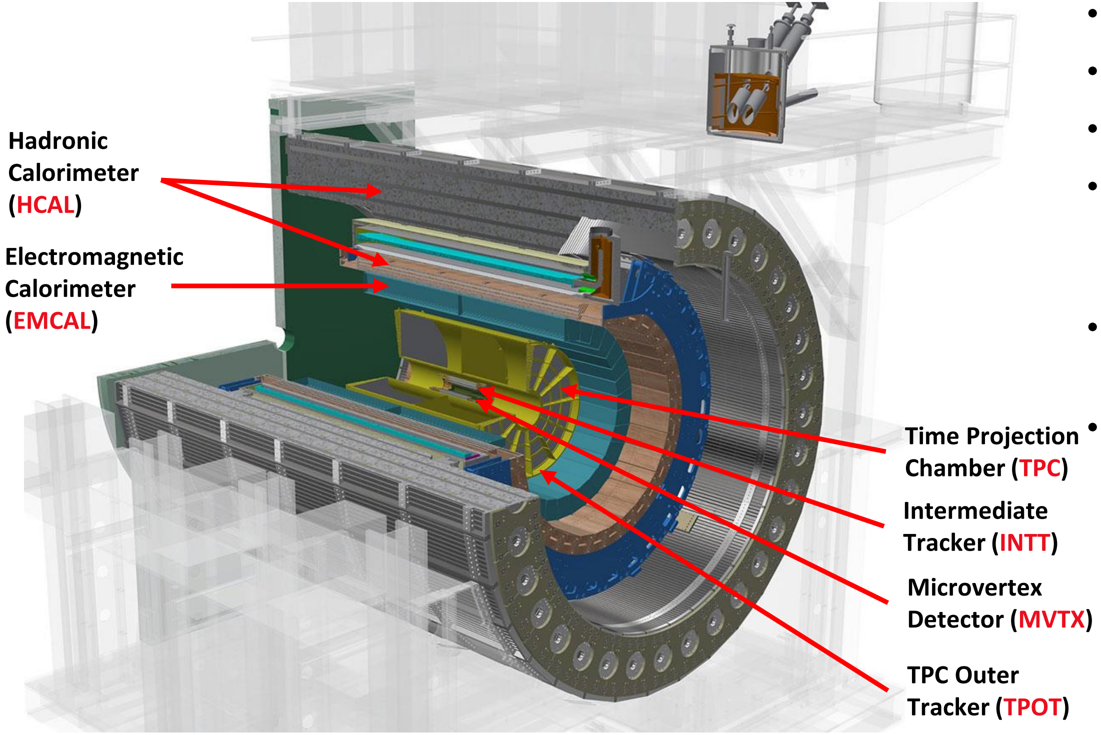
Intelligent experiments through real-time AI: Fast Data Processing and Autonomous Detector Control for sPHENIX and future EIC detectors

Jakub Kvapil for the FastML team and sPHENIX collaboration

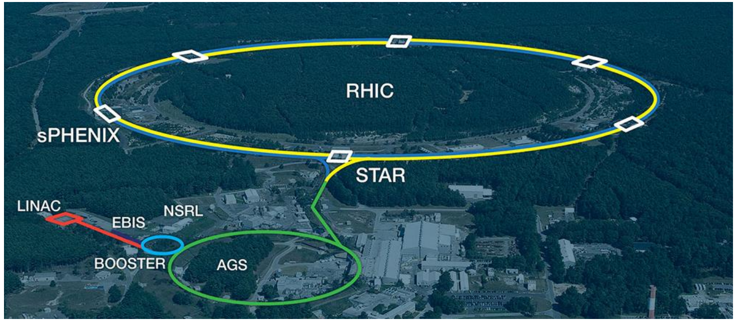
42nd International Conference on High Energy Physics
Prague, Czech Republic

Los Alamos National Laboratory (LANL)
Fermi National Laboratory (FNAL)
Massachusetts Institute of Technology (MIT)
New Jersey Institute of Technology (NJIT)
Oak Ridge National Laboratory (ORNL)
Georgia Institute of Technology (GIT)

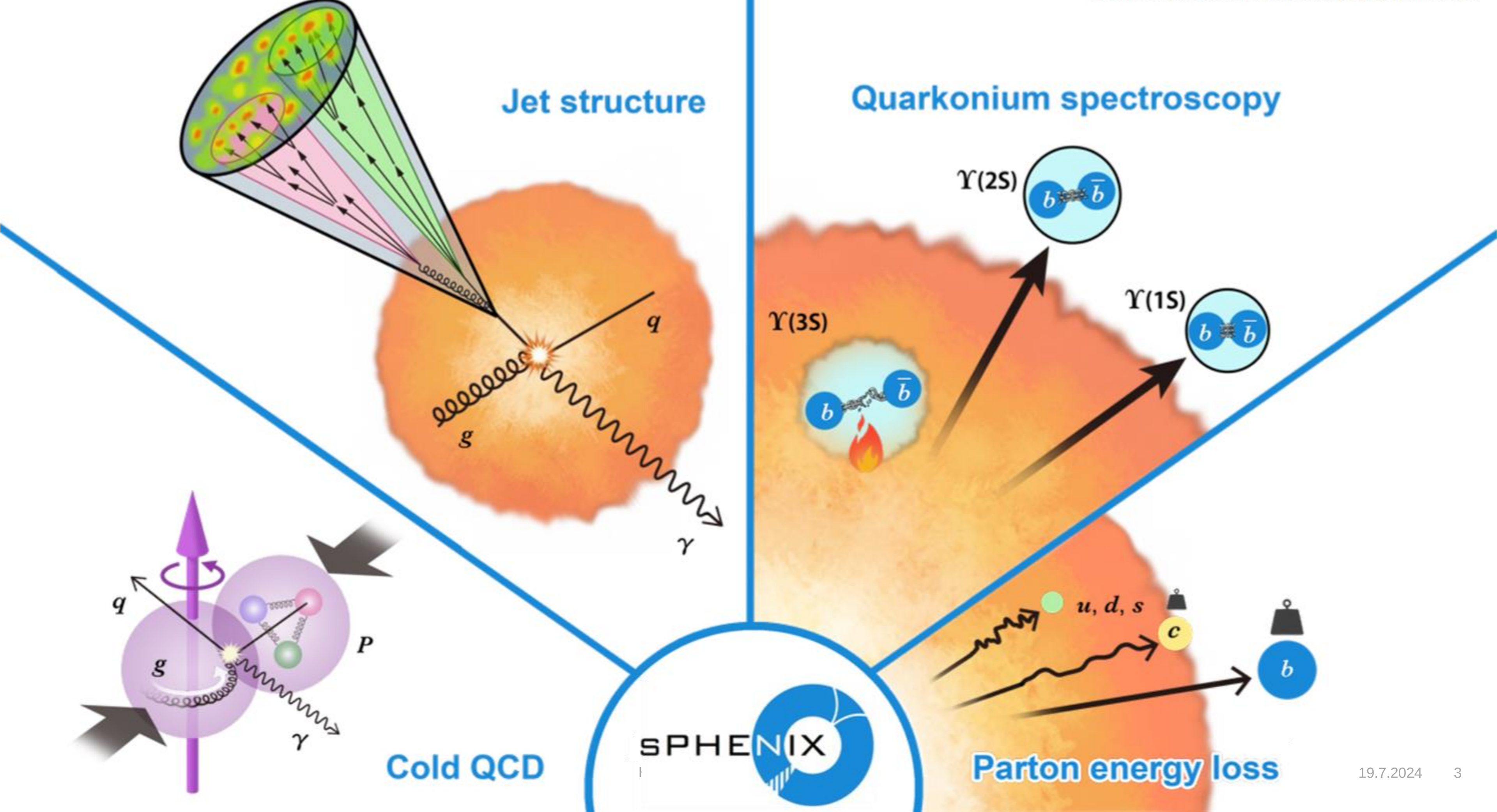
sPHENIX experiment



- Located at RHIC accelerator at BNL (USA)
- Running period 2023-2025
- 1.4 T Magnetic Field, $|\eta| \leqslant 1.1$
- Tracking detectors (MVTX, INTT, TPC, TPOT) and calorimeters (EMCAL, HCAL), and endcaps (MBD, sEPD, ZDC, SMD)
- Tracking detectors capable of streaming readout
  - Hybrid DAQ supporting 15 kHz triggering
- sPHENIX has the first mid-rapidity HCAL at RHIC!



Illustrated by Misaki Ouchida (Hokkaido University)



Motivation – Heavy Flavor

- RHIC collision rate is around 2-3 MHz, sPHENIX readout 15 kHz (DAQ - 300 Gb/s)
  - Trackers are Streaming Readout (SRO) capable, but can’t save all TPC data
- 10% trigger-enhanced SRO increases HF MB rate ~ 300 kHz
- **ML HW tagging aims to sample remaining 90% of the luminosity using the tracklet reconstruction from the silicon detectors**

1. Integrate the AI-based heavy flavour trigger system demonstrator into the sPHENIX experiment for p+p run in 2024 to R&D its feasibility, requirements, and constrains

2. Deploy future system on Electron-Ion Collider (EIC)
   - AI-based electron tagging with streaming readout to identify the (non)interesting Deep-Inelastic-Scattering (DIS) processes in the e+p/A collisions.
     - based on the measured scattering electron energy and direction

FastML - Who are we?

- Cross-discipline group of sPHENIX and LHC physicist, engineers, and computer scientists working on firmware-based ML applications
  - sPHENIX is benefiting from a 2020 Department of Energy (DOE), USA funding call
- The goals
  - Use MVTX and INTT hits to identify HF event based on topology
    - Continuously monitor the beam spot and changing detector conditions
  - Send tag downstream to readout TPC
  - Trigger signal must be done within 10 us!
    - Requires embedding ML algorithms on FPGA (we will use FELIX-712 board)

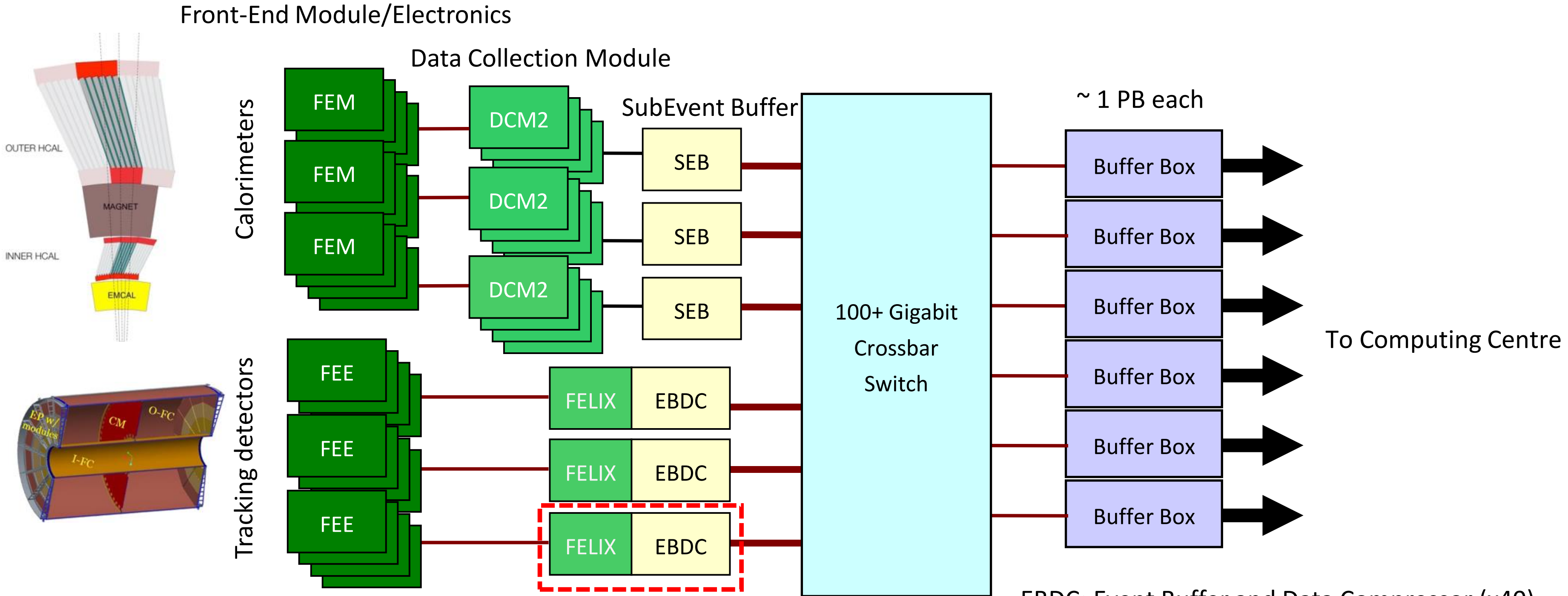
Intelligent experiments through real-time AI: Fast Data Processing and Autonomous Detector Control for sPHENIX and future EIC detectors

A proposal submitted to the DOE Office of Science

April 30, 2021 @ renewed for 2 more years in 2023

sPHENIX Readout and AI-ML HF Trigger Integration

On Detector   Rack Room



FELIX (MVTX+INTT) -> AI/ML -> Trigger

EBDC: Event Buffer and Data Compressor (x40)
- 6x MVTX, 8x INTT, 24x TPC

The ML algorithm pipeline

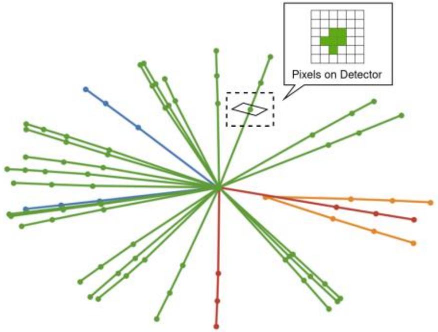
- Due to latency limitation all components need to be deployed on FPGA

1. Hit decoding – conventional logic
2. Hit clustering – conventional logic

The ML algorithm pipeline

- Due to latency limitation all components need to be deployed on FPGA

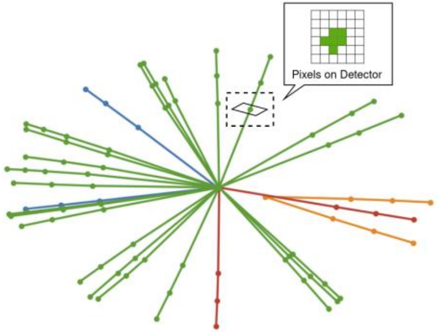
1. Hit decoding – conventional logic
2. Hit clustering – conventional logic
3. Track construction
    1. Edge candidate generation
        - Connect all clusters (nodes) together (edges) while applying geometric constrain
            - Reduced number of edges ~50% at the cost of 0.4% accuracy



The ML algorithm pipeline

- Due to latency limitation all components need to be deployed on FPGA

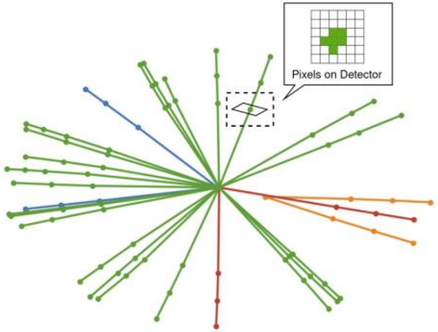
1. Hit decoding – conventional logic
2. Hit clustering – conventional logic
3. Track construction
    1. Edge candidate generation
        - Connect all clusters (nodes) together (edges) while applying geometric constrain
            - Reduced number of edges ~50% at the cost of 0.4% accuracy
    2. Edge candidate classification – GCN (arXiv:1609.02907)
        - model predicts true edge candidates



The ML algorithm pipeline

- Due to latency limitation all components need to be deployed on FPGA

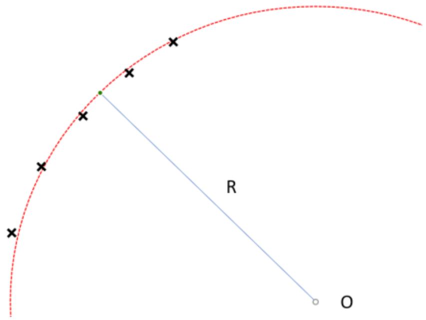
1. Hit decoding – conventional logic
2. Hit clustering – conventional logic
3. Track construction
    1. Edge candidate generation
        - Connect all clusters (nodes) together (edges) while applying geometric constrain
            - Reduced number of edges ~50% at the cost of 0.4% accuracy
    2. Edge candidate classification – GCN (arXiv:1609.02907)
        - model predicts true edge candidates
    3. Construct final track from the edges



The ML algorithm pipeline

- Due to latency limitation all components need to be deployed on FPGA

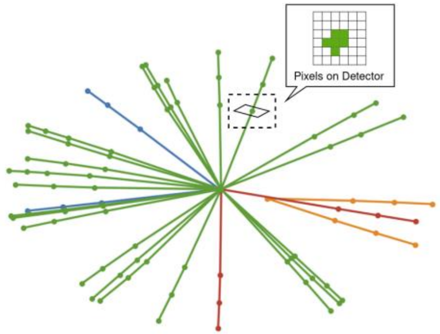
1. Hit decoding – conventional logic
2. Hit clustering – conventional logic
3. Track construction
    1. Edge candidate generation
        - Connect all clusters (nodes) together (edges) while applying geometric constrain
            - Reduced number of edges ~50% at the cost of 0.4% accuracy
    2. Edge candidate classification – GCN (arXiv:1609.02907)
        - model predicts true edge candidates
    3. Construct final track from the edges
4. Transverse momentum p_T prediction
    - Least square method ~ 14% accuracy improvement



The ML algorithm pipeline

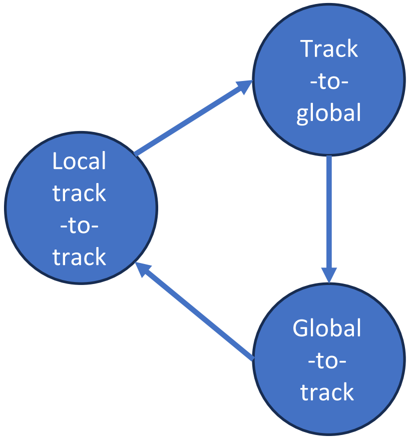
- Due to latency limitation all components need to be deployed on FPGA

1. Hit decoding – conventional logic
2. Hit clustering – conventional logic
3. Track construction
    1. Edge candidate generation
        - Connect all clusters (nodes) together (edges) while applying geometric constrain
            - Reduced number of edges ~50% at the cost of 0.4% accuracy
    2. Edge candidate classification – GCN (arXiv:1609.02907)
        - model predicts true edge candidates
    3. Construct final track from the edges
4. Transverse momentum p_T prediction
    - Least square method ~ 15% accuracy improvement
5. Trigger detection - Bipartite Graph Network with Set Transformer (BGN-ST) (10.1007/978-3-031-26409-2_4)
    - Topological selection of HF signal



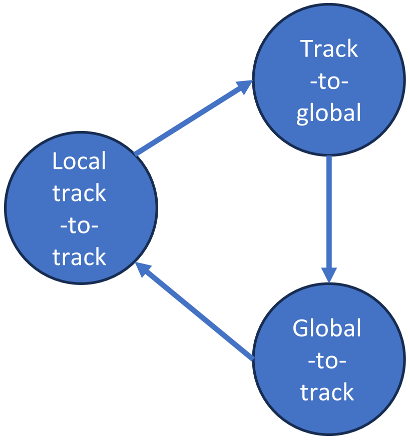
BGN-ST Architecture

- Bipartite Graph Network with Set Transformer (BGN-ST)
- Attention-based algorithm allows the modeling of following information
  1. Local track-to-track interaction
     - Determine whether two tracks share same vertex
  2. Track-to-global interaction
     - Determine collision vertex
  3. Global-to-track interaction
     - Whether track origin vertex is centered around collision vertex

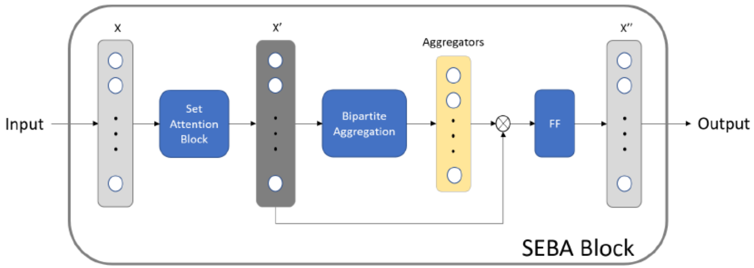


BGN-ST Architecture

- Bipartite Graph Network with Set Transformer (BGN-ST)
- Attention-based algorithm allows the modeling of following information
  1. Local track-to-track interaction
     - Determine whether two tracks share same vertex
  2. Track-to-global interaction
     - Determine collision vertex
  3. Global-to-track interaction
     - Whether track origin vertex is centered around collision vertex

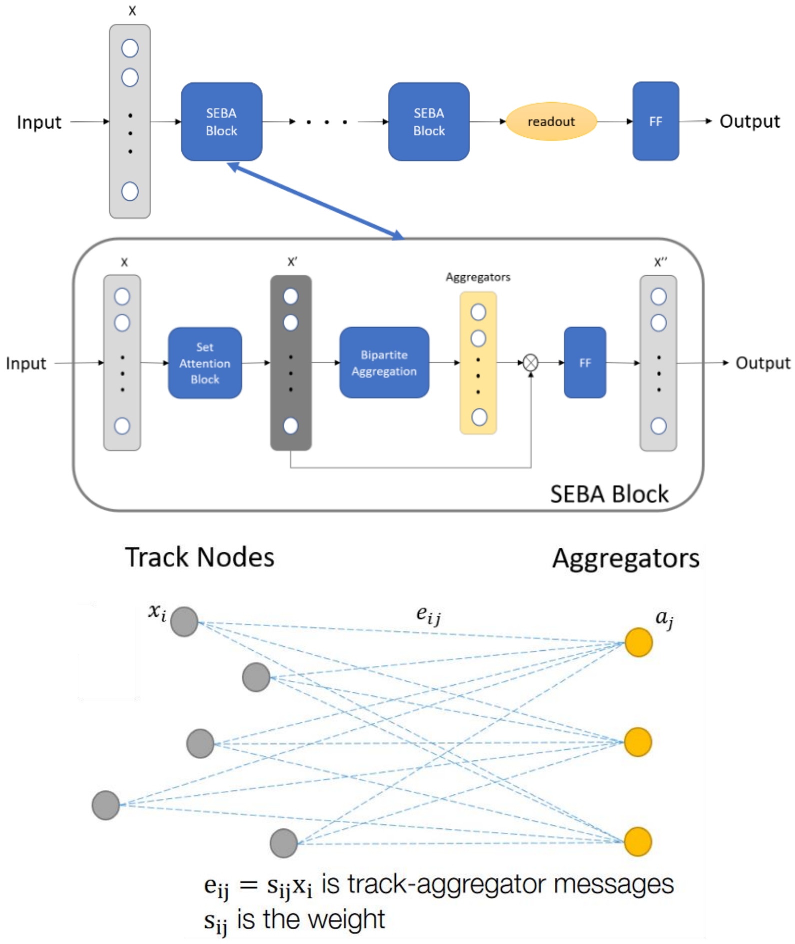


- The loop
  1. Initial track definition
  2. Weights assigned to each link (Attention)
     1. Connection to vertex information
  3. Vertex information go through feed-forward NN
  4. Update track information



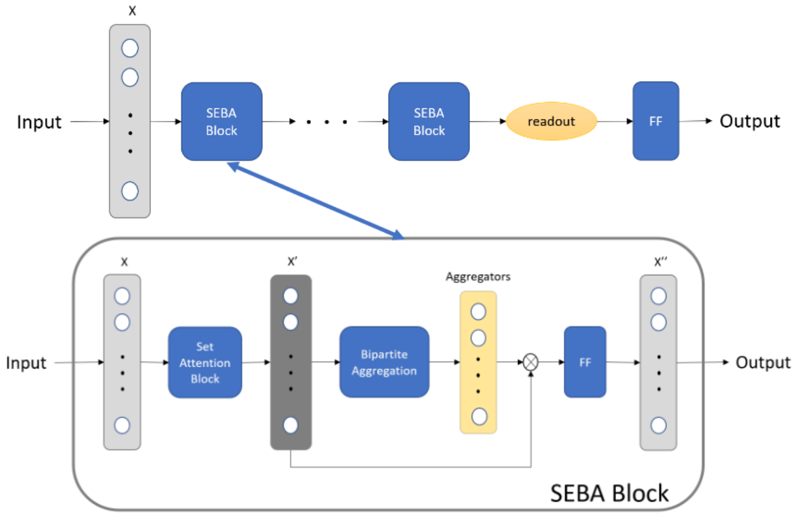
BGN-ST Architecture

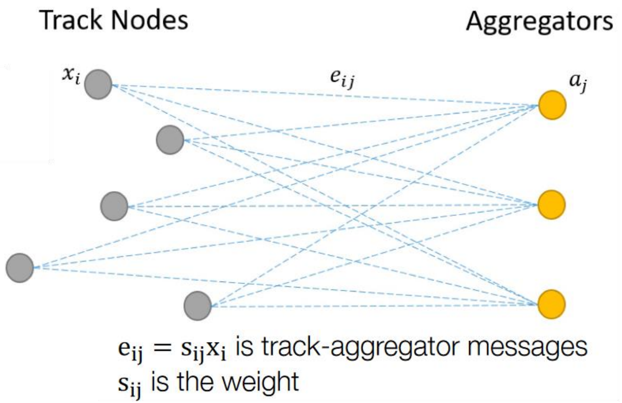
- Main element – SEBA block
  - Multiheaded attention
  - Bipartite aggregation
- Track node input vectors (total 37 features)
  1. 5 hits (MVTX + INTT)
  2. Length of each edge: $L = |\overrightarrow{x_{i+1}} - \overrightarrow{x_i}|$
  3. Angle between edges
  4. Total length of edges
  5. Track radius ($\propto$ track p_T)
- Aggregators
  - Vertices



BGN-ST Architecture

- Main element – SEBA block
  - Multiheaded attention
  - Bipartite aggregation
- Track node input vectors (total 37 features)
  1. 5 hits (MVTX + INTT)
  2. Length of each edge: $L = |\overrightarrow{x_{i+1}} - \overrightarrow{x_i}|$
  3. Angle between edges
  4. Total length of edges
  5. Track radius ($\propto$ track p_T)
- Aggregators
  - Vertices
- Model based on PyTorch and PyTorch Geometric
- Initial training on simulated data from MVTX and INTT
  - On GPU - NVIDIA Titan RTX, A500, and A6000





Set Transformer: arXiv:1810.00825
GarNet: arXiv:1902.07987
PN: arXiv:1902.08570
SAGPool: arXiv:1904.08082
GCN: arXiv:1609.02907
Graph att. Net.: arXiv:2105.14491

Model accuracy studies

1. Triggering on $D^0 \rightarrow \pi K$ (0.1% events)

Note: all Accuracies are calculated on 50% signal/background samples

- 3 MHz collision rate
- 10% HF efficiency (ext. readout)
- 1 kHz available for additional triggers
- 3000 MB rejection needed

Set Transformer: arXiv:1810.00825
GarNet: arXiv:1902.07987
PN: arXiv:1902.08570
SAGPool: arXiv:1904.08082
GCN: arXiv:1609.02907
Graph att. Net.: arXiv:2105.14491

Model accuracy studies

1. Triggering on $D^0 \rightarrow \pi K$ (0.1% events)

Note: all Accuracies are calculated on 50% signal/background samples

	With p_T prediction with LS-radius			Without p_T prediction without radius		
Model	#Parameters	Accuracy	AUC	#Parameters	Accuracy	AUC
Set Transformer	299,266	86.40%	91.92%	298,882	72.04%	78.92%
GarNet	284,210	86.22%	91.81%	284,066	72.59%	79.61%
PN+SAGPool	780,934	86.25%	92.91%	780,678	69.22%	77.18%
BGN-ST	363,426	**87.56%**	**93.22%**	363,170	**74.13%**	**81.81%**

0.1% signal/background ratio		
BG Rejection	Efficiency	Purity
90%	76%	0.75%
99%	23.2%	2.3%

- 3 MHz collision rate
- 10% HF efficiency (ext. readout)
- 1 kHz available for additional triggers
- 3000 MB rejection needed

Model accuracy studies

Set Transformer: arXiv:1810.00825
GarNet: arXiv:1902.07987
PN: arXiv:1902.08570
SAGPool: arXiv:1904.08082
GCN: arXiv:1609.02907
Graph att. Net.: arXiv:2105.14491

1. Triggering on $D^0 \to \pi K$ (0.1% events)

Note: all Accuracies are calculated on 50% signal/background samples

	With p_T prediction with LS-radius			Without p_T prediction without radius		
Model	#Parameters	Accuracy	AUC	#Parameters	Accuracy	AUC
Set Transformer	299,266	86.40%	91.92%	298,882	72.04%	78.92%
GarNet	284,210	86.22%	91.81%	284,066	72.59%	79.61%
PN+SAGPool	780,934	86.25%	92.91%	780,678	69.22%	77.18%
BGN-ST	363,426	**87.56%**	**93.22%**	363,170	**74.13%**	**81.81%**

Estimating p_T from vertex detectors resulted in 14% accuracy increase!

0.1% signal/background ratio		
BG Rejection	Efficiency	Purity
90%	76%	0.75%
99%	23.2%	2.3%

23x rate increase comparing to random selection!

- 3 MHz collision rate
- 10% HF efficiency (ext. readout)
- 1 kHz available for additional triggers
- 3000 MB rejection needed

Set Transformer: arXiv:1810.00825
GarNet: arXiv:1902.07987
PN: arXiv:1902.08570
SAGPool: arXiv:1904.08082
GCN: arXiv:1609.02907
Graph att. Net.: arXiv:2105.14491

Model accuracy studies

1. Triggering on $D^0 \rightarrow \pi K$ (0.1% events)

Note: all Accuracies are calculated on 50% signal/background samples

	With p_T prediction with LS-radius			Without p_T prediction without radius		
Model	#Parameters	Accuracy	AUC	#Parameters	Accuracy	AUC
Set Transformer	299,266	86.40%	91.92%	298,882	72.04%	78.92%
GarNet	284,210	86.22%	91.81%	284,066	72.59%	79.61%
PN+SAGPool	780,934	86.25%	92.91%	780,678	69.22%	77.18%
BGN-ST	363,426	**87.56%**	**93.22%**	363,170	**74.13%**	**81.81%**

Estimating p_T from vertex detectors resulted in 14% accuracy increase!

0.1% signal/background ratio		
BG Rejection	Efficiency	Purity
90%	76%	0.75%
99%	23.2%	2.3%

- 3 MHz collision rate
- 10% HF efficiency (ext. readout)
- 1 kHz available for additional triggers
- 3000 MB rejection needed

23x rate increase comparing to random selection!

- Trigger detection on tracks vs clusters (hit-based)
  - Accuracy: 90.22% (BGN-ST, track construction, model v2) vs 85% (GCN, hit-based)

Model accuracy studies

Set Transformer: arXiv:1810.00825
GarNet: arXiv:1902.07987
PN: arXiv:1902.08570
SAGPool: arXiv:1904.08082
GCN: arXiv:1609.02907
Graph att. Net.: arXiv:2105.14491

1. Triggering on $D^0 \rightarrow \pi K$ (0.1% events)

Note: all Accuracies are calculated on 50% signal/background samples

	With p_T prediction with LS-radius			Without p_T prediction without radius		
Model	#Parameters	Accuracy	AUC	#Parameters	Accuracy	AUC
Set Transformer	299,266	86.40%	91.92%	298,882	72.04%	78.92%
GarNet	284,210	86.22%	91.81%	284,066	72.59%	79.61%
PN+SAGPool	780,934	86.25%	92.91%	780,678	69.22%	77.18%
BGN-ST	363,426	**87.56%**	**93.22%**	363,170	**74.13%**	**81.81%**

Estimating p_T from vertex detectors resulted in 14% accuracy increase!

0.1% signal/background ratio		
BG Rejection	Efficiency	Purity
90%	76%	0.75%
99%	23.2%	2.3%

- 3 MHz collision rate
- 10% HF efficiency (ext. readout)
- 1 kHz available for additional triggers
- 3000 MB rejection needed

23x rate increase comparing to random selection!

- Trigger detection on tracks vs clusters (hit-based)
  - Accuracy: 90.22% (BGN-ST, track construction, model v2) vs 85% (GCN, hit-based)

The charm sample is plentiful, the aim is on rare beauty events

Model accuracy studies

Set Transformer: arXiv:1810.00825
GarNet: arXiv:1902.07987
PN: arXiv:1902.08570
SAGPool: arXiv:1904.08082
GCN: arXiv:1609.02907
Graph att. Net.: arXiv:2105.14491

1. Triggering on $D^0 \rightarrow \pi K$ (0.1% events)

Note: all Accuracies are calculated on 50% signal/background samples

	With p_T prediction with LS-radius			Without p_T prediction without radius		
Model	#Parameters	Accuracy	AUC	#Parameters	Accuracy	AUC
Set Transformer	299,266	86.40%	91.92%	298,882	72.04%	78.92%
GarNet	284,210	86.22%	91.81%	284,066	72.59%	79.61%
PN+SAGPool	780,934	86.25%	92.91%	780,678	69.22%	77.18%
BGN-ST	363,426	**87.56%**	**93.22%**	363,170	**74.13%**	**81.81%**

Estimating p_T from vertex detectors resulted in 14% accuracy increase!

0.1% signal/background ratio		
BG Rejection	Efficiency	Purity
90%	76%	0.75%
99%	23.2%	2.3%

- 3 MHz collision rate
- 10% HF efficiency (ext. readout)
- 1 kHz available for additional triggers
- 3000 MB rejection needed

23x rate increase comparing to random selection!

- Trigger detection on tracks vs clusters (hit-based)
  - Accuracy: 90.22% (BGN-ST, track construction, model v2) vs 85% (GCN, hit-based)

3. Triggering on Beauty decays, (0.05% events)

- No pileup
  - Accuracy: 97.38% (BGN-ST, track construction, model v2)
  - Clusters -> Edge Candidate Generation -> Trigger prediction: Accuracy 91.53% (Graph Attention Network, hit-based)
  - Clusters -> Trigger prediction: Accuracy 90.57% (GarNet, hit-based)

Set Transformer: arXiv:1810.00825
GarNet: arXiv:1902.07987
PN: arXiv:1902.08570
SAGPool: arXiv:1904.08082
GCN: arXiv:1609.02907
Graph att. Net.: arXiv:2105.14491

Model accuracy studies

1. Triggering on $D^0 \to \pi K$ (0.1% events)

Note: all Accuracies are calculated on 50% signal/background samples

	With p_T prediction with LS-radius			Without p_T prediction without radius		
Model	#Parameters	Accuracy	AUC	#Parameters	Accuracy	AUC
Set Transformer	299,266	86.40%	91.92%	298,882	72.04%	78.92%
GarNet	284,210	86.22%	91.81%	284,066	72.59%	79.61%
PN+SAGPool	780,934	86.25%	92.91%	780,678	69.22%	77.18%
BGN-ST	363,426	**87.56%**	**93.22%**	363,170	**74.13%**	**81.81%**

Estimating p_T from vertex detectors resulted in 14% accuracy increase!

0.1% signal/background ratio		
BG Rejection	Efficiency	Purity
90%	76%	0.75%
99%	23.2%	2.3%

- 3 MHz collision rate
- 10% HF efficiency (ext. readout)
- 1 kHz available for additional triggers
- 3000 MB rejection needed

23x rate increase comparing to random selection!

– Trigger detection on tracks vs clusters (hit-based)
  - Accuracy: 90.22% (BGN-ST, track construction, model v2) vs 85% (GCN, hit-based)

3. Triggering on Beauty decays, (0.05% events)

– No pileup
  - Accuracy: 97.38% (BGN-ST, track construction, model v2)
  - Clusters -> Edge Candidate Generation -> Trigger prediction: Accuracy 91.53% (Graph Attention Network, hit-based)
  - Clusters -> Trigger prediction: Accuracy 90.57% (GarNet, hit-based)

Large accuracy increase reconstructing tracks!

Attention provides slight improvement for clusters

Model accuracy studies

Set Transformer: arXiv:1810.00825
GarNet: arXiv:1902.07987
PN: arXiv:1902.08570
SAGPool: arXiv:1904.08082
GCN: arXiv:1609.02907
Graph att. Net.: arXiv:2105.14491

1. Triggering on $D^0 \rightarrow \pi K$ (0.1% events)

Note: all Accuracies are calculated on 50% signal/background samples

	With p_T prediction with LS-radius			Without p_T prediction without radius		
Model	#Parameters	Accuracy	AUC	#Parameters	Accuracy	AUC
Set Transformer	299,266	86.40%	91.92%	298,882	72.04%	78.92%
GarNet	284,210	86.22%	91.81%	284,066	72.59%	79.61%
PN+SAGPool	780,934	86.25%	92.91%	780,678	69.22%	77.18%
BGN-ST	363,426	**87.56%**	**93.22%**	363,170	**74.13%**	**81.81%**

Estimating p_T from vertex detectors resulted in 14% accuracy increase!

0.1% signal/background ratio		
BG Rejection	Efficiency	Purity
90%	76%	0.75%
99%	23.2%	2.3%

- 3 MHz collision rate
- 10% HF efficiency (ext. readout)
- 1 kHz available for additional triggers
- 3000 MB rejection needed

23x rate increase comparing to random selection!

- Trigger detection on tracks vs clusters (hit-based)
  - Accuracy: 90.22% (BGN-ST, track construction, model v2) vs 85% (GCN, hit-based)

3. Triggering on Beauty decays, (0.05% events)
   - No pileup
     - Accuracy: 97.38% (BGN-ST, track construction, model v2)
     - Clusters -> Edge Candidate Generation -> Trigger prediction: Accuracy 91.53% (Graph Attention Network, hit-based)
     - Clusters -> Trigger prediction: Accuracy 90.57% (GarNet, hit-based)
   - Pileup (~350 hits + 65 noise)
     - Clusters -> Trigger Prediction: Accuracy 88.52% (GarNet, hit-based)

Large accuracy increase reconstructing tracks!

Attention provides slight improvement for clusters

Pileup has a small effect!

Generation of the FPGA IP core – two parallel efforts

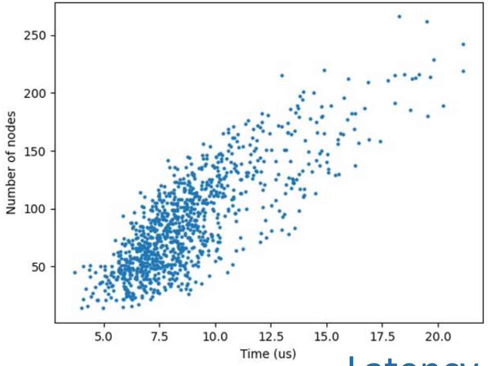
1. Team lead by the Georgia Institute of Technology (GIT)
   - Direct translation using FlowGNN (ArXiv:2204.13103)
   - Goal: 100-200 nodes, 200-500 edges
   1. Implementation of edge classification
      - 92 nodes, 142 edges
      - Measured Start-to-end latency
        - 150 us @ 130 MHz, edge classification v1
        - 8.82 us @ 285 MHz, edge classification v2
   2. Implementation of hit-based model
      - Measured Start-to-end latency
        - 9.2 us @ 180 MHz

Utilization (Alveo U280)	
LUT	194K (14.9%)
FF	214K (8.2%)
BRAM	406 (20.2%)
DSP	488 (5.4%)

Note: target board FLX-712 is half the size of Alveo

1k graphs

Node size



Latency

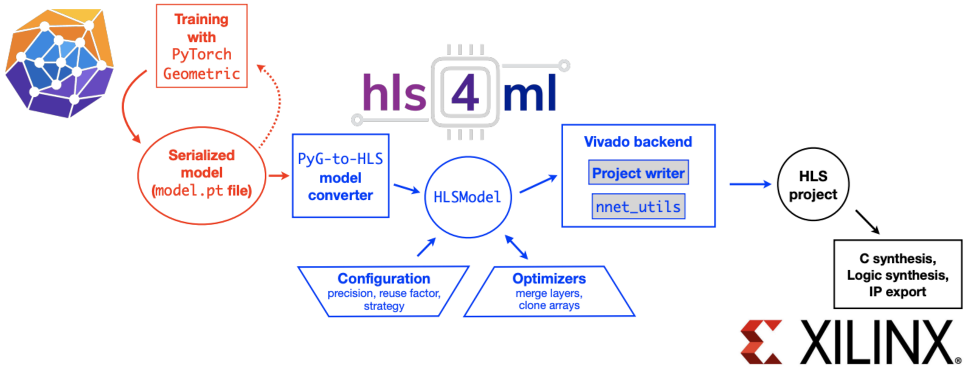
- Detailed latency breakdown and parallelism exploration ongoing
  - The effects of FlowGNN parameters

Close discussion between model developers and FPGA engineers

Generation of the FPGA IP core – two parallel efforts

2. Team lead by the Massachusetts Institute of Technology (MIT) and Fermilab (FNAL)
   - Based on High Level Synthesis for Machine Learning (hls4ml), a generalized python framework for machine learning inference in FPGAs

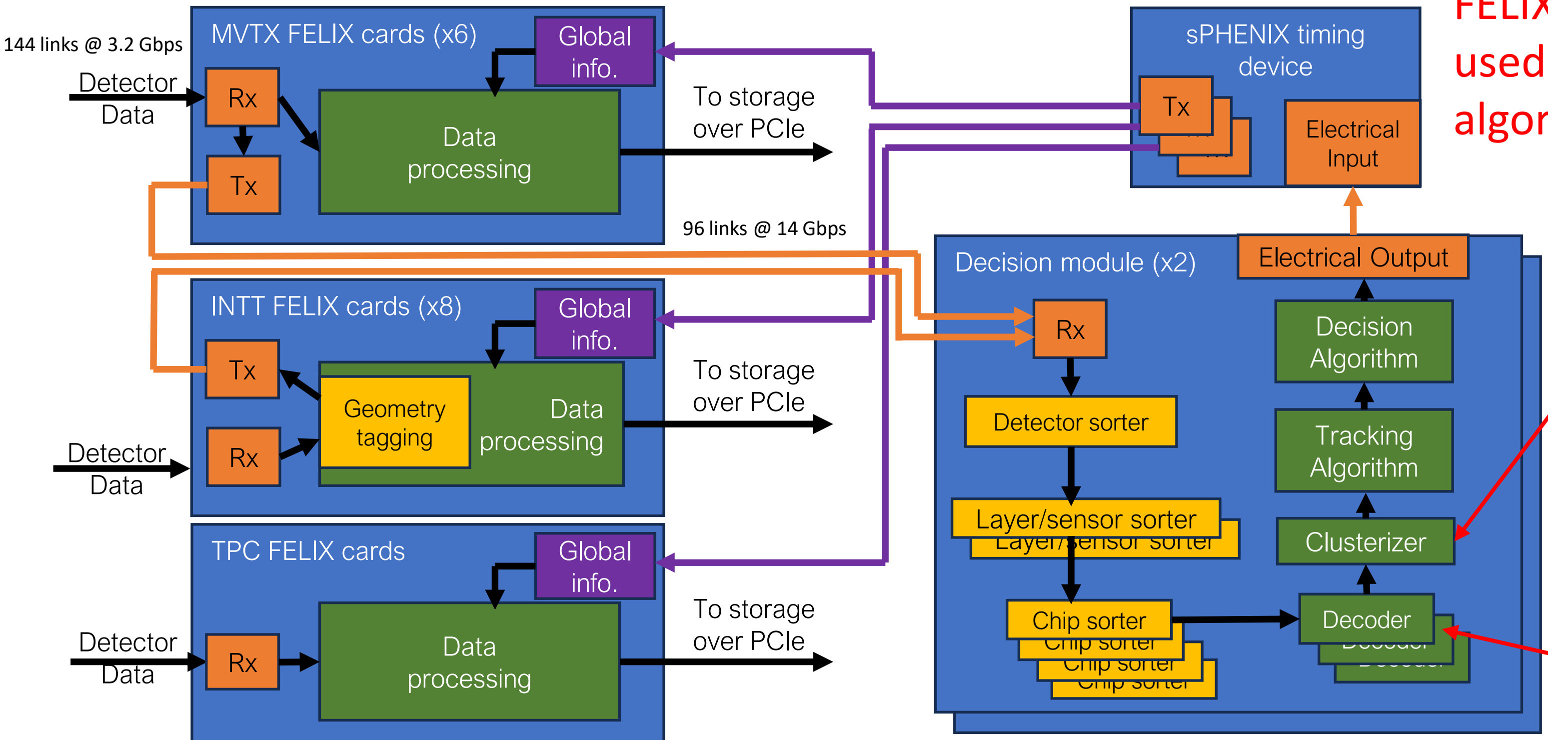
- Unfortunately, PyTorch Geometric is currently not supported in hls4ml
  - Used in our attention-based model (Estimated 3-4 us latency for edge classification v1)
  - Would require extensive customization of current hls4ml -> against its philosophy
  - Developers are working on PyTorch Geometric support and will be released in 2025!
- We moved focus for hit-based beauty model that is based on GarNet which is currently supported in hls4ml



arXiv:2112.02048

arXiv:2103.05579

Realizing the firmware



Schematic courtesy of Cameron Dean

FELIX-712 board is used to host the GNN algorithms

- Clusters can be assembled as they arrive due to column-by-column ALPIDE readout
- 13.5 µm cluster resolution from 21 bits

- Low occupancy allows online FPGA based decoding

Summary and Outlook

- The model has been developed and tested on HF event simulation for sPHENIX
  - provides good precision while analyzing two detector hemispheres independently
  - Current focus on further improvement in model ability to handle pileup event – accuracy and latency studies
- IP core (FPGA code) generation by two teams – FlowGNN and hls4ml
  - Huge progress and improvement of the utilization and latency
  - Iteration with model developers
- All components in the FPGA pipeline are developed
  - Final push to finalize development of each FPGA component putting them together on a single FPGA
    - Very challenging in meeting the utilization and timing constrains
  - FLX-712 boards to serve as AI engine installed in sPHENIX counting house
- A new FLX-182 board arrived to BNL which will be the base for EIC development
  - Probe the possibility of using off-the-shelf card (Alveo etc.)

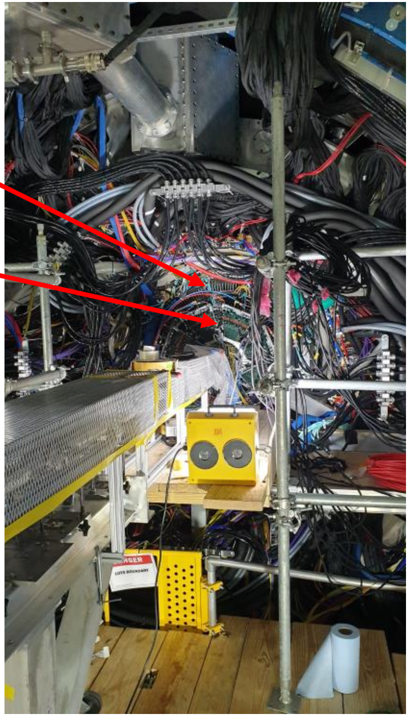
Thank you for your attention

Artificial intelligence and machine learning have the potential to revolutionize our approach collecting, reconstructing and understanding data, and thereby maximizing the discovery potential in the new era of nuclear physics experiments.

MVTX and INTT

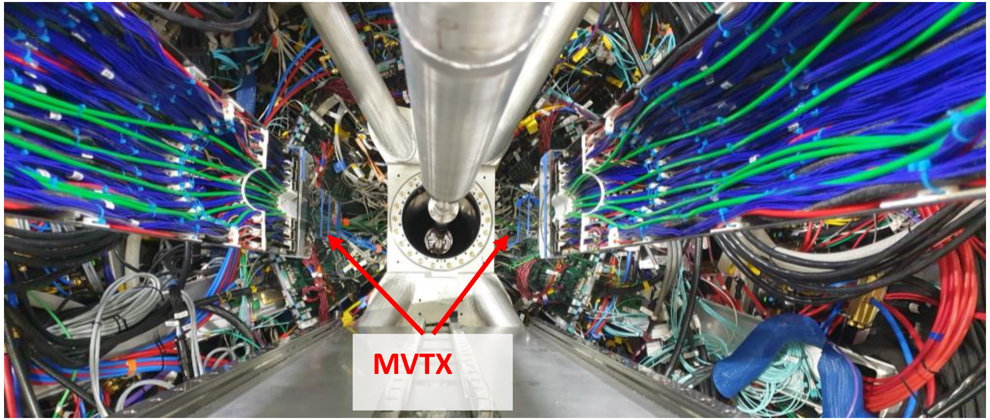
TPC

INTT



MVTX - Active area ~1685 cm^2

- Based on ALICE ITS2 ALPIDE chips, with ATLAS FELIX backend
  - Monolithic Active Pixel Sensors
  - Very fine pitch (27 μm x 29 μm)
  - Event Time resolution ~ 5 μs
  - 3 layers, 48 staves total, 9 chips per stave ~ 230M total channels

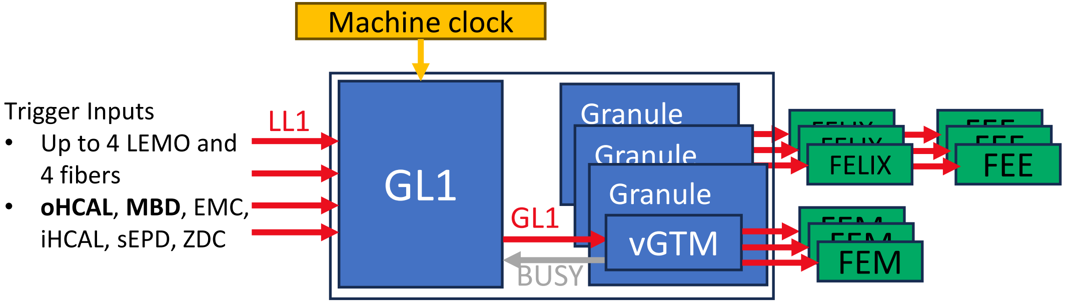


INTT

- Silicon Strip Detector
  - Hamamatsu silicon modules
  - Pitch 78 μm x 16 (or 20) mm
  - Excellent Time resolution ~100 ns
    (100 ns is the RHIC BC time)
  - 2 layers, 56 ladders total, 360k channels

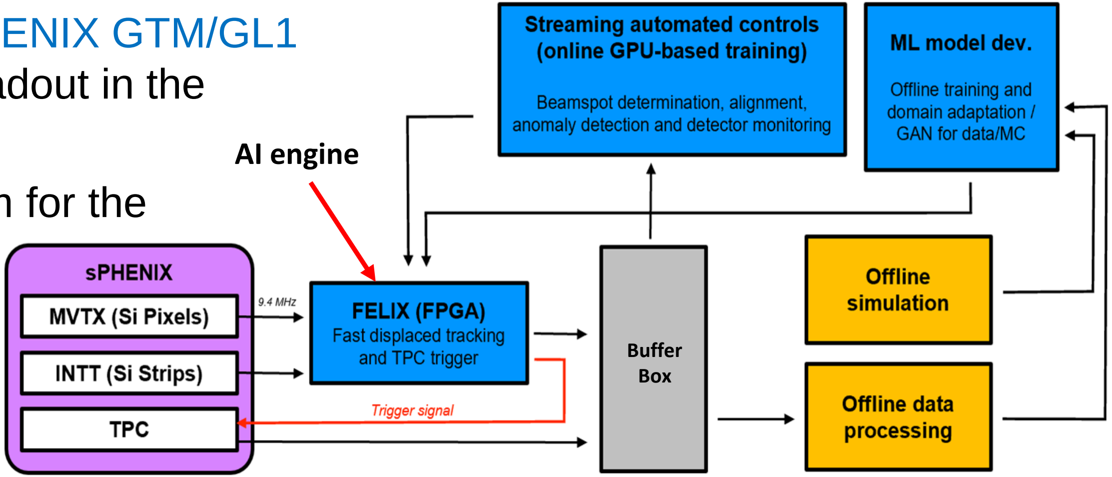
The timing and trigger distribution

- The Global Level 1 Trigger (GL1) and the machine clock is distributed via Granule Timing Module (GTM)
  - GL1 trigger is used by calorimeters and the TPC
  - GL1 transmits clock and trigger to the vGTM, which then transmits it to the FEE
    - vGTM is the adapter to a given detector
    - GL1 is maintaining the BUSY received from vGTM



The DAQ–AI Data Flow

- Motivation to use FELIX board:
  - To reuse the PCIe implementation (16-lane Gen-3) and software tools provided by the FELIX developers
  - on-board FPGA is a Kintex Ultrascale XCKU115FLVF1924-2E
- The decision signal of heavy flavor event from the AI-Engine will be sent out via the LEMO connectors to the sPHENIX GTM/GL1 system to initiate the TPC readout in the triggered mode
- GPU based feed-back system for the beamspot monitoring



The latency constrains

- The TPC buffers can hold up to 30 μs of data
  - The goal of this project is to aim for 10 μs collision-trigger latency to capture the TPC stream

- The latency breakdown
1. MVTX readout window 5 μs – not fixed interaction-readout latency!
2. IR -> Counting house ~0.3 μs (81 m fibres)
3. FELIX -> AI data forward, decoder buffers ~ 0.6 μs (@240 MHz)
4. Clusterizer + Trigger decision (currently 9.2 μs for hit-based model!)
5. AI -> GTM -> TPC FELIX (negligible, all three sits in Counting house)

Motivation – The challenges

- Real-time selection of rare decays of HF particles
  - requires continuous monitoring and adjustment of the
    - beam trajectory (“beam spot”) – in time periods of seconds to hours, the position and shape can change (this will affect the HF the topology)
    - detector alignment, conditions and anomalies
- Adapt AI to continuous learning and changing conditions -> adaptive learning
  - Development of real-time autonomous closed loop adaptive learning system

Predicted timeline

2021

- Project funded by DOE FOA
- Initial simulations constructed
- First data for algorithm training

2022

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

2023

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

2024

- Deploy device at sPHENIX pp/pA run

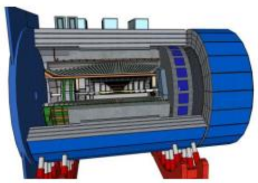
- Take advantage of new technology if required

- Deploy device at EIC

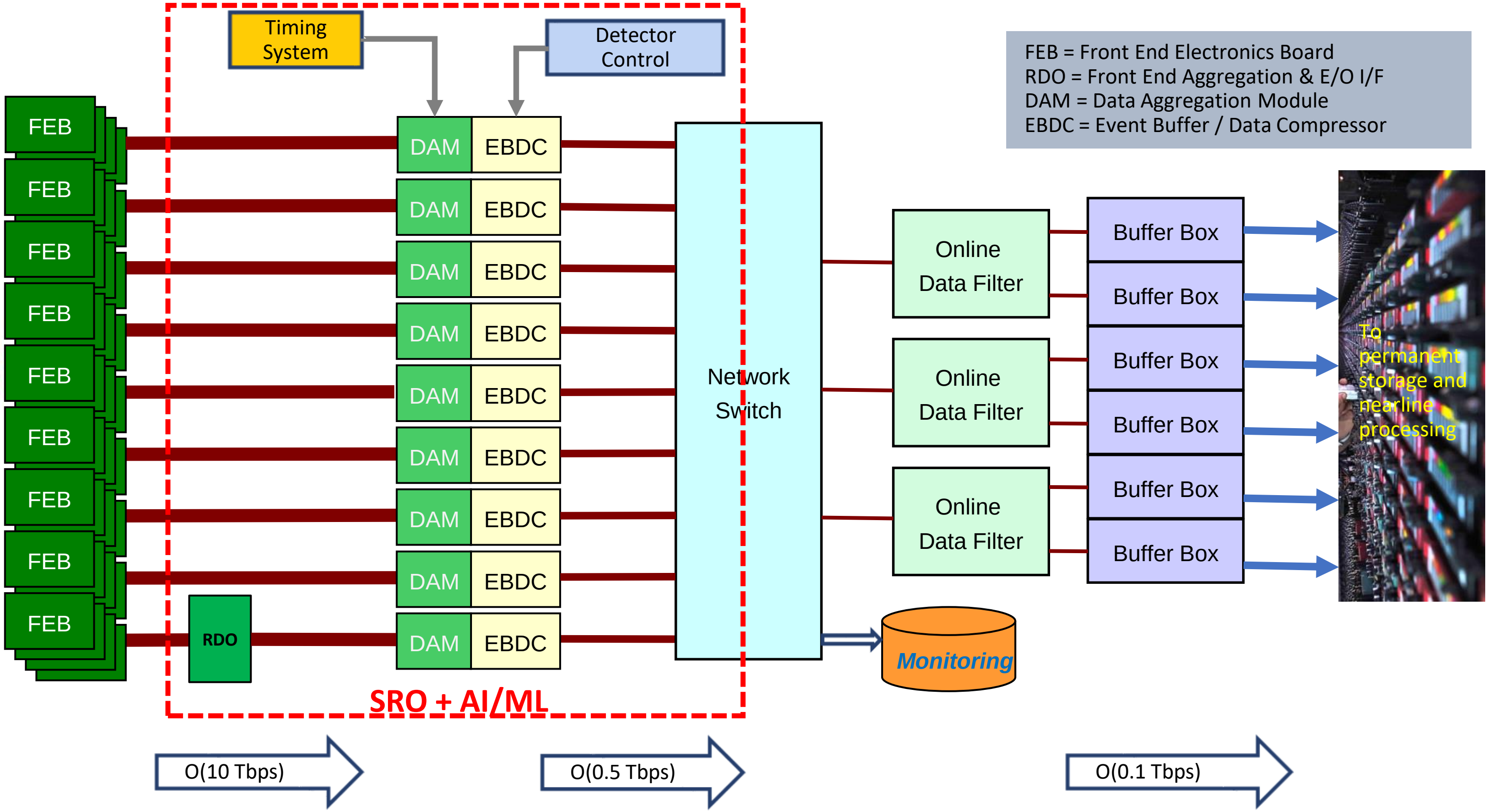
We are here!

From sPHENIX to ePIC: Streaming + AI/ML DAQ

from Jo's talk at ePIC collab. mtg



(sPHENIX) ePIC detector



O(2 Pbps)