Jefferson Lab
Thomas Jefferson National Accelerator Facility

Development of ML FPGA filter for particle identification and tracking in real time

Sergey Furletov
(Jefferson Lab)

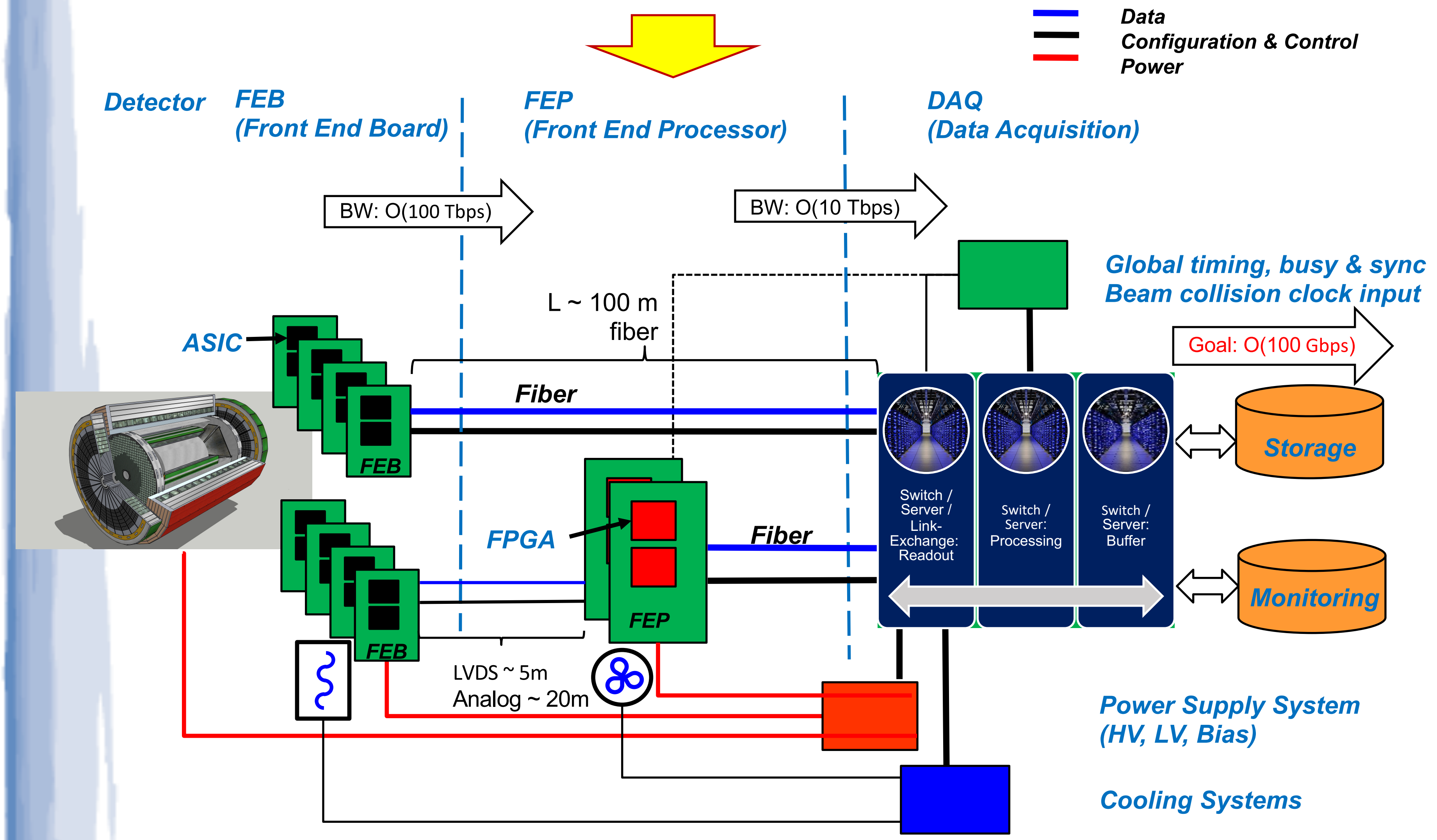
<u>*Team :*</u>
F. Barbosa, L. Belfore, N. Branson, N. Brei, C. Dickover, C. Fanelli,
D. Furletov, L. Jokhovets, D. Lawrence, C. Mei, D. Romanov, K. Shivu

24th IEEE Real Time Conference - ICISE, Quy Nhon, Vietnam

23 Apr 2024

EIC streaming readout as motivation

Jefferson Lab
Thomas Jefferson National Accelerator Facility

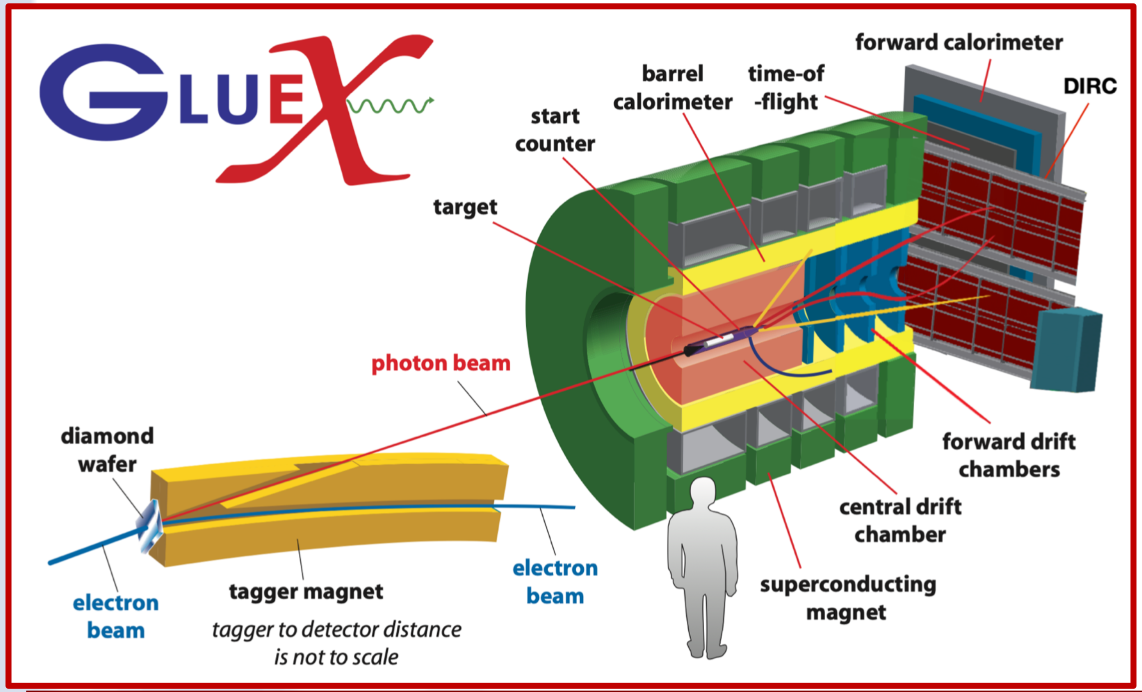


- The correct location for the ML on the FPGA filter is called "FEP" in this figure.
- This gives us a chance to reduce traffic earlier.
- Allows us to touch physics: ML brings intelligence to L1.
- However, it is now unclear how far we can go with physics at the FPGA.
- Initially, we can start in pass-through mode.
- Then we can add background rejection.
- Later we can add filtering processes with the largest cross section.
- In case of problems with output traffic, we can add a selector for low cross section processes.
- The ML-on-FPGA solution complements the purely computer-based solution and mitigates DAQ performance risks.

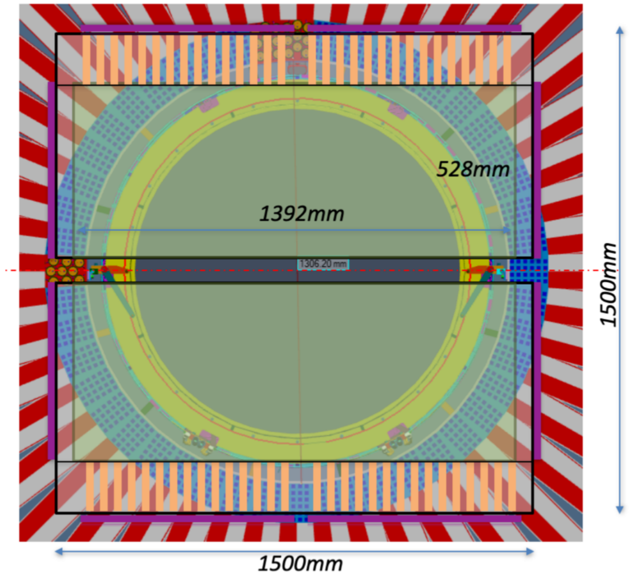
Motivation for GlueX

Jefferson Lab
Thomas Jefferson National Accelerator Facility

- *Real-time data processing is a frontier field in experimental particle physics.*
- *The growing computational power of modern FPGA boards allows us to add more sophisticated algorithms for real-time data processing.*
- *Many tasks, such as tracking and particle identification, could be solved using modern Machine Learning (ML) algorithms which are naturally suited for FPGA architectures.*
- *The work described in this report aims to test ML-FPGA algorithms in a triggered data acquisition system, as well as in streaming data acquisition, such as in the future EIC collider.*
- *The first target is the GlueX experiment, with a plan to build a Transition Radiation Detector (TRD) based on GEM technology (GEM-TRD), to improve the electron-pion separation in the GlueX experiment. It will allow to study precisely reactions with electron-positron pairs in the final states.*



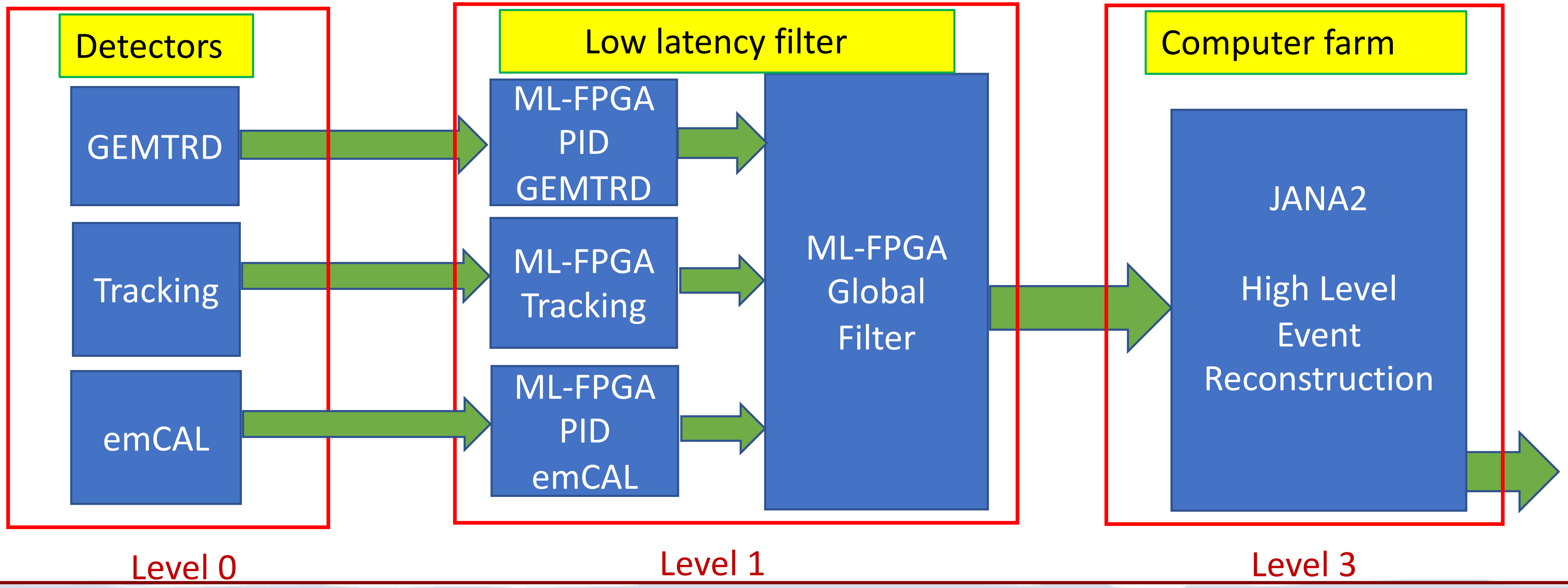
GEMTRD



- *GEM-TRD is supposed to be installed in front of the DIRC detector.*
- *Hall D is dedicated to the operation with a linearly-polarized photon beam produced by ~12 GeV electrons from CEBAF at Jefferson Lab.*
- *Typical L1 trigger rate 40-70 kHz*
- *Data rate 0.7 – 1.2 GB/s*
- *L1 Trigger latency 3.5 us.*

Generic EIC R&D project RD15, ML-(on)-FPGA

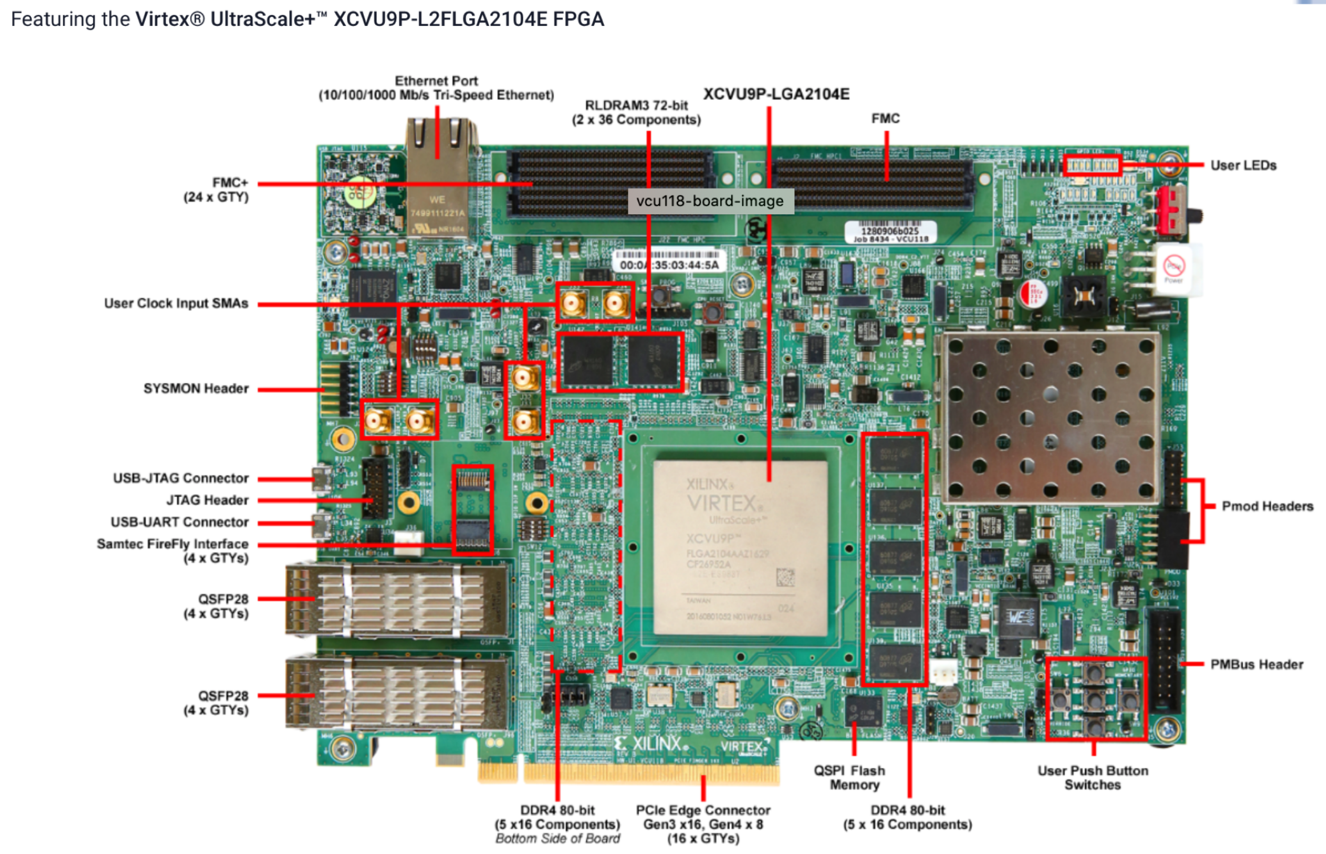
- *Usually, several PID detectors are used in an experiment.*
- *For example, the GEM-TRD and e/m-calorimeter, both provide separation of electrons and hadrons.*
- *Summation and processing of joint data from both detectors at the early stages will increase the identification power of these detectors compared to independent identification.*
- *To test the "global PID" performance we work on developing the ML-FPGA setup for real-time data pre-processing.*
- *The setup consists of several PID and tracking detectors: emCAL, GEMTRD, GEM tracker.*
- *Preprocessed data from both detectors including decision on the particle type will be transferred to another ML-FPGA board with neural network for global PID decision.*
- *The global filter transfers data to off-line computer farm, running JANA2 software.*



FPGA test board for ML

Jefferson Lab
Thomas Jefferson National Accelerator Facility

- At an early stage in this project, as hardware to test ML algorithms on FPGA , we use a standard Xilinx evaluation boards rather than developing a customized FPGA board. These boards have functions and interfaces sufficient for proof of principle of ML-FPGA.
- The Xilinx evaluation board includes the Xilinx XCVU9P and 6,840 DSP slices. Each includes a hardwired optimized multiply unit and collectively offers a peak theoretical performance in excess of 1 Tera multiplications per second.
- Second, the internal organization can be optimized to the specific computational problem. The internal data processing architecture can support deep computational pipelines offering high throughputs.
- Third, the FPGA supports high speed I/O interfaces including Ethernet and 180 high speed transceivers that can operate in excess of 30 Gbps.

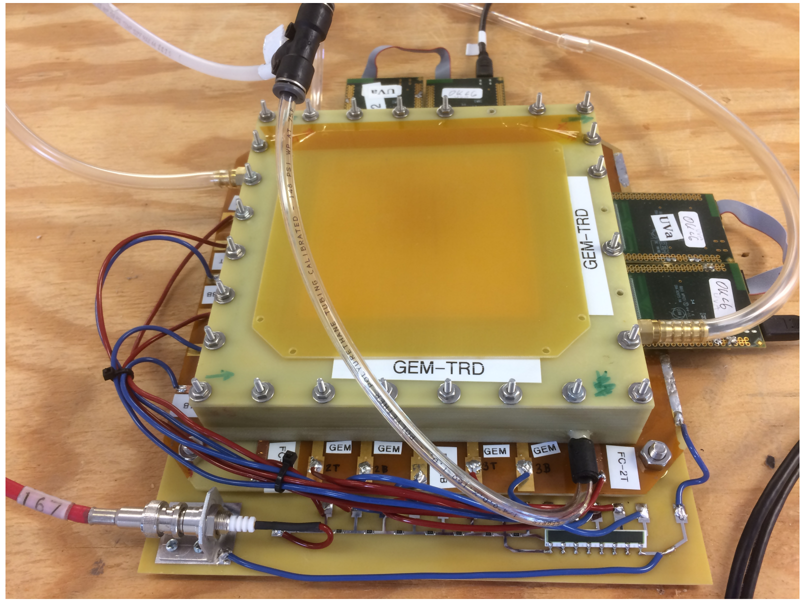


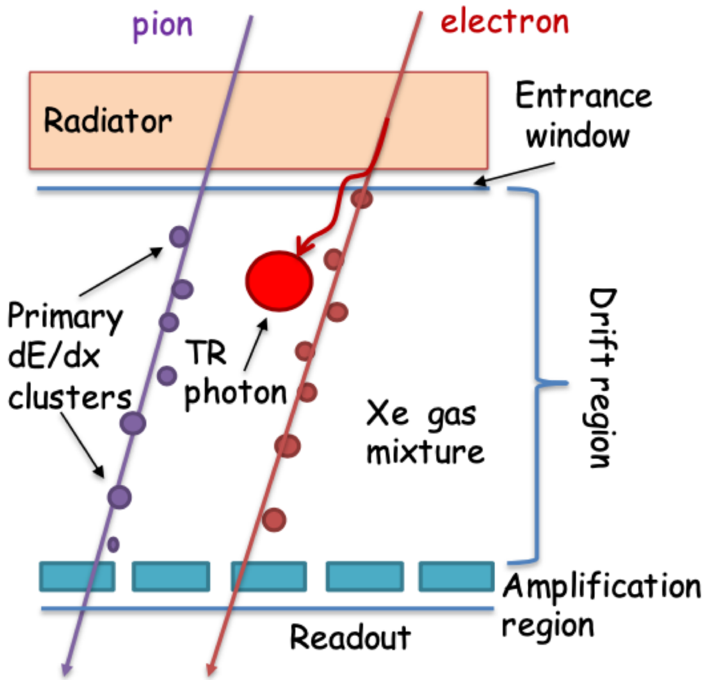
Xilinx Virtex® UltraScale+™

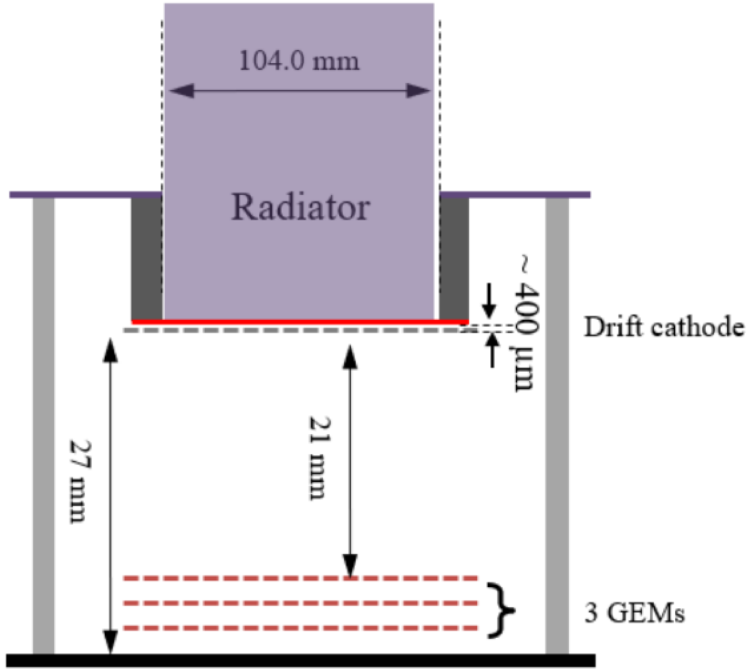
GEM-TRD prototype for EIC R&D

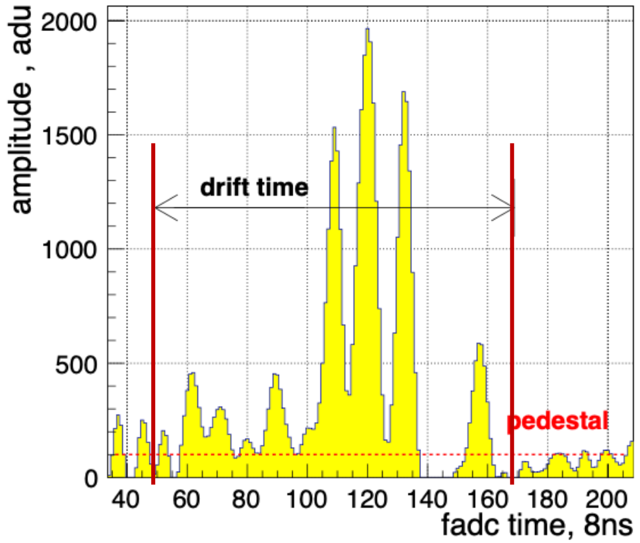
- To demonstrate the operating principle of the ML FPGA, we use the existing setup
- from the EIC detector R&D project
- A test module was built at the University of Virginia
- The prototype of GEMTRD/T module has a size of 10 cm × 10 cm with a corresponding to a total of 512 channels for X/Y coordinates.
- The readout is based on flash ADC system developed at JLAB (fADC125) @125 MHz sampling.

- GEM-TRD provides e/hadron separation and tracking







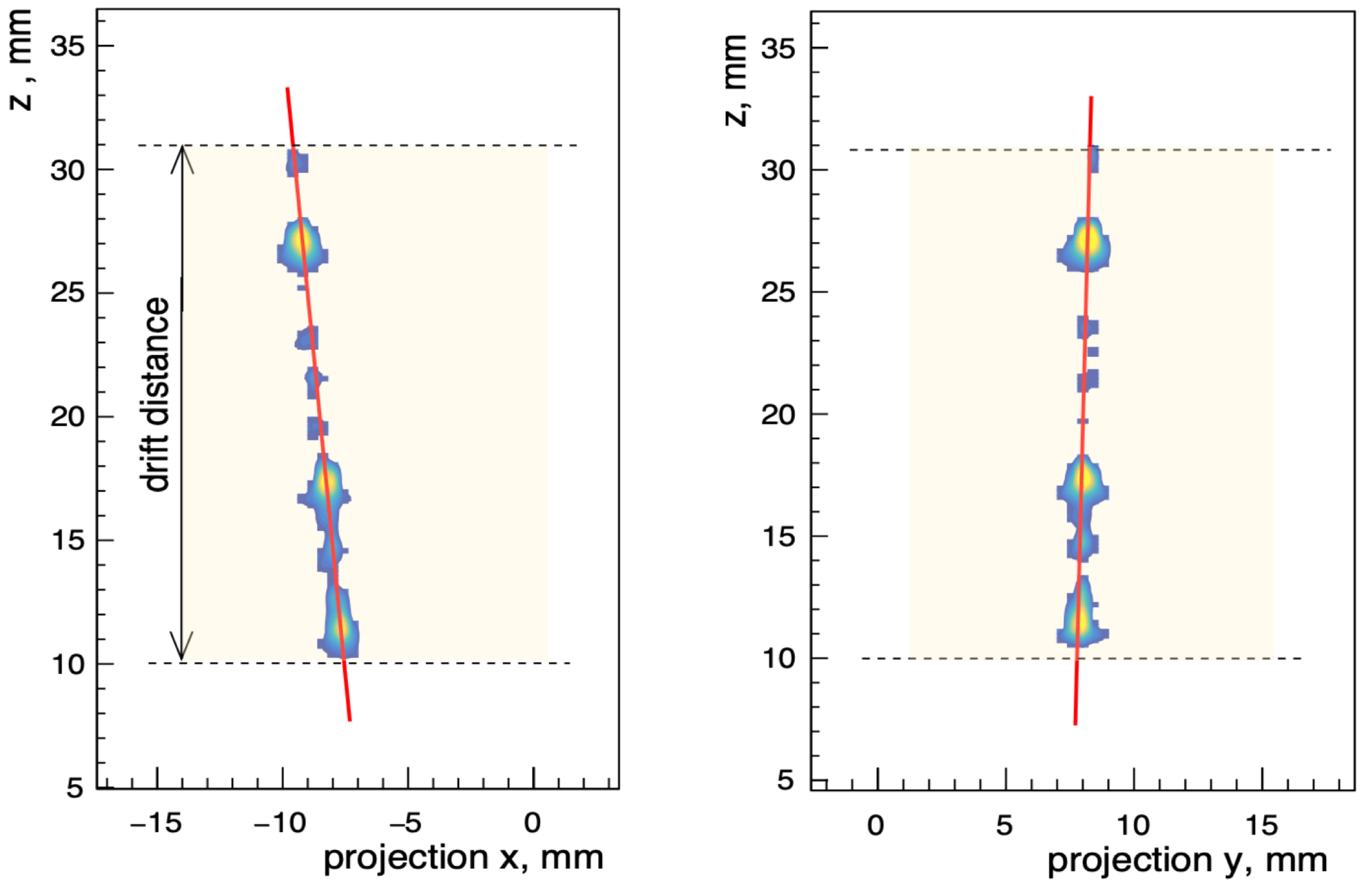


GEM-TRD principle

Jefferson Lab
Thomas Jefferson National Accelerator Facility

- ❑ *The e/pion separation in the GEM-TRD detector is based on counting the ionization along the particle track.*
- ❑ *For electrons, the ionization is higher due to the absorption of transition radiation photons*
- ❑ *So, particle identification with TRD consists of several steps:*
  - The first step is to cluster the incoming signals and create "hits".
  - The next is "pattern recognition" - sorting hits by track.
  - Finding a track
  - Ionization measurement along a track
  - As a bonus, TRD will provide a track segment for the global tracking system.

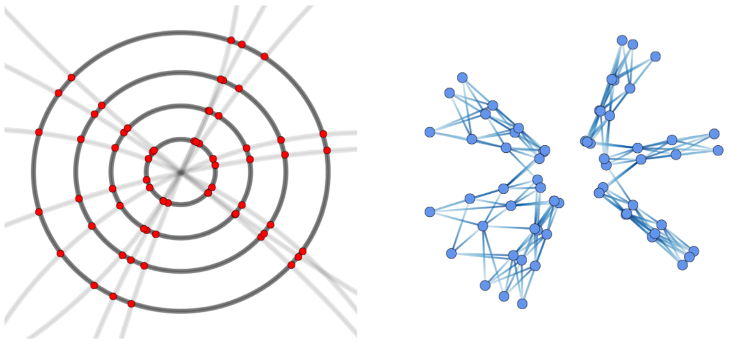
GEM-TRD can work as micro TPC, providing 3D track segments



GEMTRD tracks

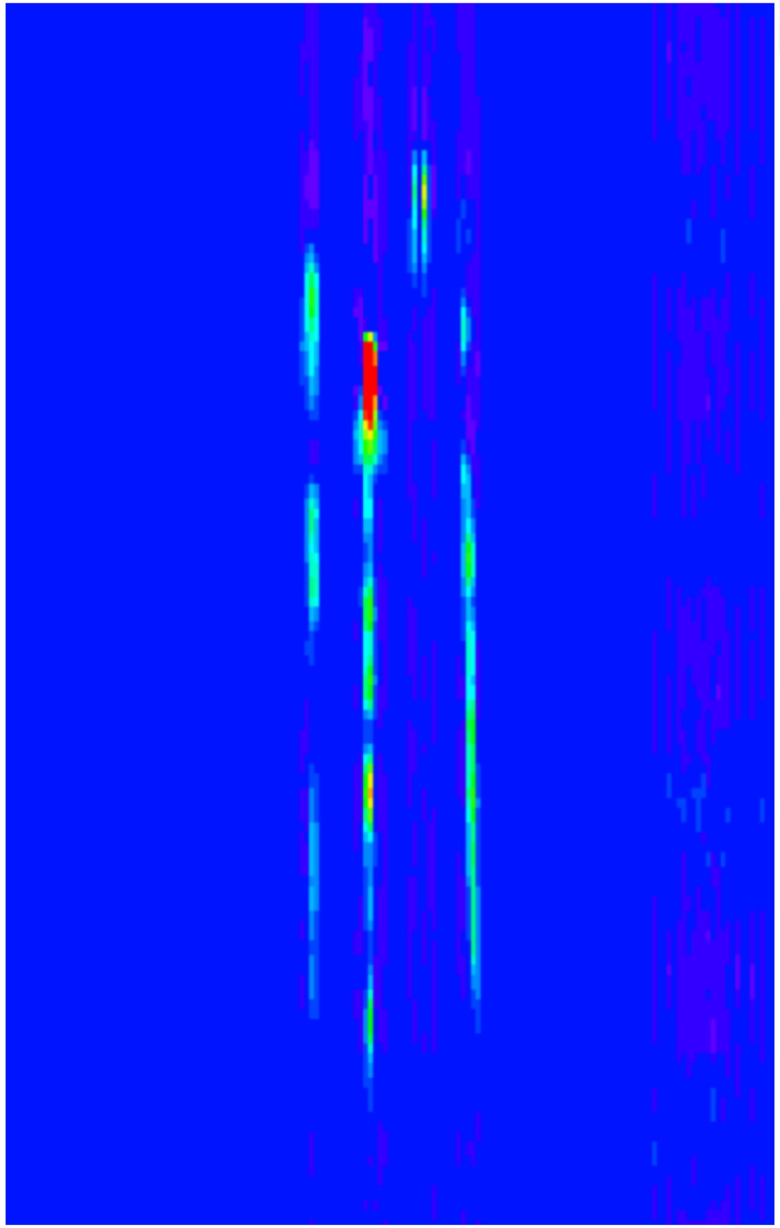
Jefferson Lab
Thomas Jefferson National Accelerator Facility

- *In a real experiment, GEMTRD will have multiple tracks.*
- *So we also need a fast algorithm for pattern recognition*
- *As well as for track fitting.*
- *The decision was made to try the Graph Neural Network (GNN) for pattern recognition.*
- *And a recurrent neural network – LSTM, for track fitting.*



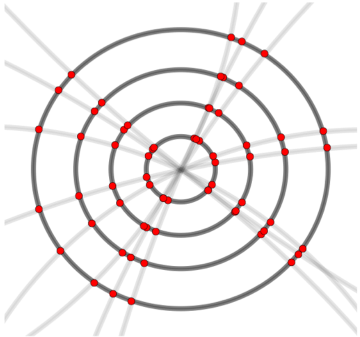
Javier Duarte
arXiv:2012.01249v2 [hep-ph] 7 Dec 2020

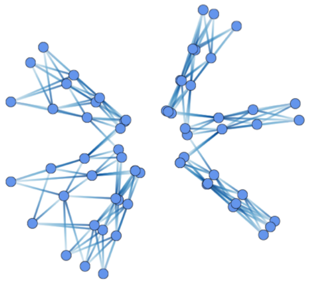
- HEP advanced tracking algorithms at the exascale **(Project Exa.TrkX)**
- https://exatrkx.github.io/



GEMTRD tracks

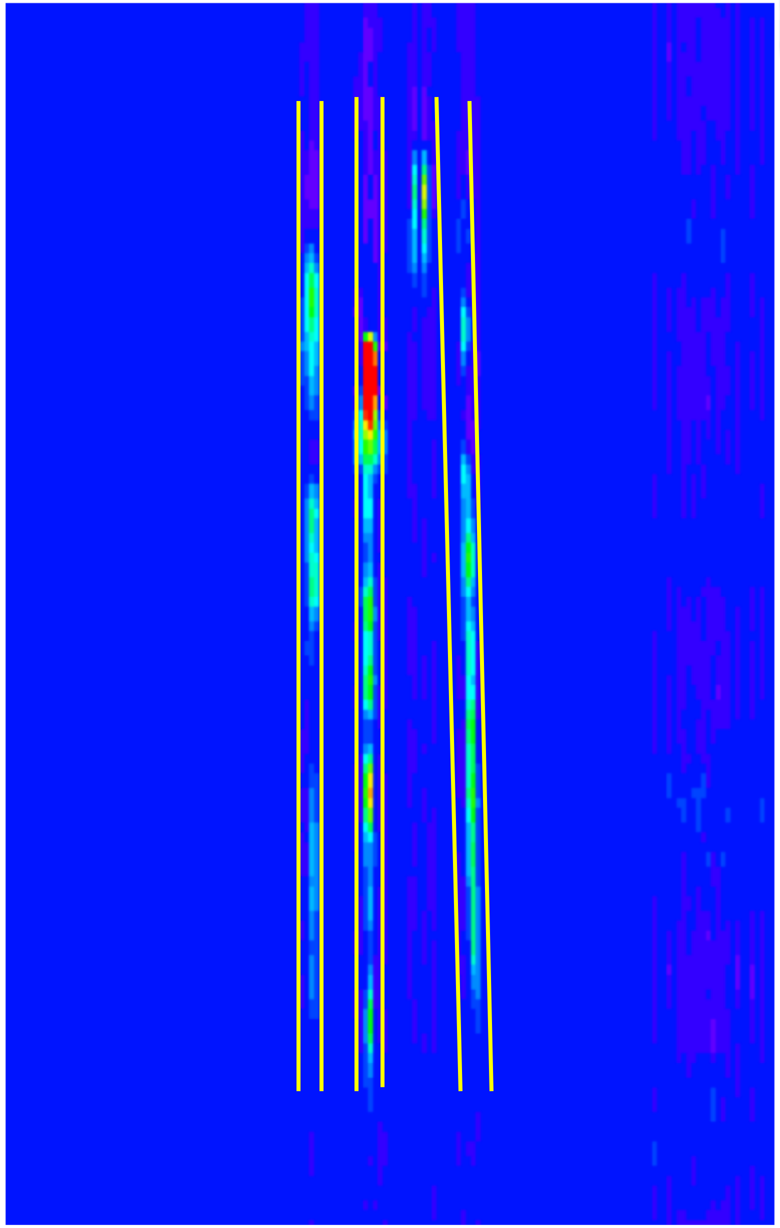
- In a real experiment, GEMTRD will have multiple tracks.
- So we also need a fast algorithm for pattern recognition
- As well as for track fitting.
- The decision was made to try the Graph Neural Network (GNN) for pattern recognition.
- And a recurrent neural network – LSTM, for track fitting.
- PID is based on measuring ionization along the track.





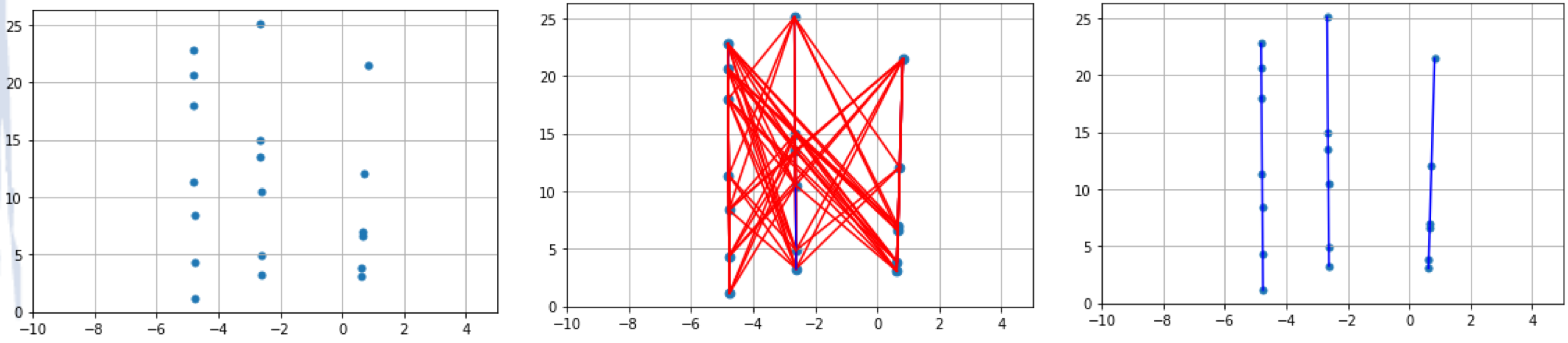
Javier Duarte
arXiv:2012.01249v2 [hep-ph] 7 Dec 2020

- HEP advanced tracking algorithms at the exascale **(Project Exa.TrkX)**
- https://exatrkx.github.io/



GNN for pattern recognition

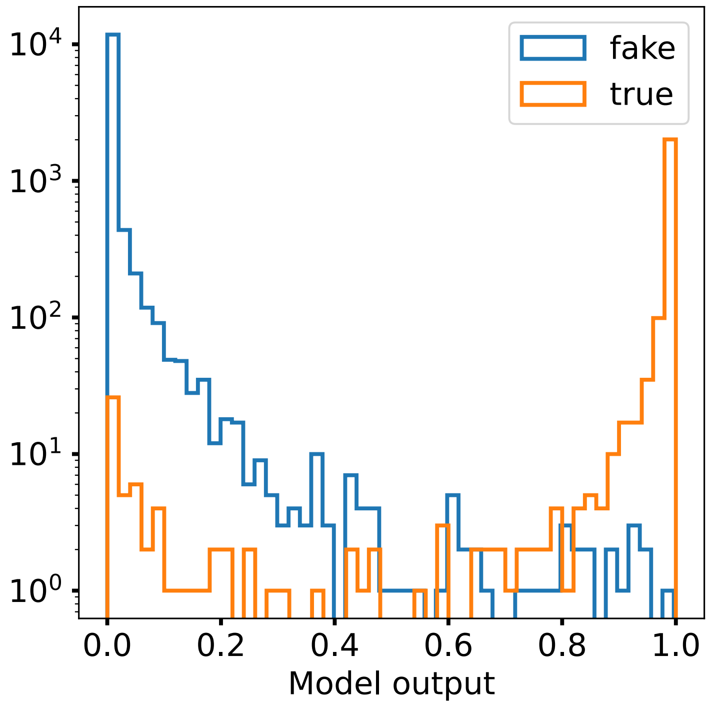
- *Graph Neural Networks (GNNs) designed for the tasks of hit classification and segment classification.*
  - These models read a graph of connected hits and compute features on the nodes and edges.
- *The input and output of GNN is a graph with a number of features for nodes and edges.*
  - In our case we use the edge classification
- *A complete graph on N vertices contains N(N - 1)/2 edges.*
  - This will require a lot of resources which are limited in FPGA.
- *To keep resources under control, we can construct the graph for a specific geometry and limit the minimum particle momentum.*
- *In our case we have a straight track segments, with a quite narrow angular distribution ~15 degree.*
- *Thus, for the input hits (left), we connect only those edges that satisfy our geometry and the momentum of most tracks (middle)*
- *The trained GNN processes the input graph and sets the probability for each edge as output.*
- *The right plot shows edges with a probability greater than 0.7*

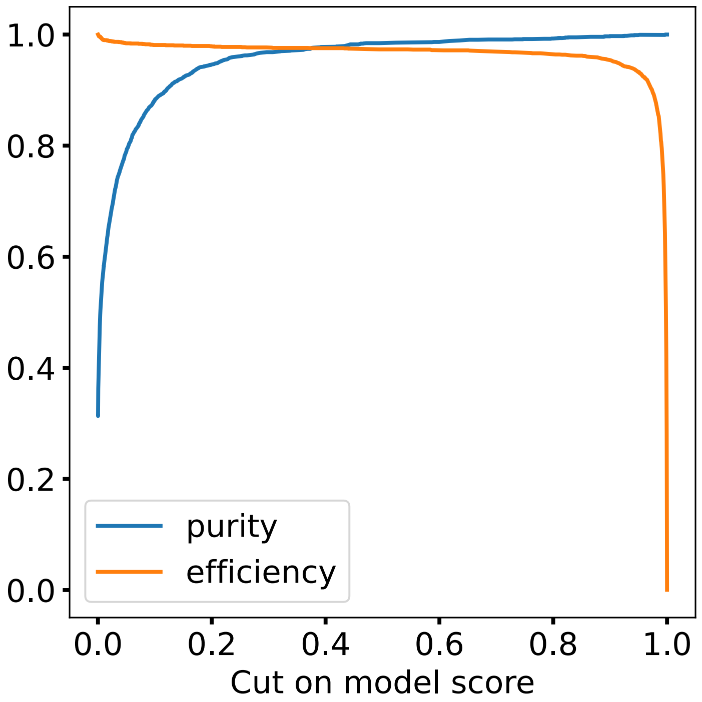


GNN performance

Jefferson Lab
Thomas Jefferson National Accelerator Facility

- *This type of graph neural network is not yet supported in HLS4ML.*
- *So we did a manual conversion first to C++ and then to Verilog using Vitis_HLS.*
- *This neural network has not been optimized, so it consumes a lot of resources - 70% of DSPs, (4651 of 6840).*
  - At the moment it can serve up to 21 hits and 42 edges, or , in our case (GEM-TRD), it will be 3-5 tracks.
- *However, it performs all calculations in ~3 µs (left plot) (thanks to Ben Raydo), providing good purity and efficiency (right plot).*



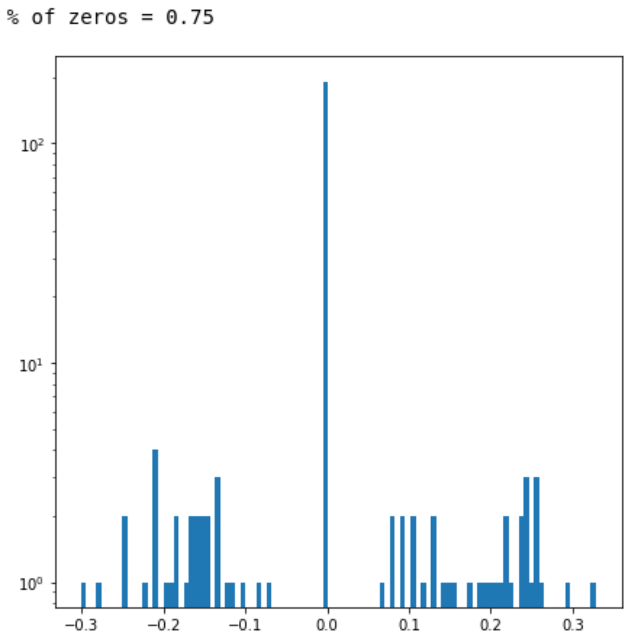


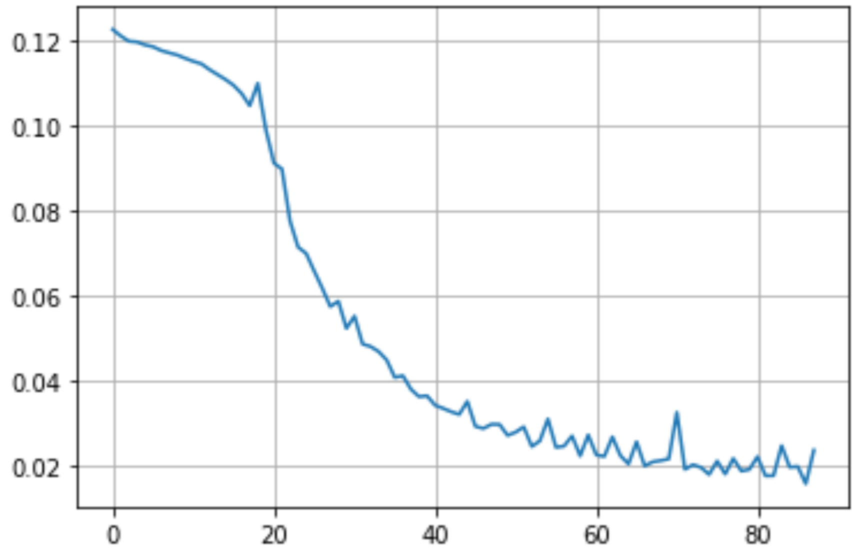
Modules & Loops	Issue Type	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
gnn2dfs2		-	589	2.945E3	-	590	-	no	42	4424	394036	2519454	0
toGraph		-	499	2.495E3	-	497	-	dataflow	42	4424	391308	2515320	0
fromGraph		-	331	1.655E3	-	1	-	yes	0	0	197686	1673583	0
gnn2dfs_loc_1		-	496	2.480E3	-	496	-	no	42	4422	172620	785082	0
toGraph_Block_split100_proc205		-	480	2.400E3	-	480	-	no	0	2	7226	49627	0
VITIS_LOOP_1365_1		-	63	315.000	3	-	21	no	-	-	-	-	-
VITIS_LOOP_1400_3		-	22	110.000	3	1	21	yes	-	-	-	-	-

RNN/LSTM for track fit

Jefferson Lab
Thomas Jefferson National Accelerator Facility

- *The hits sorted by tracks from the pattern recognition GNN are fed into another neural network trained to fit the tracks.*
- *We tested DNN and RNN/LSTM neural networks. (thanks to Dylan Rankin for help)*
- *DNN is faster, but LSTM seems to be more reliable in the case of a stochastic distribution of hits on the track.*
  - The work on optimization of NN is ongoing.
- *The LSTM network after pruning consumes 19% of the DSP resources and has a latency of 1 μs.*





+ Latency (clock cycles):
  * Summary:

Latency min	Latency max	Interval min	Interval max	Pipeline Type
213	213	208	208	function

== Utilization Estimates

* Summary:

Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	6	-
FIFO	-	-	-	-	-
Instance	64	4271	23258	163672	-
Memory	-	-	-	-	-
Multiplexer	-	-	-	955	-
Register	-	-	2323	-	-
Total	64	4271	25581	164633	0
Available SLR	1440	2280	788160	394080	320
Utilization SLR (%)	4	187	3	41	0
Available	4320	6840	2364480	1182240	960
Utilization (%)	1	62	1	13	0

== Utilization Estimates

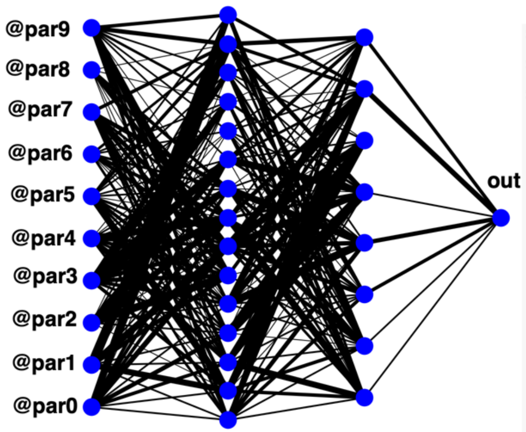
* Summary:

Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	6	-
FIFO	-	-	-	-	-
Instance	64	1308	12199	53194	-
Memory	-	-	-	-	-
Multiplexer	-	-	-	955	-
Register	-	-	2147	-	-
Total	64	1308	14346	54155	0
Available SLR	1440	2280	788160	394080	320
Utilization SLR (%)	4	57	1	13	0
Available	4320	6840	2364480	1182240	960
Utilization (%)	1	19	~0	4	0

MLP neural network for PID

Jefferson Lab
Thomas Jefferson National Accelerator Facility

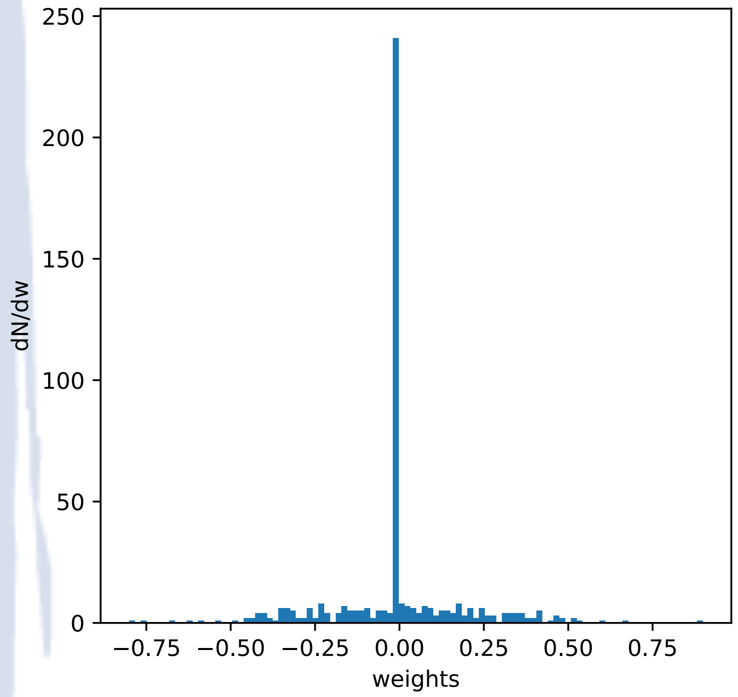
- After the track is fit, the ionization along the track can be counted.
- The distance along the track is divided into 10-20 bins, and the ionization energy in these bins is fed to the input of the MLP neural network.
- Typically neural network weights often have many zeros, thus, it is possible to reduce the size of the network by removing weights close to zero (~50%)
- The network performance near the working value of 90% efficiency.

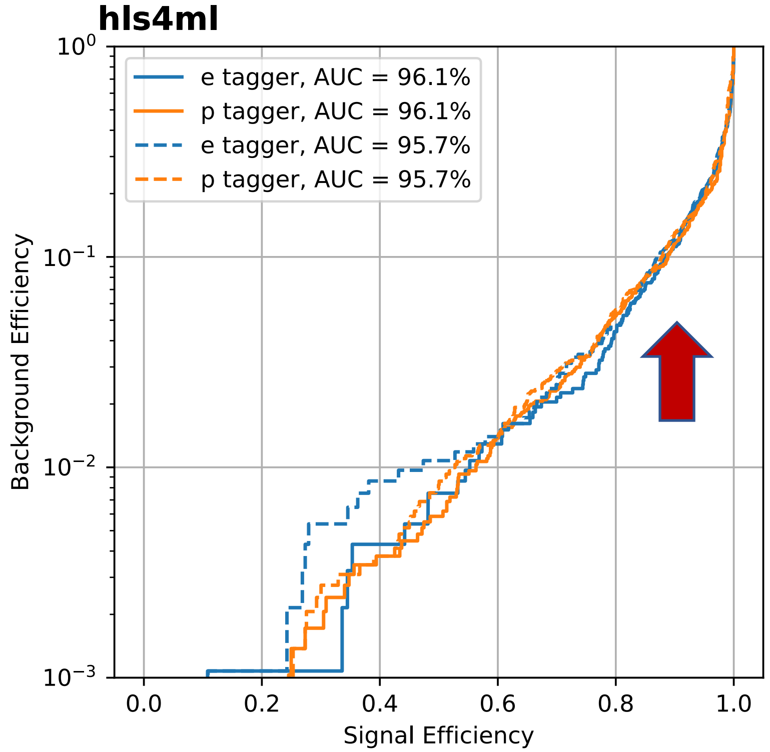



```
================================================================
== Performance Estimates
================================================================
+ Timing (ns):
    * Summary:
    +--------+--------+----------+------------+
    |  Clock | Target| Estimated| Uncertainty|
    +--------+--------+----------+------------+
    |ap_clk  |    5.00|     3.968|        0.62|
    +--------+--------+----------+------------+

+ Latency (clock cycles):
    * Summary:
    +-----+-----+-----+-----+----------+
    |  Latency  |  Interval | Pipeline |
    | min | max | min | max |   Type   |
    +-----+-----+-----+-----+----------+
    |   13|   13|    1|    1| function |
    +-----+-----+-----+-----+----------+

================================================================
== Utilization Estimates
================================================================
* Summary:
+-----------------+---------+-------+-------+--------+-----+
|       Name      | BRAM_18K| DSP48E|   FF  |   LUT  | URAM|
+-----------------+---------+-------+-------+--------+-----+
|DSP              |        -|      -|      -|       -|    -|
|Expression       |        -|      -|      0|       6|    -|
|FIFO             |        -|      -|      -|       -|    -|
|Instance         |       16|    233|   1241|   11742|    -|
|Memory           |        -|      -|      -|       -|    -|
|Multiplexer      |        -|      -|      -|      36|    -|
|Register         |        -|      -|   1235|       -|    -|
+-----------------+---------+-------+-------+--------+-----+
|Total            |       16|    233|   2476|   11784|    0|
+-----------------+---------+-------+-------+--------+-----+
|Utilization (%)  |       ~0|      3|     ~0|      ~0|    0|
+-----------------+---------+-------+-------+--------+-----+
```

Latency = 65ns

II = 5ns

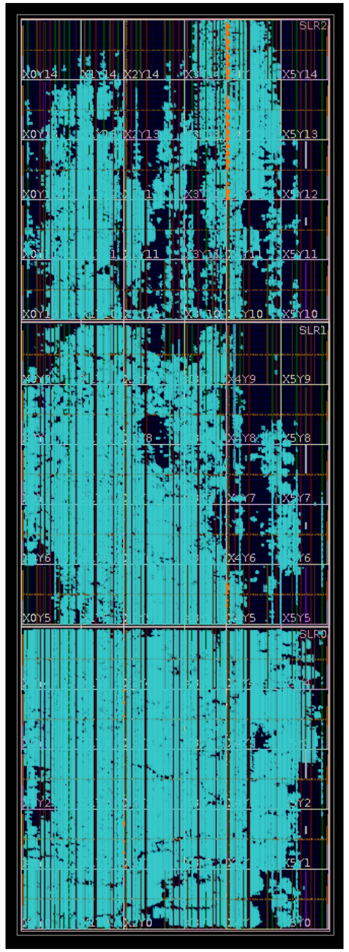


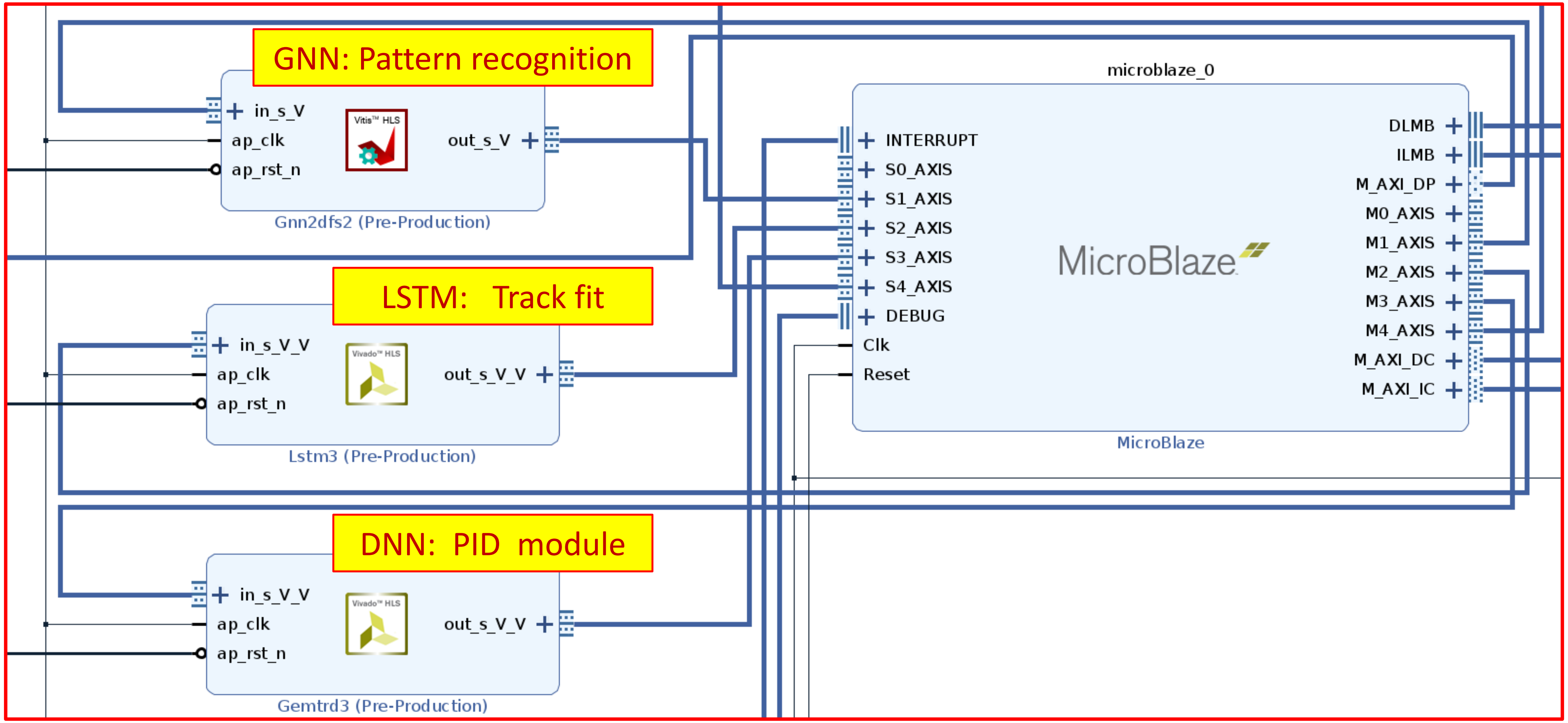


DSP utilization 3%

FPGA test bench (vcu118 board)

Jefferson Lab
Thomas Jefferson National Accelerator Facility



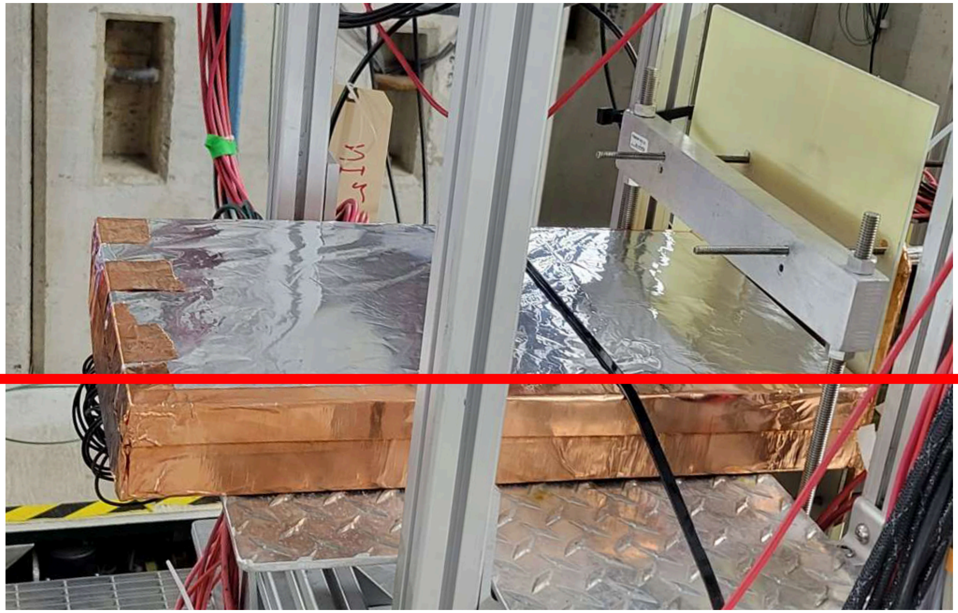


- ❑ *Several version of IPs were synthesized and tested on FPGAs.*
- ❑ *The logic test was performed with the MicroBlaze processor.*
- ❑ *I/O data transfer is carried out through the ETH interface with the TCP/IP core.*

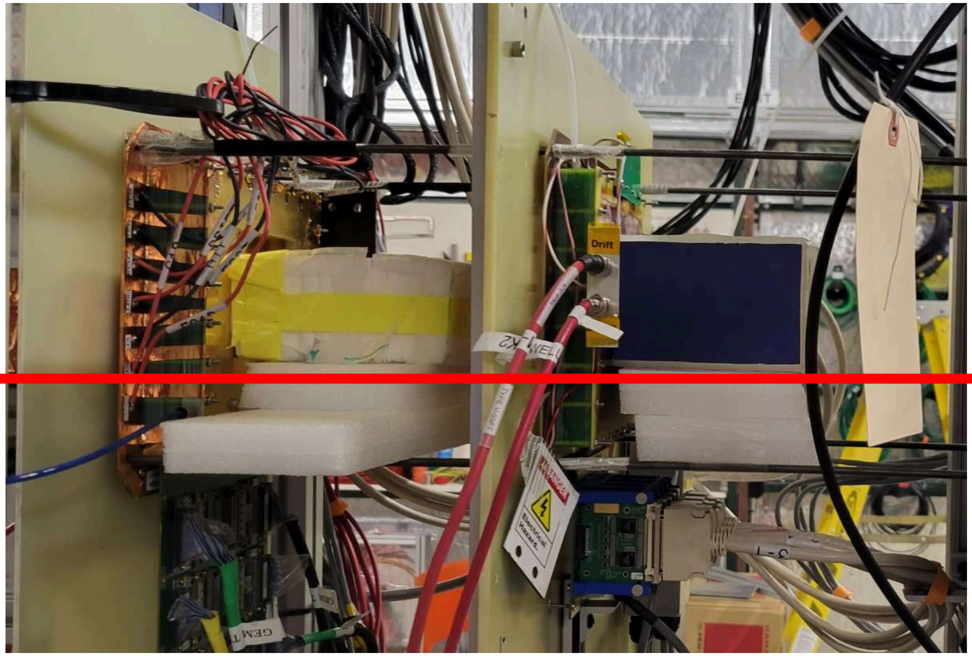
Beam test at FermiLab

Jefferson Lab
Thomas Jefferson National Accelerator Facility

Calorimeter



GEM-TRD   Micromegas TRD



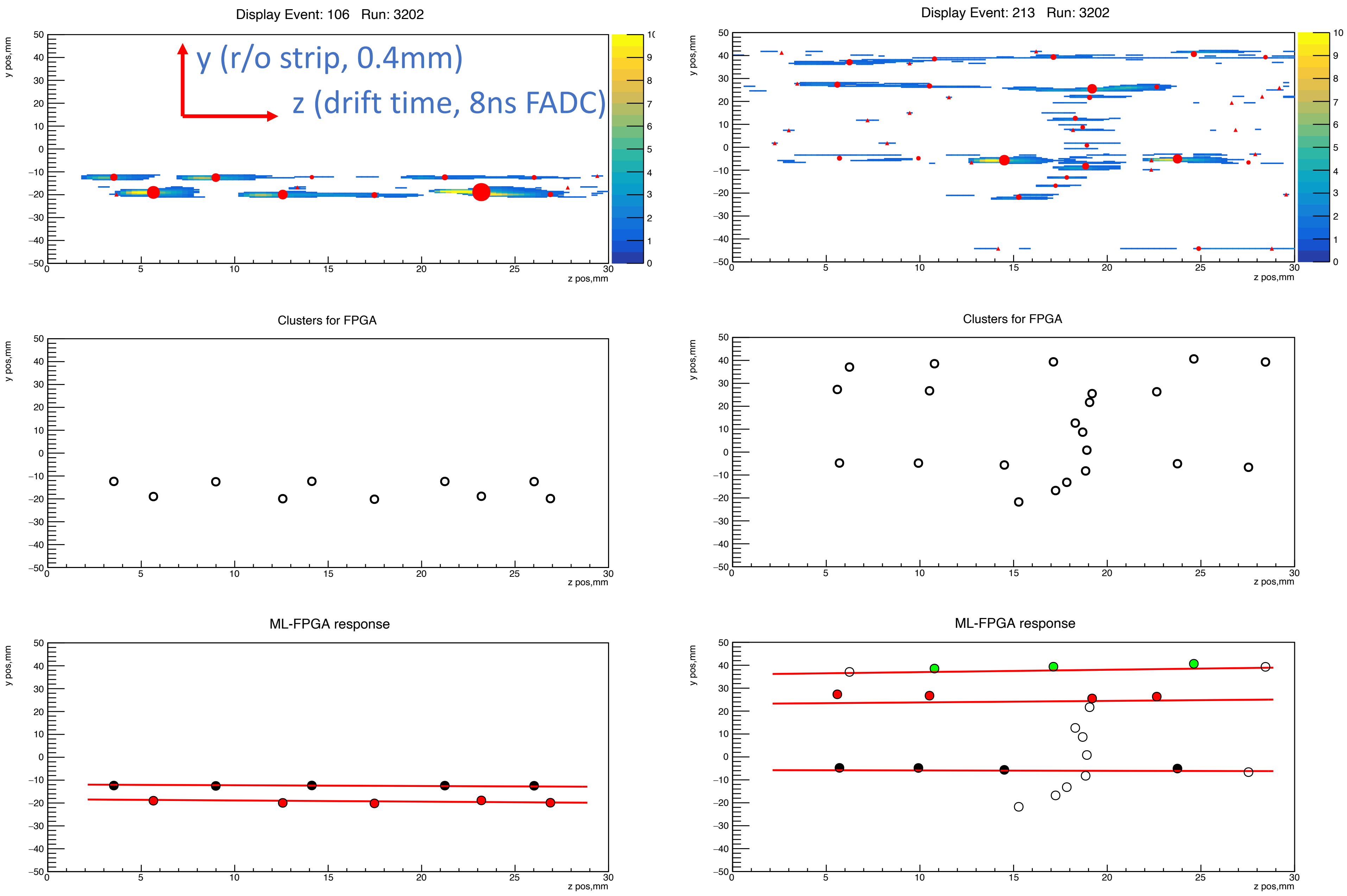
Beam

Lead Tungstate ($PbWO_4$) crystals

- *FermiLab test beam :*
  - Primary beam: protons 120 GeV
  - 4.2 seconds = length of spill
  - 60 seconds = approximate rep rate of spill
  - Beam intensity: Particles per spill : 10K – 1M (pps)

Tracking performance

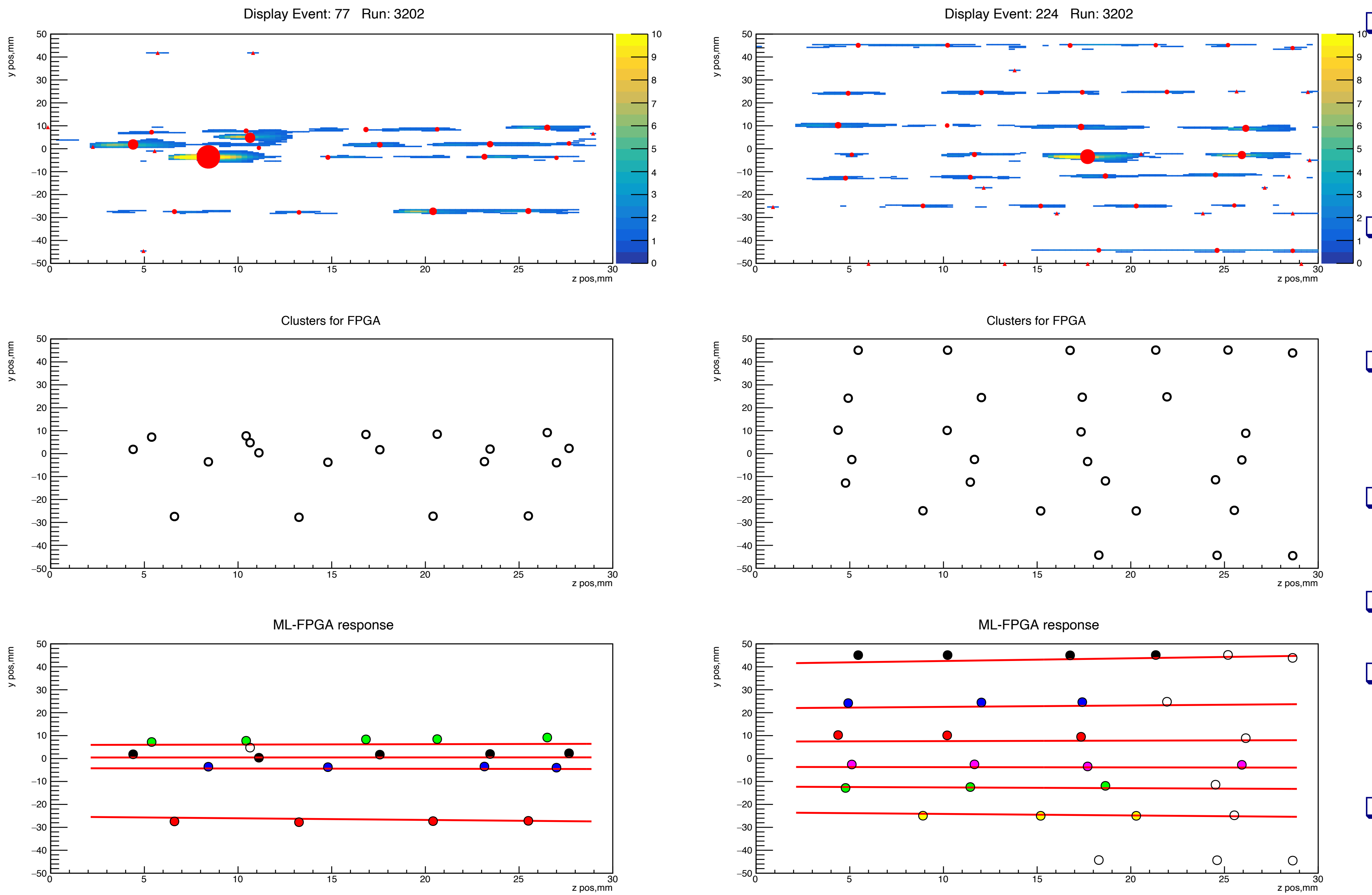
Jefferson Lab
Thomas Jefferson National Accelerator Facility



- *Top rows: show ionization along the track in GEMTRD detector.*
  - Red circles are reconstructed clusters using some dE/dx threshold. The size is proportional to energy.
- *Middle rows: after filtering out the noisy clusters, the coordinates of the clusters are sent to the FPGA/GNN for pattern recognition.*
- *Bottom rows: GNN provides labeling of clusters (by color in the figure), the same colors belong to the same track.*
- *Then clusters of the same color (tag) are sent to the track fitting module: LSTM.*
- *The results of track fitting are represented by lines in the figures.*
- *The next step is to count all the ionization in the corridor around the track and send it to the PID module (DNN).*
- *As a bonus, GEMTRD provides a track segment for the global tracking system.*

Tracking performance 2

Jefferson Lab
Thomas Jefferson National Accelerator Facility



- *Top rows: show ionization along the track in GEMTRD detector.*
  - Red circles are reconstructed clusters using some dE/dx threshold. The size is proportional to energy.
- *Middle rows: after filtering out the noisy clusters, the coordinates of the clusters are sent to the FPGA/GNN for pattern recognition.*
- *Bottom rows: GNN provides labeling of clusters (by color in the figure), the same colors belong to the same track.*
- *Then clusters of the same color (tag) are sent to the track fitting module: LSTM.*
- *The results of track fitting are represented by lines in the figures.*
- *The next step is to count all the ionization in the corridor around the track and send it to the PID module (DNN).*
- *As a bonus, GEMTRD provides a track segment for the global tracking system.*

Latency and rates (very preliminary)

Jefferson Lab
Thomas Jefferson National Accelerator Facility

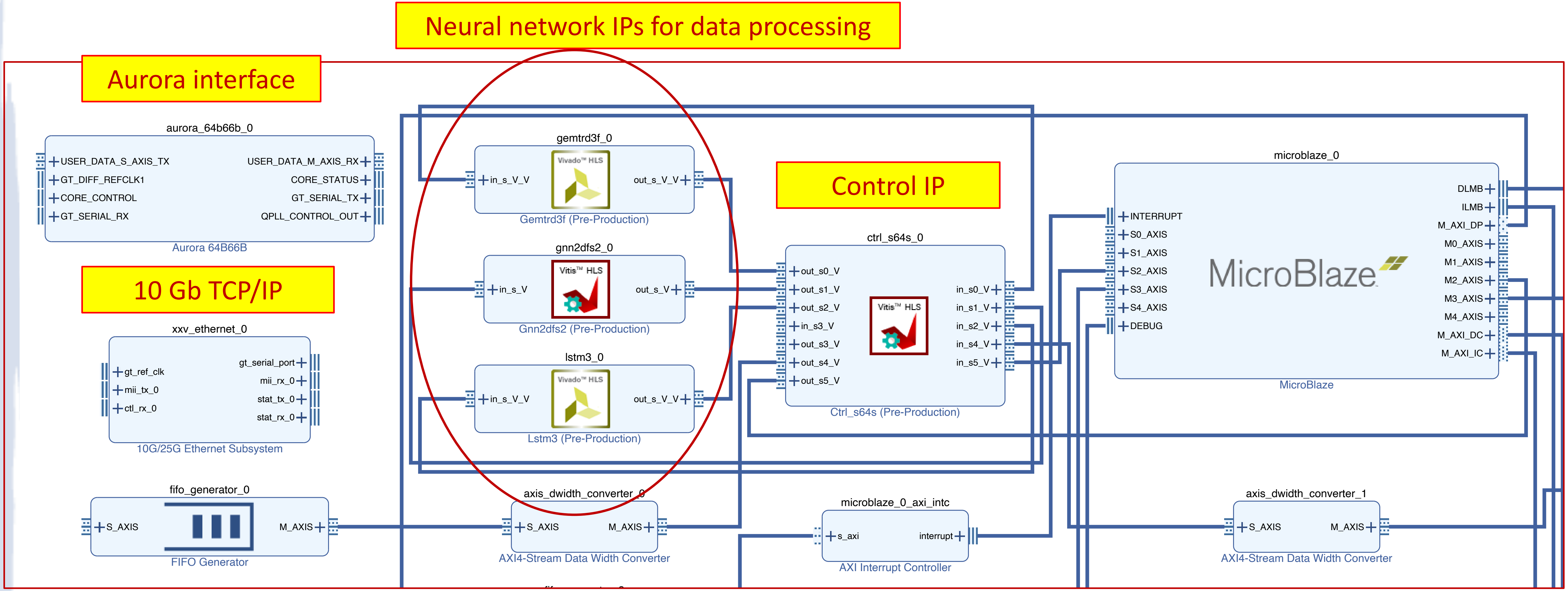
- *Although the system worked in principle, overall performance was quite poor:*
  - the board could process data at a speed of about ten hertz.
- *The latency was determined by MicroBlaze's participation in the data exchange.*
- *So the next step was to synthesize the Control IP with the functionality of a C program running on MicroBlaze.*
- *The IP block was synthesized directly using Vitis_HLS and the overall latency was reduced to 20 μs. (~50kHz).*
- *Control IP block primarily performs serial I/O.*
- *Therefore, it consists of long loops designed to accommodate the maximum data size.*
- *In reality, the average data size is much smaller, so the actual speed should be higher.*
- *This was confirmed in measurements - peak performance reached 80 kHz.*
- *This is the first version, not yet optimized and II violations have not been fixed.*

Modules & Loops	Issue Type	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
ctrl_s64s	II Violation	-	4178	2.089E4	-	4179	-	no	8	5	4184	22984	0
VITIS_LOOP_399_2		-	4	20.000	1	1	4	yes	-	-	-	-	-
VITIS_LOOP_443_3		-	1024	5.120E3	1	1	1024	yes	-	-	-	-	-
VITIS_LOOP_464_4		-	1025	5.125E3	3	1	1024	yes	-	-	-	-	-
VITIS_LOOP_475_5	II Violation	-	45	225.000	6	2	21	yes	-	-	-	-	-
VITIS_LOOP_479_7	II Violation	-	43	215.000	4	2	21	yes	-	-	-	-	-
VITIS_LOOP_484_9_VITIS_LOOP_484_10		-	45	225.000	5	1	42	yes	-	-	-	-	-
VITIS_LOOP_503_11		-	7	35.000	5	1	4	yes	-	-	-	-	-
VITIS_LOOP_508_12		-	21	105.000	1	1	21	yes	-	-	-	-	-
VITIS_LOOP_523_13		-	27	135.000	3	1	26	yes	-	-	-	-	-
VITIS_LOOP_540_14		-	21	105.000	1	1	21	yes	-	-	-	-	-
VITIS_LOOP_542_15		-	22	110.000	3	1	21	yes	-	-	-	-	-
VITIS_LOOP_562_16	II Violation	-	804	4.020E3	45	40	20	yes	-	-	-	-	-
VITIS_LOOP_626_20		-	44	220.000	3	2	21	yes	-	-	-	-	-
VITIS_LOOP_642_21		-	1025	5.125E3	3	1	1024	yes	-	-	-	-	-

New board design with Control IP

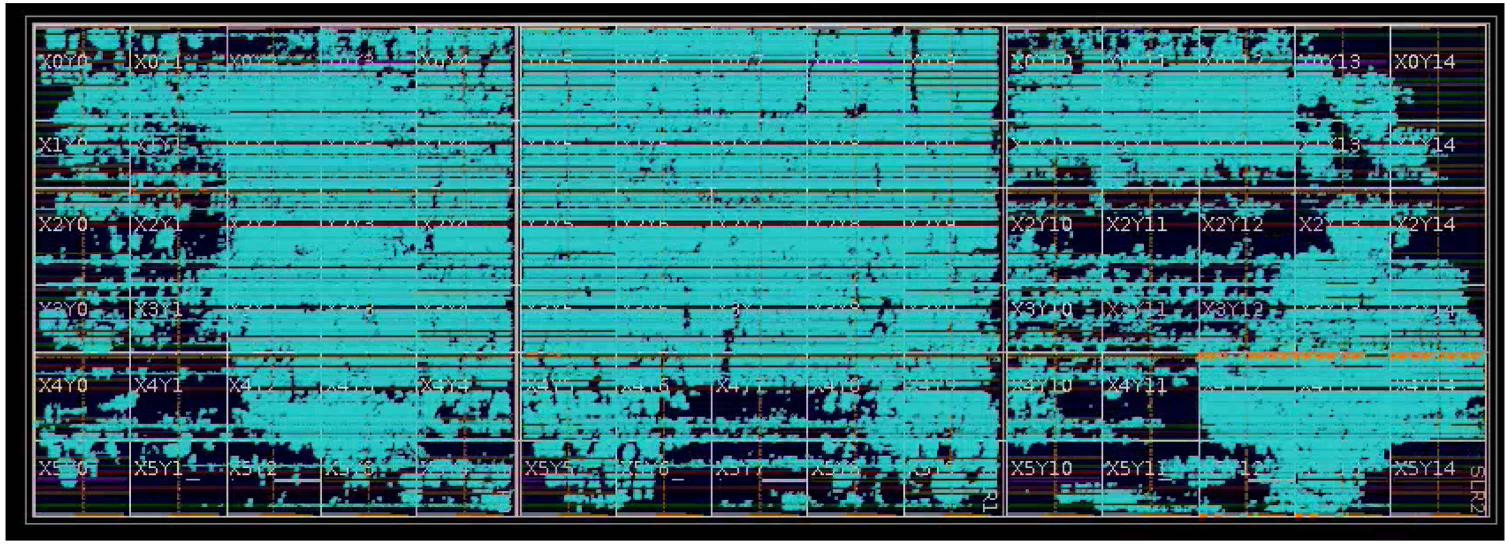
Jefferson Lab
Thomas Jefferson National Accelerator Facility

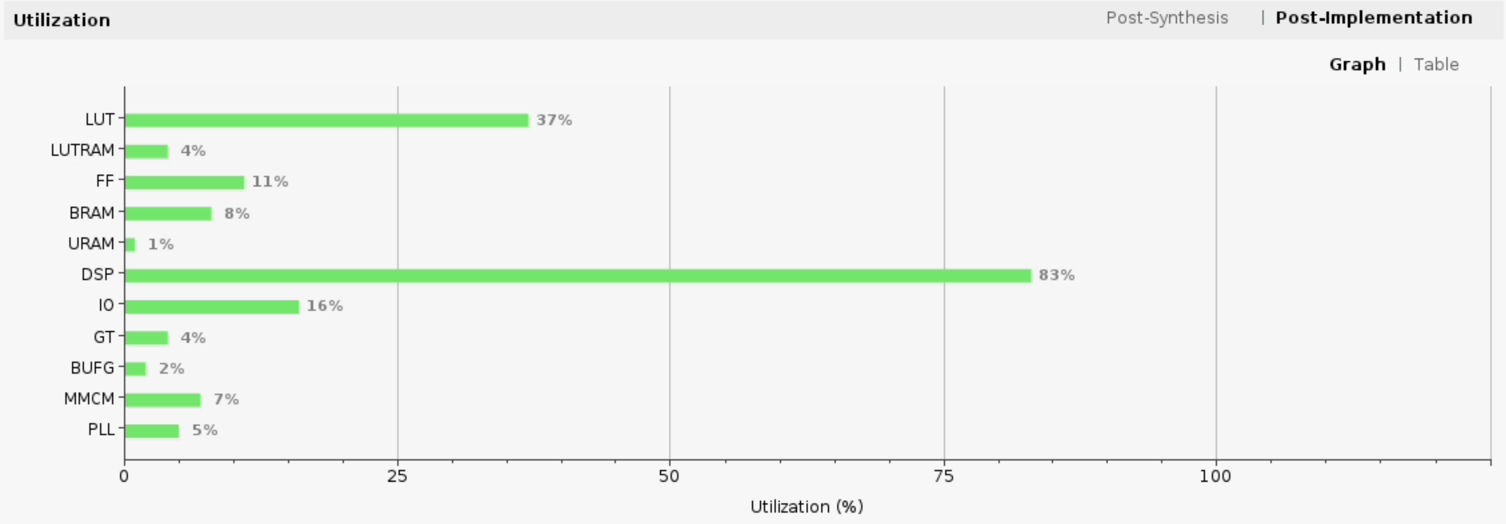
- *All data I/O operations are performed by Control IP*
- *Microblaze is only used to configure the board and monitor data processing.*
- *Aurora interface provides communication with a second FPGA board that processes the calorimeter data (CNN).*
- *10 Gigabit Ethernet uses TCP/IP, receives data from detectors (DAQ) and sends pre-processed data to the computer (farm).*



FPGA board resources for GEMTRD

- Neural networks use a lot of FPGA resources.
- Therefore, one VCU118 board can only process data from GEMTRD.
  - See pictures on the right
- The calorimeter uses CNN to process its data and currently occupies the entire VCU118 board.
- Calorimeter FPGA board has its own 10 Gb ethernet and Aurora interfaces.





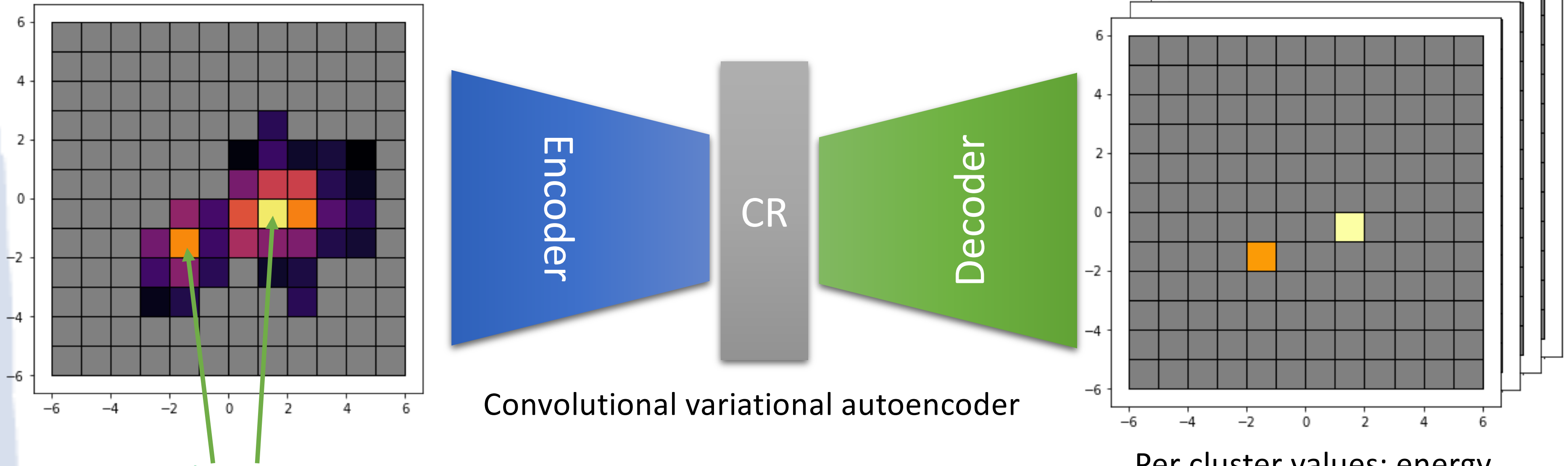
ML for Calorimeter

Calorimeter parameters reconstruction

Jefferson Lab
Thomas Jefferson National Accelerator Facility

By Dmitry Romanov

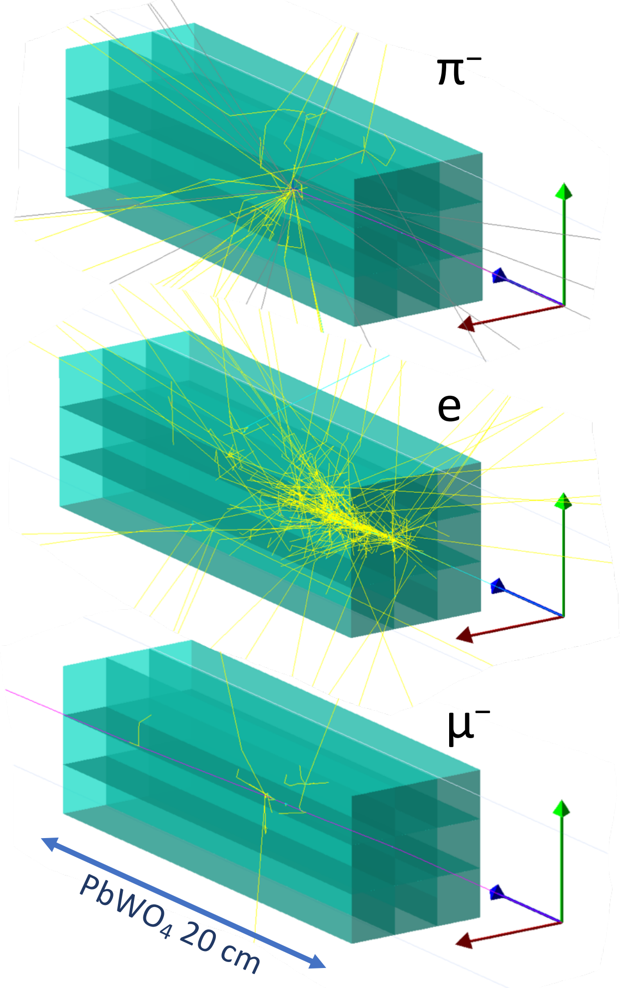
Modules deposits



Convolutional variational autoencoder

Per cluster values: energy, pid, x, y, features

Geant 4 simulation



Examples of events with e and π^- showers and μ^- passing through.

- Convolutional VAE as a backbone
- Modules deposits as inputs
- Per cluster output of multiple values:
- Energy, e/ π, coordinates, features

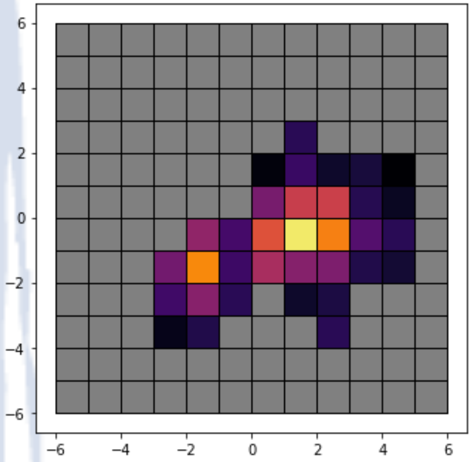
CNN for calorimeter reconstruction

- *In this work we used a convolutional encoder with a decoder consisting of dense layers, which provide e-π separation scores as the output.*
- *Synthesized with HLS4ML*
- *This was done to minimize a network size in FPGA and due to current limitation of HSL4ML of supported network layer types.*
- *FPGA synthesis with reuse factor of 2 has a latency of 0.7µs and an interval of 125 clocks. It uses 74% of DPS resources*

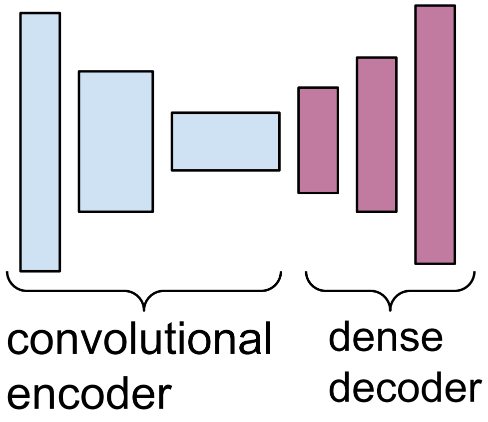
```
+--------+--------+----------+------------+
|  Clock | Target| Estimated| Uncertainty|
+--------+--------+----------+------------+
|ap_clk  |    5.00|     4.303|        0.62|
+--------+--------+----------+------------+

Latency (clock cycles):
  * Summary:
  +-----+-----+-----+-----+----------+
  |  Latency  |  Interval | Pipeline |
  | min | max | min | max |   Type   |
  +-----+-----+-----+-----+----------+
  |  139|  139|  125|  125| dataflow |
  +-----+-----+-----+-----+----------+
```

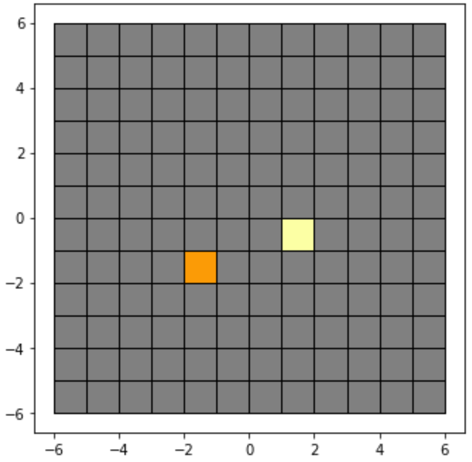
Actual values	Predicted results	
	e	π
e	98.8 %	1.2 %
π	2.9 %	97.1 %

```
+-----------------------+---------+--------+---------+---------+-----+
|         Name          | BRAM_18K| DSP48E|    FF    |   LUT    | URAM|
+-----------------------+---------+--------+---------+---------+-----+
|DSP                    |        -|       -|        -|        -|    -|
|Expression             |        -|       -|        0|        2|    -|
|FIFO                   |      404|       -|     8999|    15698|    -|
|Instance               |       61|    5124|    55854|   243846|    -|
|Memory                 |        -|       -|        -|        -|    -|
|Multiplexer            |        -|       -|        -|        -|    -|
|Register               |        -|       -|        -|        -|    -|
+-----------------------+---------+--------+---------+---------+-----+
|Total                  |      465|    5124|    64853|   259546|    0|
+-----------------------+---------+--------+---------+---------+-----+
|Available SLR          |     1440|    2280|   788160|   394080|  320|
+-----------------------+---------+--------+---------+---------+-----+
|Utilization SLR (%)    |       32|     224|        8|       65|    0|
+-----------------------+---------+--------+---------+---------+-----+
|Available              |     4320|    6840|  2364480|  1182240|  960|
+-----------------------+---------+--------+---------+---------+-----+
|Utilization (%)        |       10|      74|        2|       21|    0|
+-----------------------+---------+--------+---------+---------+-----+
```



Input data



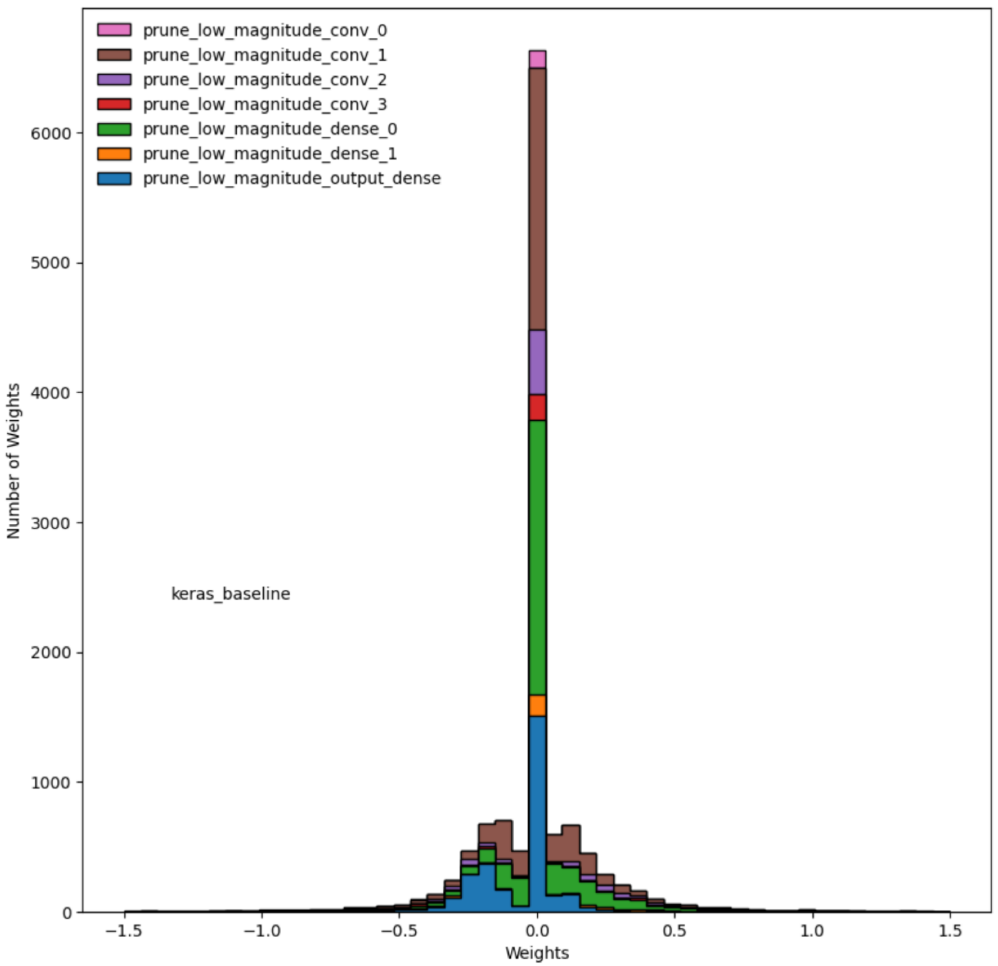


output

Calorimeter CNN optimization with HLS4ML

```
hls_config['Model']['Precision'] = 'ap_fixed<20,10>'
```

```
Layer prune_low_magnitude_conv_0: % of zeros = 0.5
Layer prune_low_magnitude_conv_1: % of zeros = 0.5
Layer prune_low_magnitude_conv_2: % of zeros = 0.5
Layer prune_low_magnitude_conv_3: % of zeros = 0.5
Layer prune_low_magnitude_dense_0: % of zeros = 0.5
Layer prune_low_magnitude_dense_1: % of zeros = 0.5
Layer prune_low_magnitude_output_dense: % of zeros = 0.5
Layer prune_low_magnitude_fused_convbn_0: % of zeros = 0.0
Layer prune_low_magnitude_fused_convbn_1: % of zeros = 0.0
Layer prune_low_magnitude_fused_convbn_2: % of zeros = 0.0
Layer prune_low_magnitude_fused_convbn_3: % of zeros = 0.0
Layer prune_low_magnitude_dense_0: % of zeros = 0.0
Layer prune_low_magnitude_dense_1: % of zeros = 0.0
Layer output_dense: % of zeros = 0.0
```

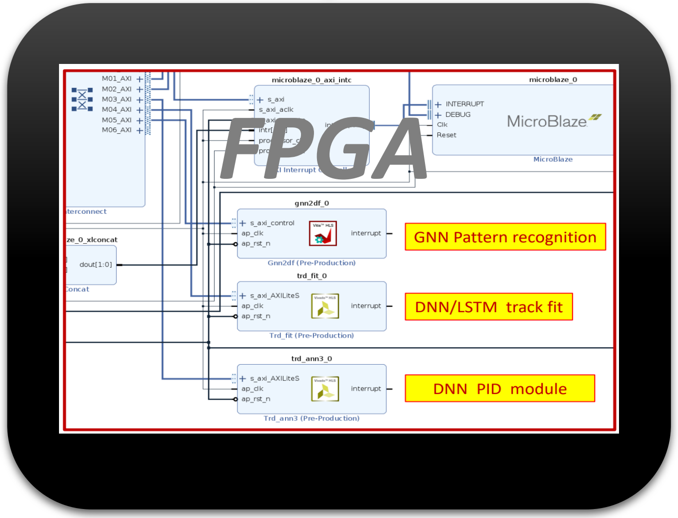


JANA2 for ML on FPGA

Pre-processed data from the FPGA is transferred over the network (TCP/IP) to a computer running JANA2 software.

JANA4ML4FPGA

Detector



Validation software

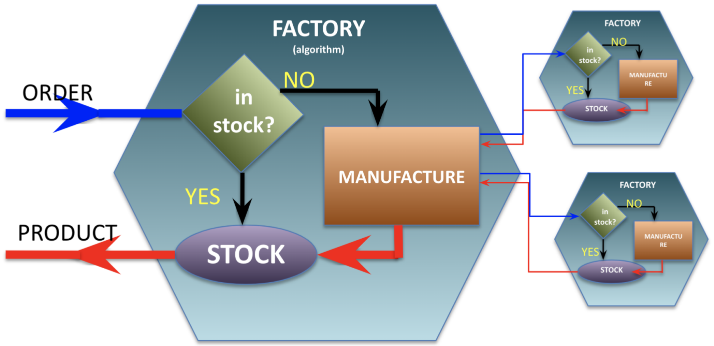
JANA2

(**J**Lab **ANA**lysis framework)

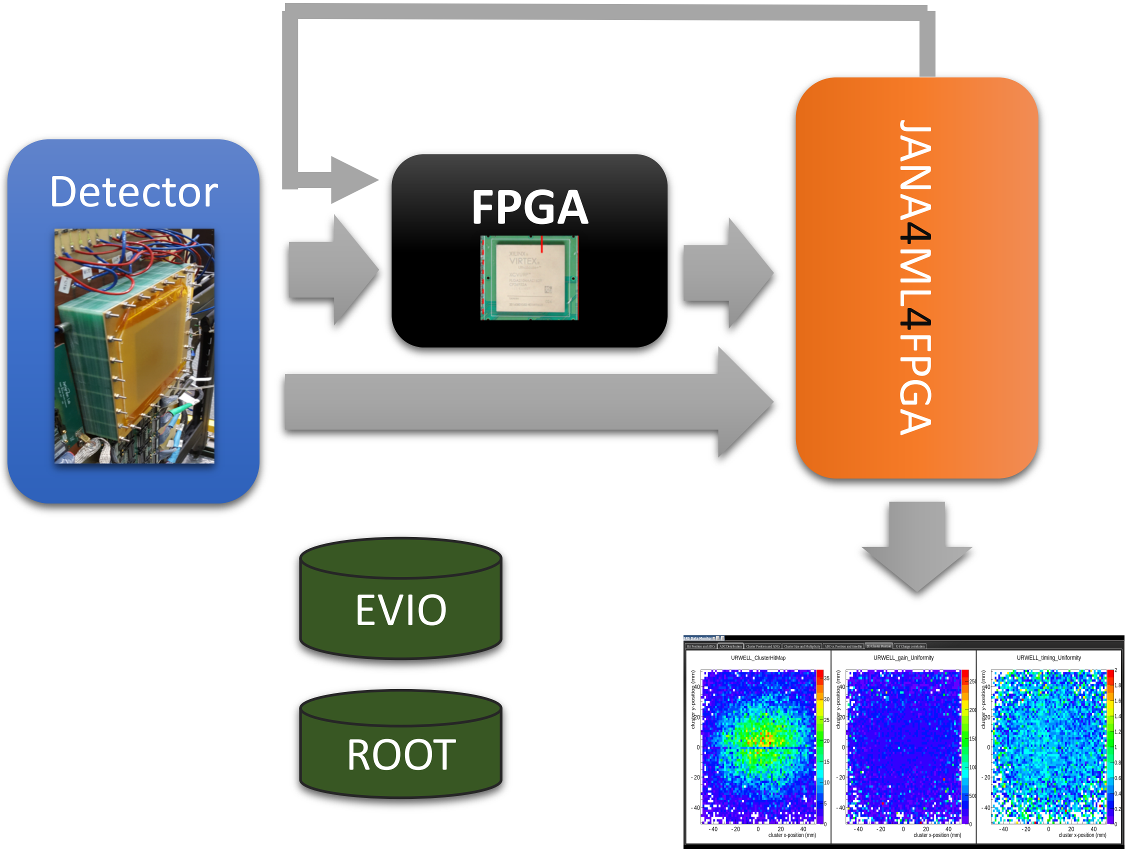
- JANA2 is a multi-threaded modular event reconstruction framework being developed at Jlab for online and offline processing

- JANA2 is a rewrite based on modern coding and CS practices. Developed for modern NP experiments with streaming readout, heterogeneous computing and AI

- JANA2 is the main framework chosen for EIC. Used for ePIC collaboration reconstruction and further Detector 2. Used in multiple Jlab experiments and prototypes



JANA4ML4FPGA



Goals:

- Read and write EVIO
- Write flat ROOT files
- Receive EVIO by TCP (and save)
- Receive network streams
- Receive FPGA data
- Simulate sending detector data
- Data Quality Monitor
- AI streaming preprocessing
- Conventional preprocessing

Outlook

Jefferson Lab
Thomas Jefferson National Accelerator Facility

- *An FPGA-based Neural Network application would offer online event preprocessing and allow for data reduction based on physics at the early stage of data processing.*
- *The ML-on-FPGA solution complements the purely computer-based solution and mitigates DAQ performance risks.*
- *FPGA provides extremely low-latency neural-network inference.*
- *Open-source HLS4ML software tool with Xilinx® Vivado® High Level Synthesis (HLS) accelerates machine learning neural network algorithm development.*

- *The ultimate goal is to build a real-time event filter based on physics signatures.*

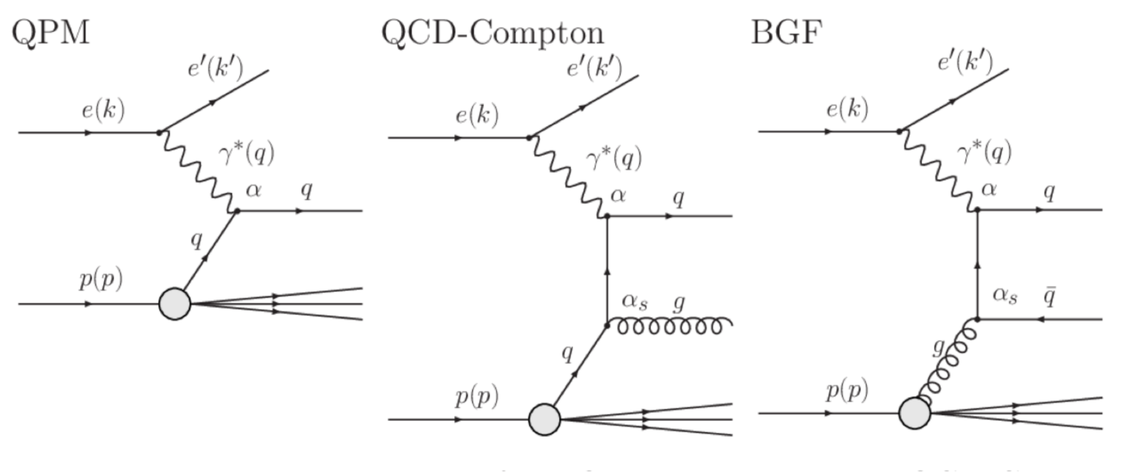


Figure 2.1: Feynman diagrams of the Quark Parton Model, QCD-Compton and Boson Gluon Fusion processes in NC DIS.

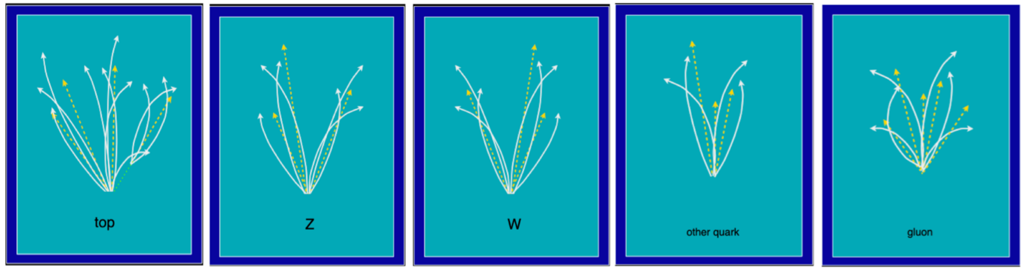
Published in 2007

Measurement of multijet events at low x_{Bj} and low Q^2 with the ZEUS detector at HERA

T. Gosau

Case study: jet tagging

Study a multi-classification task: discrimination between highly energetic (boosted) ***q, g, W, Z, t*** initiated jets



t→bW→bqq	**Z→qq**	**W→qq**	**q/g background**
3-prong jet	2-prong jet	2-prong jet	no substructure and/or mass ~ 0

Signal: reconstructed as one massive jet with substructure

Jet substructure observables used to distinguish signal vs background [*]

[*] D. Guest at al. PhysRevD.94.112002, G. Kasieczka et al. JHEP05(2017)006, J. M. Butterworth et al. PhysRevLett.100.242001, etc..

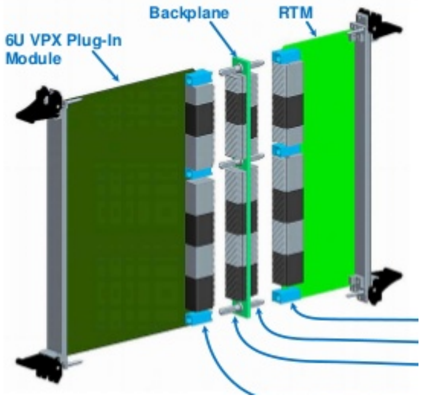
11.01.2019 Jennifer Ngadiuba - hls4ml: deep neural networks in FPGAs 25

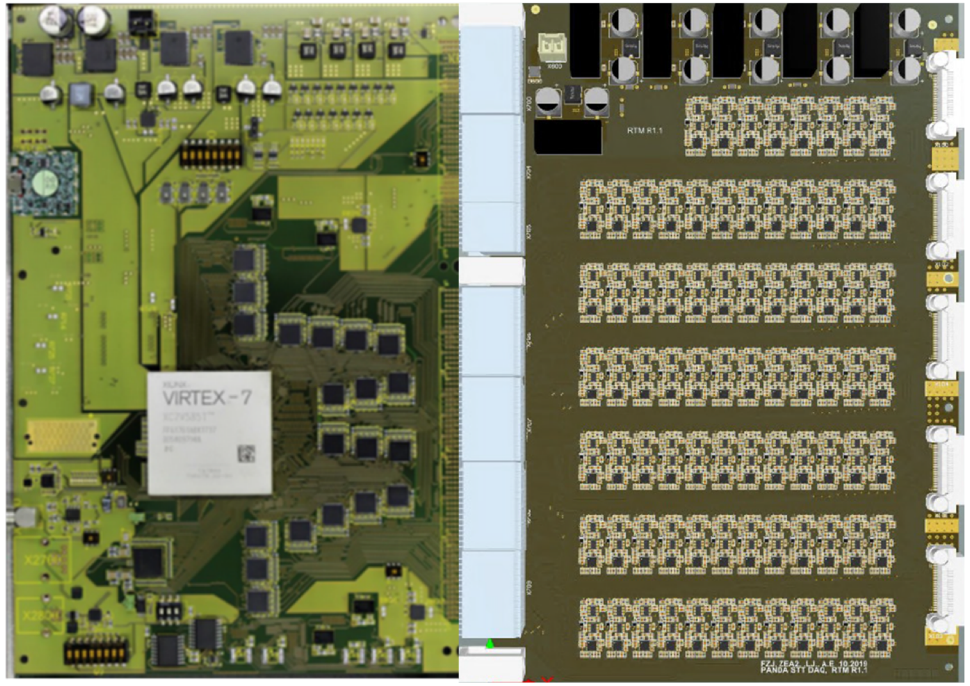
Backup

ADC based DAQ for PANDA STT

Jefferson Lab
Thomas Jefferson National Accelerator Facility

Level 0 Open VPX Crate
ADC based DAQ for PANDA STT (one of approaches):
- 160 channels (shaping, sampling and processing) per payload slot, 14 payload slots+2 controllers;
- **totally 2200 channels per crate**;
- time sorted output data stream (arrival time, energy,...)
- noise rejection, pile up resolution, base line correction, ..





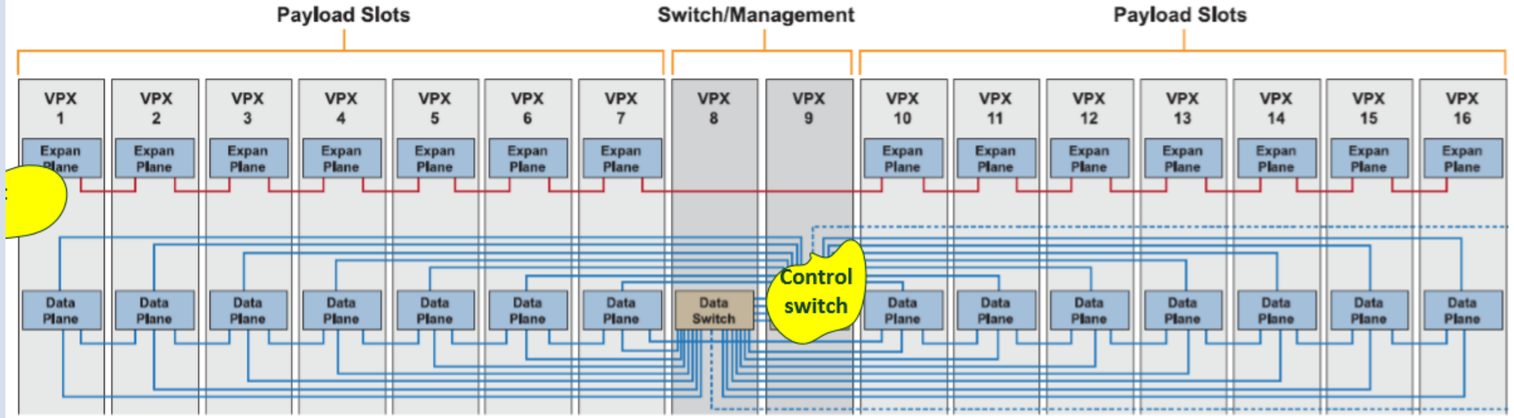
- 40 4-channel ADCs (configurable up to 1 GSPS);
- Single Virtex7 FPGA

- 160 Amplifiers;
- 5 connectors for 32-pins samtec cables

✦ *All information from the straw tube tracker is processed in one unit.*

✦ *Allows to build a complete STT event.*

✦ *This unit can also be used for calorimeters readout and processing.*



https://doi.org/10.1088/1748-0221/17/04/C04022
2022_JINST_17_C04022

Powerful Backplane up to 670 GBs

L. Jokhovets, P Kulessa ..
JÜLICH Forschungszentrum

A Brief Intro to Artificial Neural Network on FPGA

Jefferson Lab
Thomas Jefferson National Accelerator Facility

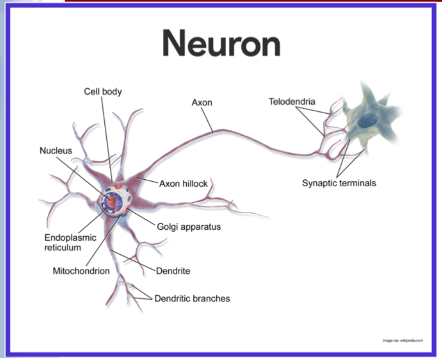
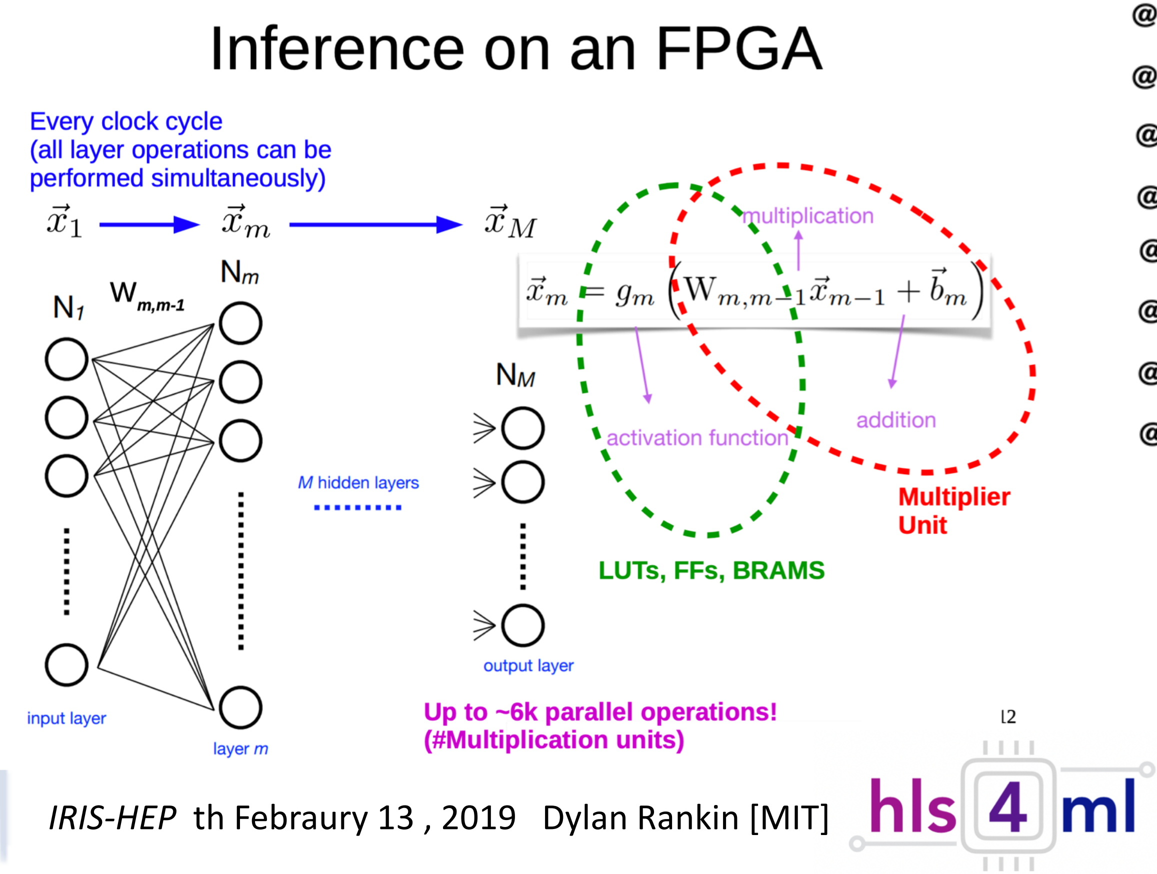
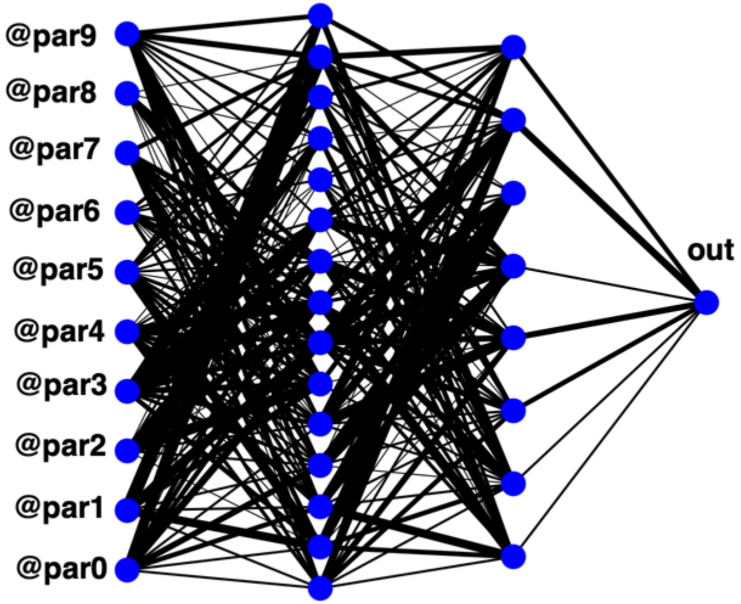
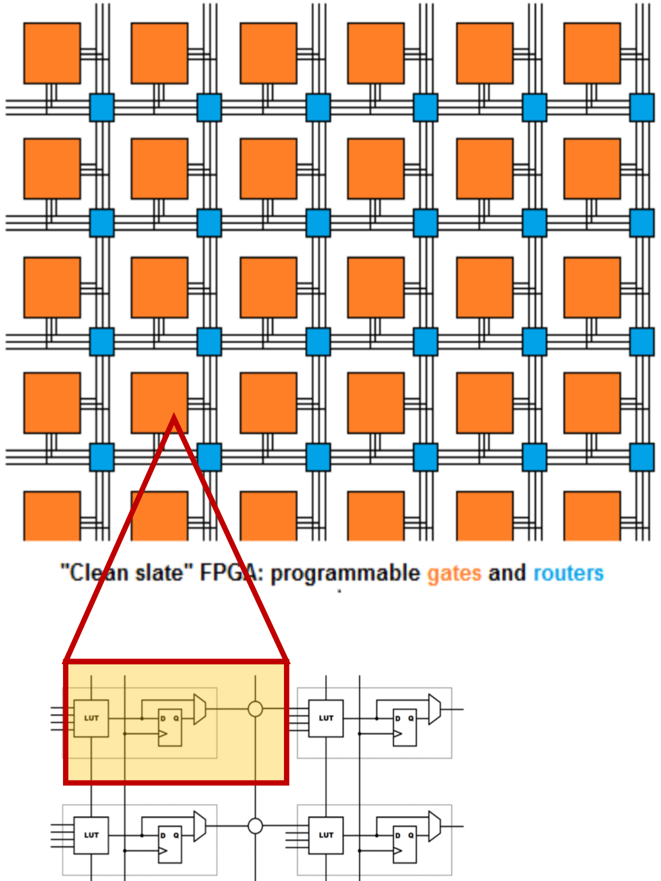


Image: https://nurseslabs.com/nervous-system/

- FPGA Field Programmable Gate Array .
- It can perform logical operation in parallel







- Modern FPGAs have DSP slices - specialized hardware blocks placed between gateways and routers that perform mathematical calculations.
- The number of DSP slices can be up to 6000-12000 per chip.

Image from: https://www.embeddedrelated.com/showarticle/195.php

Optimization with hls4ml package

Jefferson Lab
Thomas Jefferson National Accelerator Facility

- A package hls4ml is developed based on High-Level Synthesis (HLS) to build machine learning models in FPGAs.

article: J. Duarte *et al* 2018 *JINST* **13** P07027