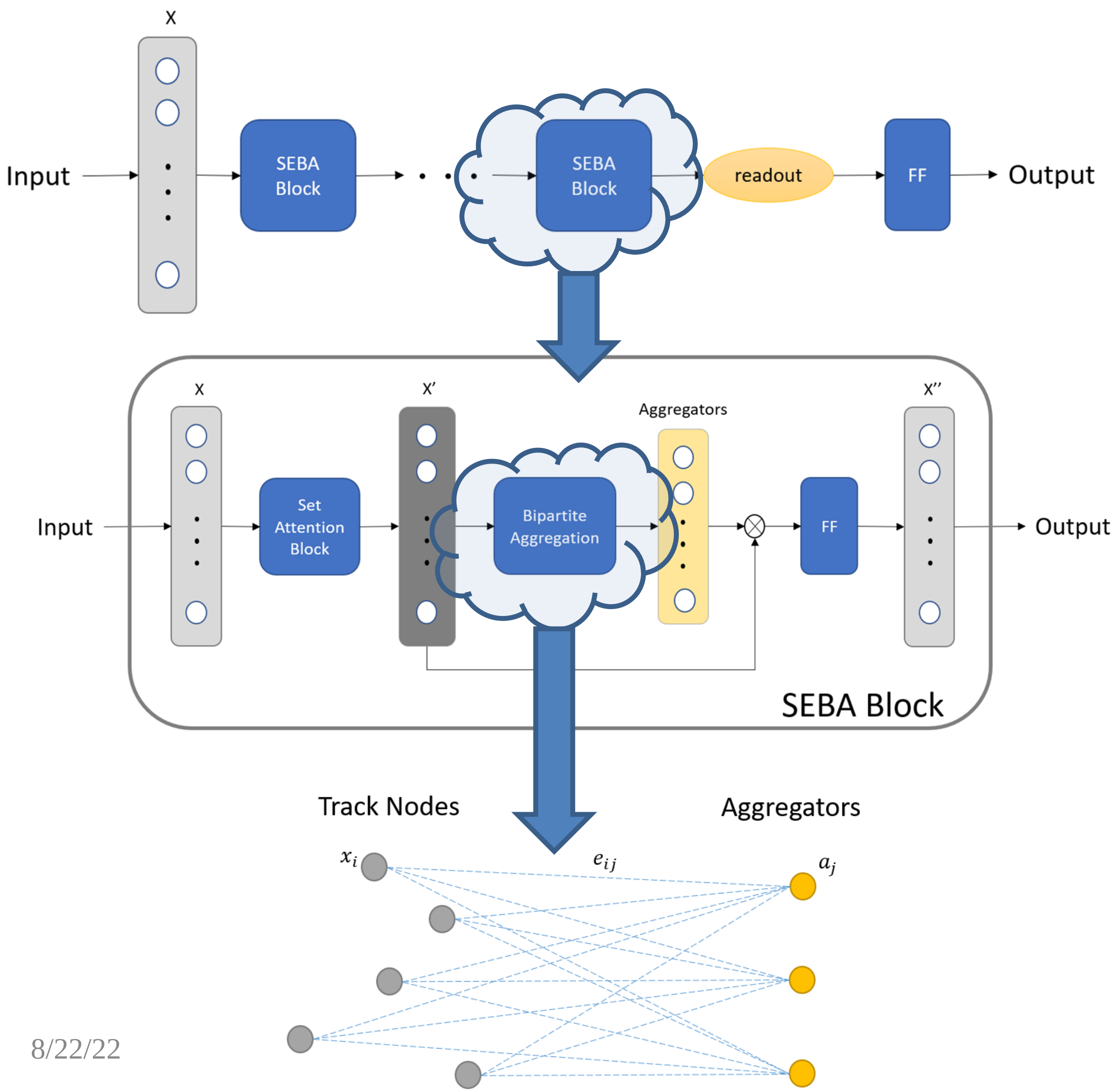
Link prediction
Two tracks are connected if they have the same original vertex

$$\hat{A}$$

$$\begin{pmatrix} 1 & 0.74 & \cdots & 0.28 \\ 0.74 & 1 & & 0.8 \\ \vdots & & \ddots & \vdots \\ 0.8 & 0.53 & & 0.17 \\ 0.28 & 0.47 & \cdots & 1 \end{pmatrix}$$



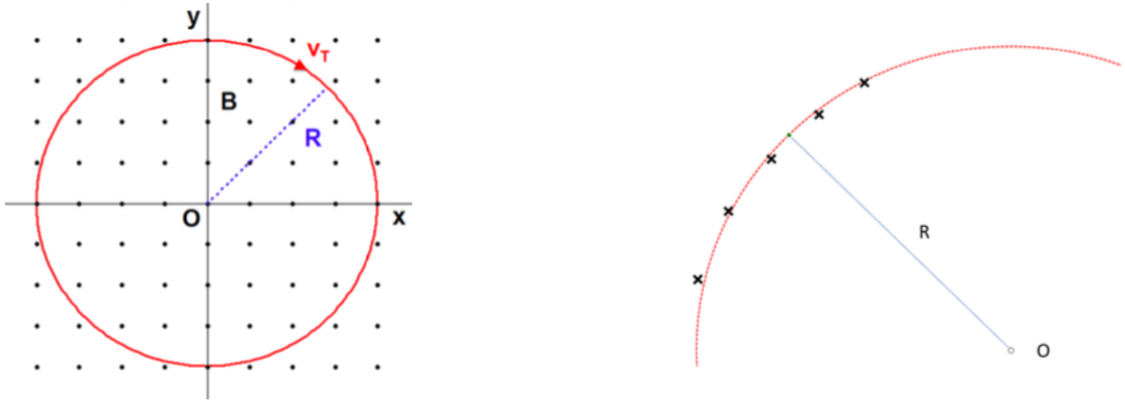
Trigger Detection Network Architecture



- SEBA - set encoder with Bipartite aggregator
- Readout Functions:
  - Mean Pooling
  - Max Pooling
  - Pool on the concatenation of node embedding from this GNN layer

Physics-Aware Graph Neural Networks for trigger

- Each track represents the trace of a particle. Can we estimate some high-level physics information, e.g., Particle Mass, Momentums, and particle ID?
- We demonstrate the physics-momentum guided GNN improves accuracy by 15~16% over those without it.



$$P_T = 0.3BR, where\ B\ is\ the\ magnetic\ field\ strength$$

Fig. 3: The left figure shows that a positively charged particle will undergo a circular motion clockwise with a radius R in the uniform magnetic field B along the $+z$ direction. The right figure shows an example track with a fitted circle. The black cross markers represent five hits on the example track; the red dashed curve approximate a particle track and is the fitted circle with a radius R.

Experiment Results

Table 5: Ablation Study of Activations

Activation	Accuracy	AUC
ReLU	90.74%	96.87%
Tanh	90.19%	96.58%
Potential	90.41%	96.75%
Softmax	**92.18%**	**97.68%**

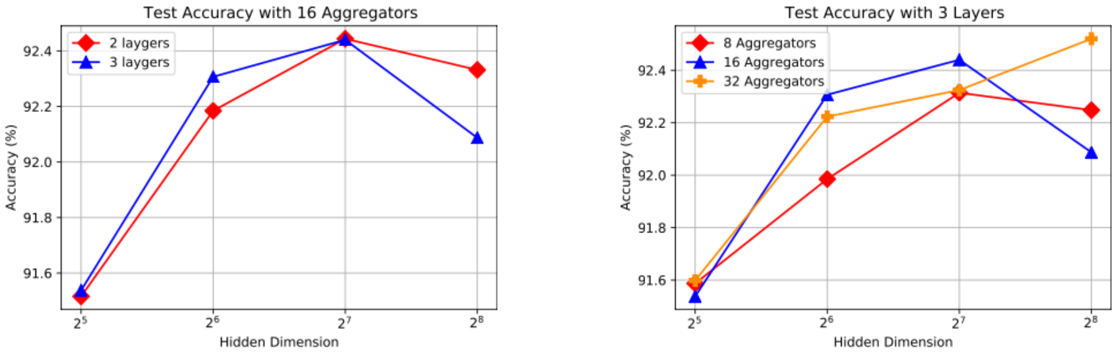


Fig. 7: Accuracy performance in respect to hidden dimension for two/three-layer models and different number of aggregators.

Table 2: Comparison to Baseline Models with Estimated Radius.

	with LS-radius			without radius		
Model	#Parameters	Accuracy	AUC	#Parameters	Accuracy	AUC
Set Transformer	300,802	84.17%	90.61%	300,418	69.80%	76.25%
GarNet	284,210	90.14%	96.56%	284,066	75.06%	82.03%
PN+SAGPool	780,934	86.25%	92.91%	780,678	69.22%	77.18%
BGN-ST	355,042	**92.18%**	**97.68%**	354,786	**76.45%**	**83.61%**

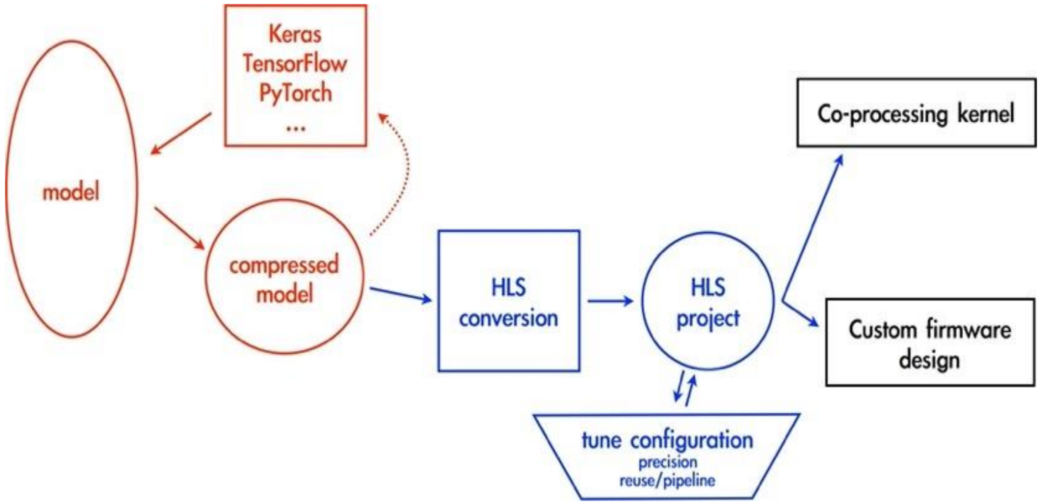
Year 2022

Year 2021

FPGA Implementation

`hls4ml` is a software package for creating HLS implementations of neural networks.

https://hls-fpga-machine-learning.github.io/hls4ml/



FPGA Performance

Pipeline Stage	Number of Parameters	Accuracy	Kernel Time (μs)	Speedup
Clustering	-	99.2%	85	1152×
Tracking	745	92.8%	23	280×
Triggering	2441	68.1%	35	21×
Full Pipeline	3186	68.0%	140	750×

Deep Learning Training and Inference Product Hardware

Stage 1: Off-line Training

Stage 2: On-line Training /Transfer Learning

Stage 3: Real-Time Inference

Stage 4: Light-Weight Real-Time Deep Learning & Inference Engine

Stage 5: Real-Time Sensing & Decision Making Engine

NVIDIA GPU Clou

1

+

2

=

3

4

High Performance Computing | Edge Computing | Embedded System | Controller

Electricity Cost High | Electricity Cost Medium | Electricity Cost Low

GPU servers

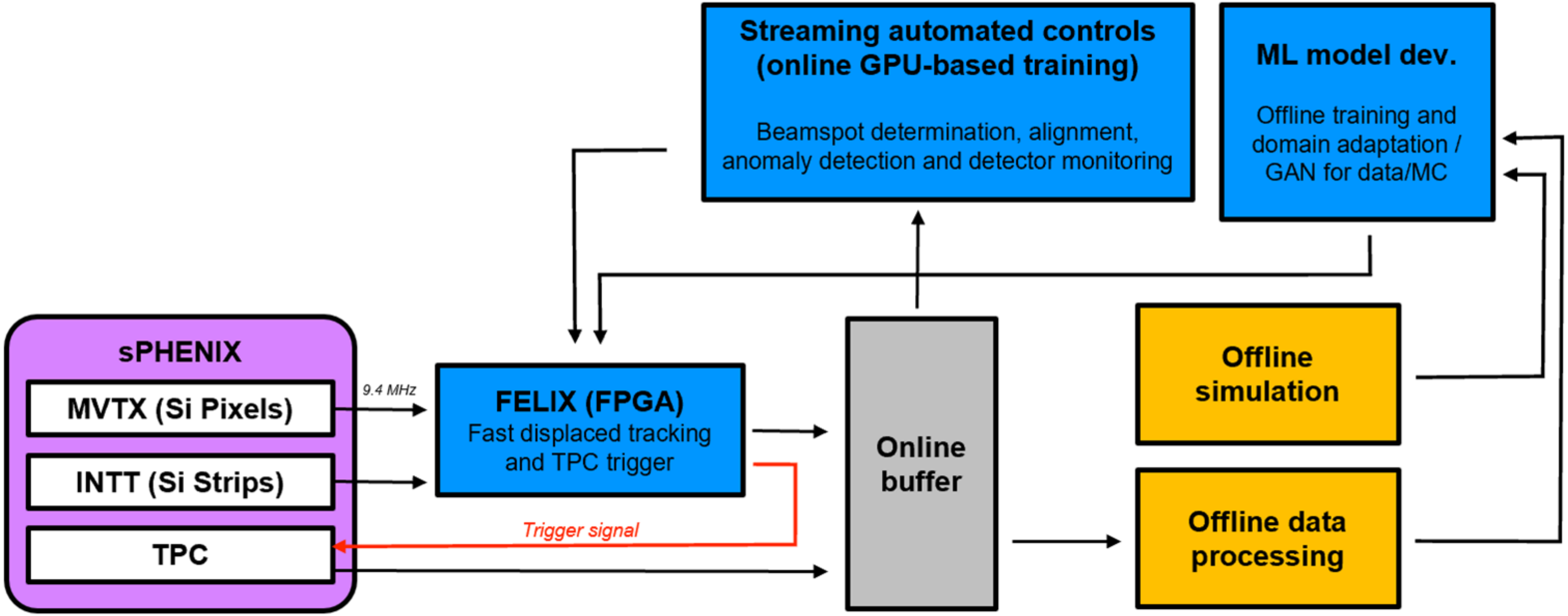
High-end Embedded System

AI-Engine Board and Standalone Embedded System

Future Plan: Integrated into the sPHENIX Readout Upgrade (DOE Project led by LANL)

AI-based real-time system: Fast Data Processing and Smart Trigger

- Identify heavy quark events in p+p and p+Au collision events



Conclusion, Accomplishments and Milestone

1. Implement the Trigger Detection Algorithm based on advanced GNN
2. Implement Physics-aware pipeline for decision making
3. Extremely fast GNN algorithm on FPGA (3KHz/second for end-to-end pipeline), 20 times faster than GPU (2021).
4. With the Support of HLS4ML, the trigger software runs on a server and embedded system with FPGA (2022)

Year 2 milestones

- Simulation Dataset with MVTX+INTT (1~5 million events) and retrained models (Done)
- FPGA implementation for new models with MVTX and INTT (in Progress)
- Fast prototype design for online triggering hardware (Done)
- Design and implement embedded system with both training (on GPU) and inference (on FPGA)

Year 3 milestones:

- sPhenix trigger to be deployed for upcoming sPhenix experiment run at 2023.

Acknowledgement

- Thank DOE Office of Science, Dr. Michelle Shinn for funding this project, and every contributor for working on this project!