Fast ML on FPGA for Particle Identification and Tracking

Sergey Furletov
(Jefferson Lab)

F. Barbosa, L. Belfore, N. Brei, C. Dickover, C. Fanelli,
D. Furletov, D. Lawrence, C. Mei, D. Romanov, K. Shivu

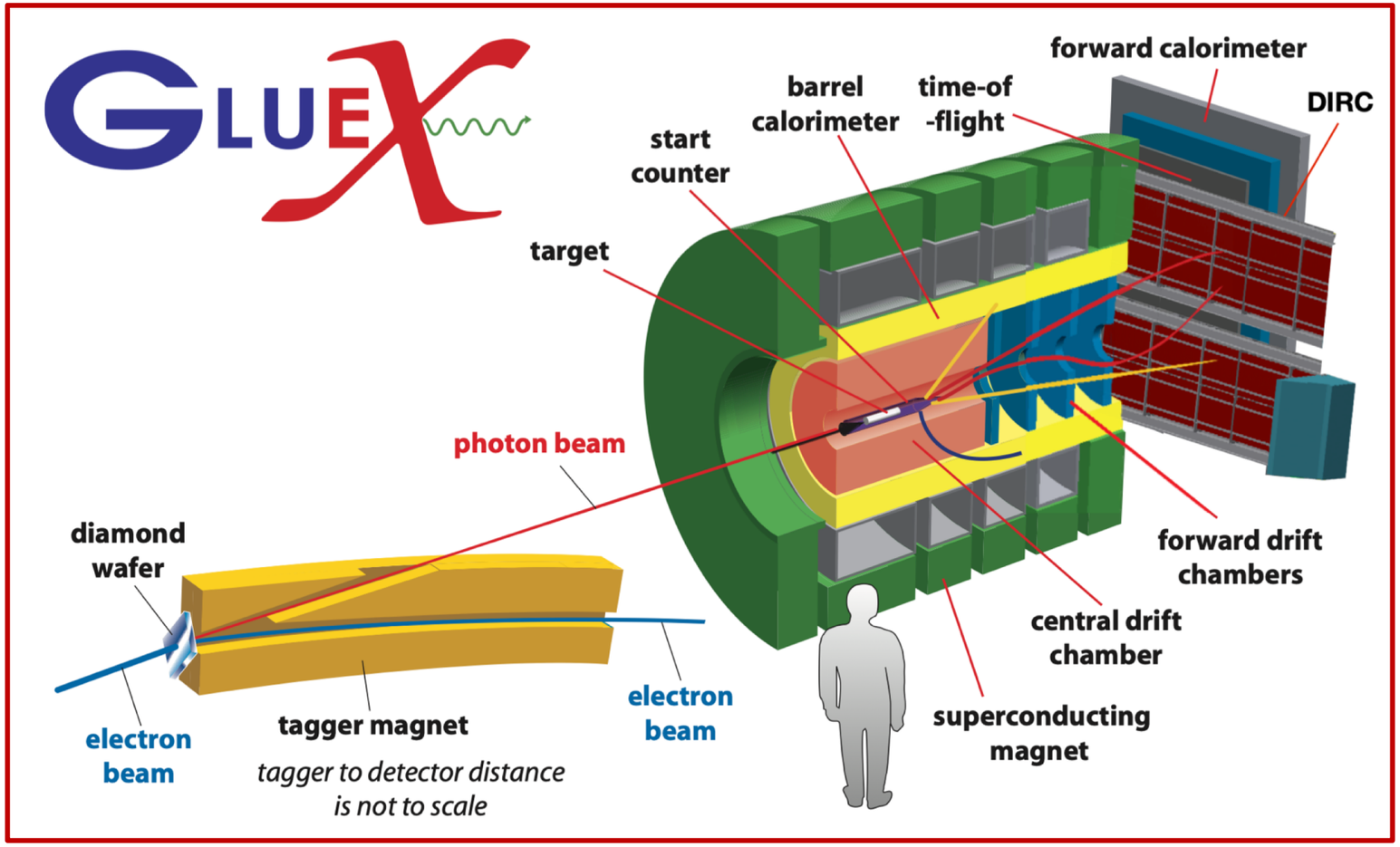
25th IEEE Real Time Conference

May 25 – 29, 2026 La Biodola - Isola d'Elba (Italy)

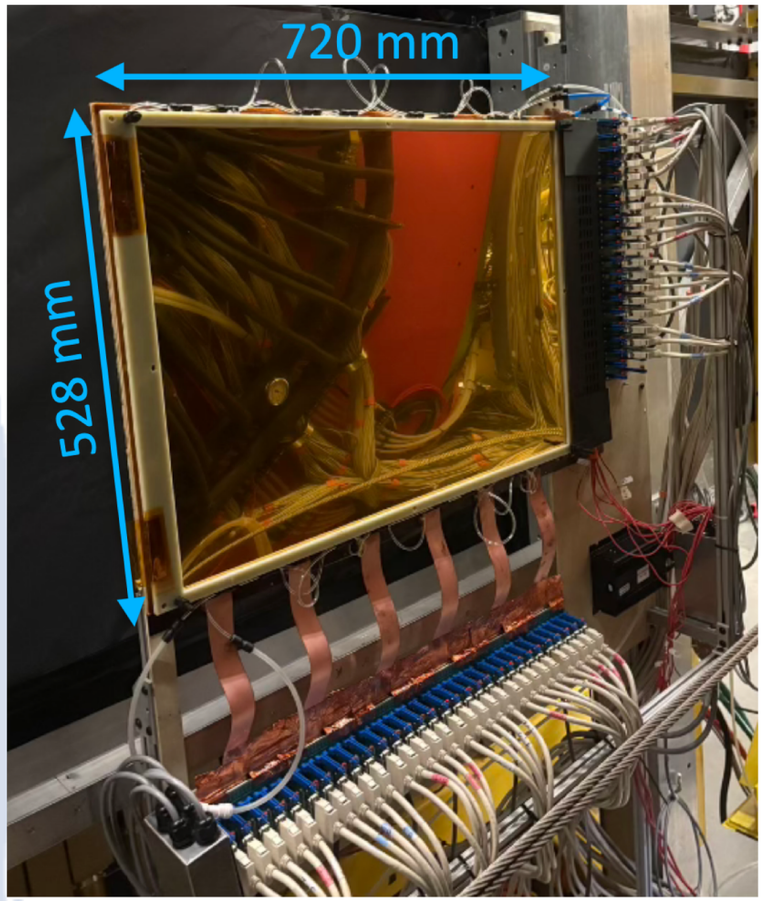
GlueX experiment at Jefferson Lab

- *GlueX is a particle physics experiment located at the Thomas Jefferson National Accelerator Facility (JLab) accelerator in Newport News, Virginia.*
- *The planned installation of a Transition Radiation Detector (TRD) is a central upgrade for the GlueX-III experiment. This system improves particle identification (PID) by efficiently separating electrons from heavier hadrons like pions in high-background environments.*
- *This upgrade also includes the development of new readout electronics capable of performing data processing using machine learning in real time.*

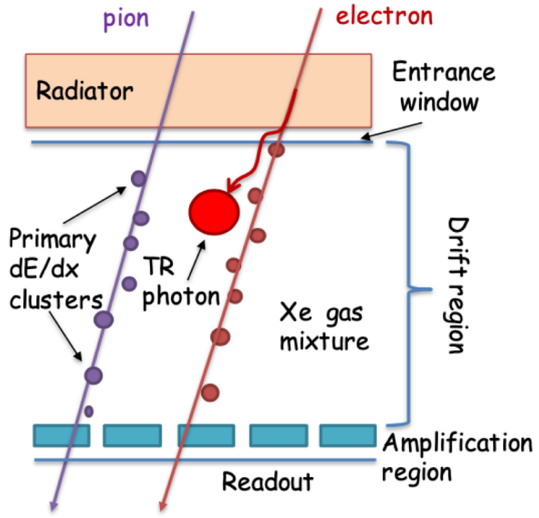
- *Typical L1 trigger rate 40-70 kHz*
- *Data rate 0.7 – 1.2 GB/s*
- *L1 Trigger latency 3.5 us μs.*
  - (increase up to 8 μs is possible.)

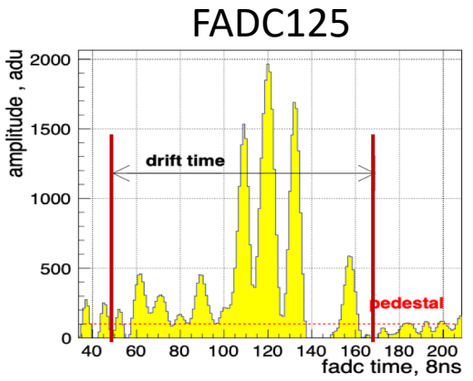


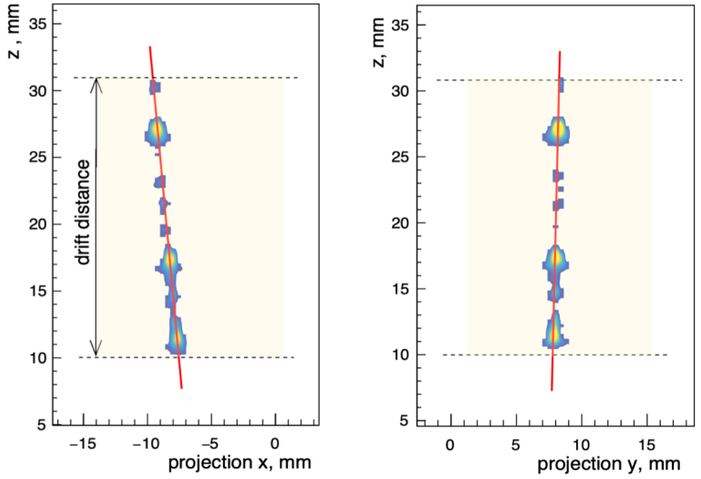
GEM-TRD principle



Fleece Radiator





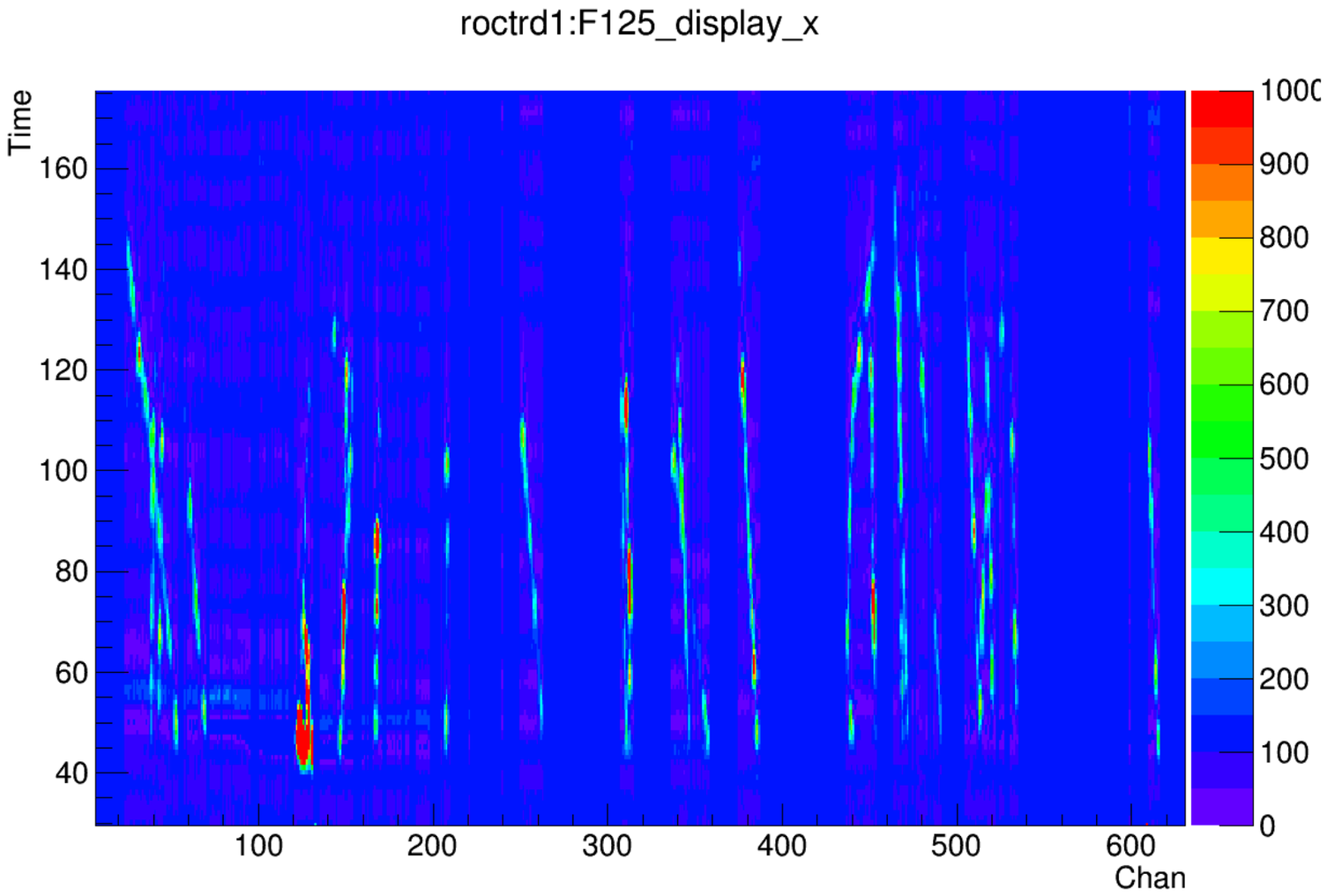


- *The e/pion separation in the GEM-TRD detector is based on counting the ionization along the particle track.*
- *For electrons, the ionization is higher due to the absorption of transition radiation photons*

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

GEM-TRD event display

Projection X

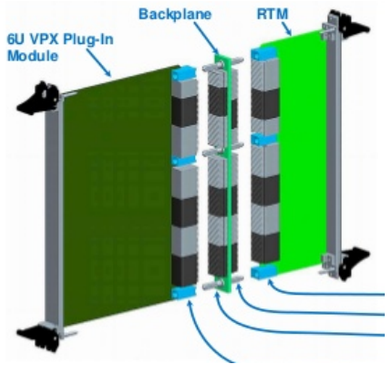


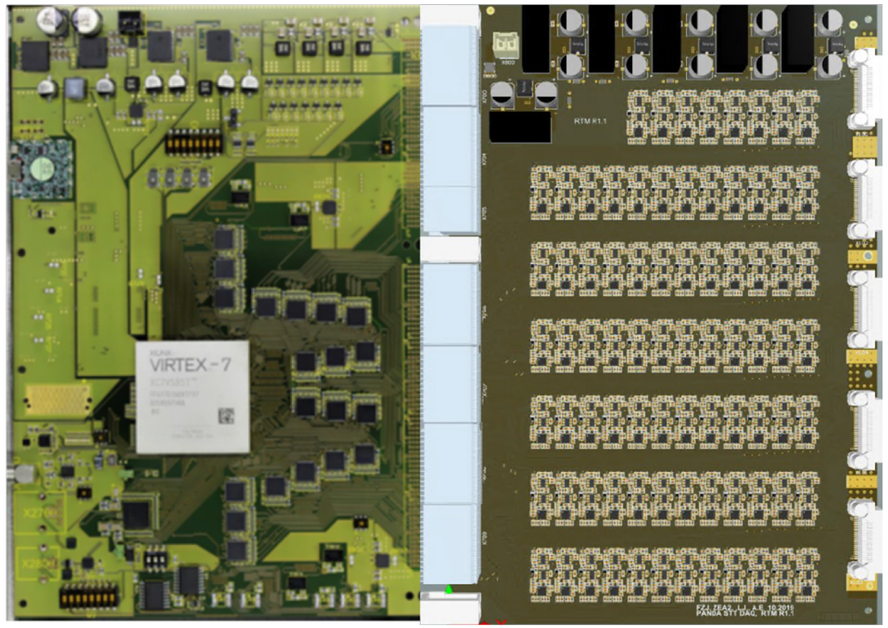
Readout electronics : OpenVPX FADC

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, ..

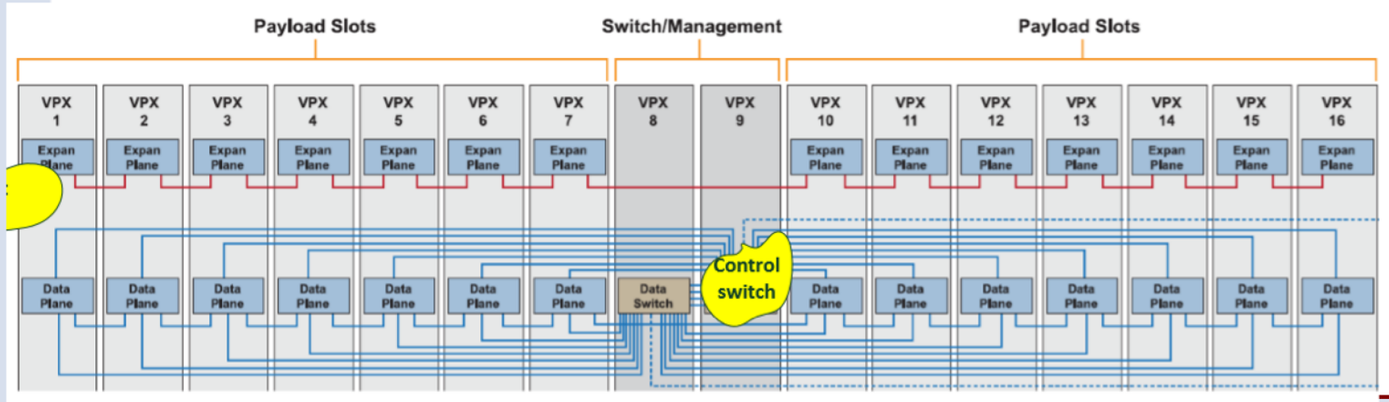




- 40 4-channel ADCs (configurable up to 1 GSPS);
- Single Virtex7 FPGA

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

- *FADC:166 MHz ,12 bit*
- *2200 chan. per crate*
- *The Virtex-7 FPGA enables real-time signal processing and feature extraction.*



Powerful Backplane up to 670 GBs

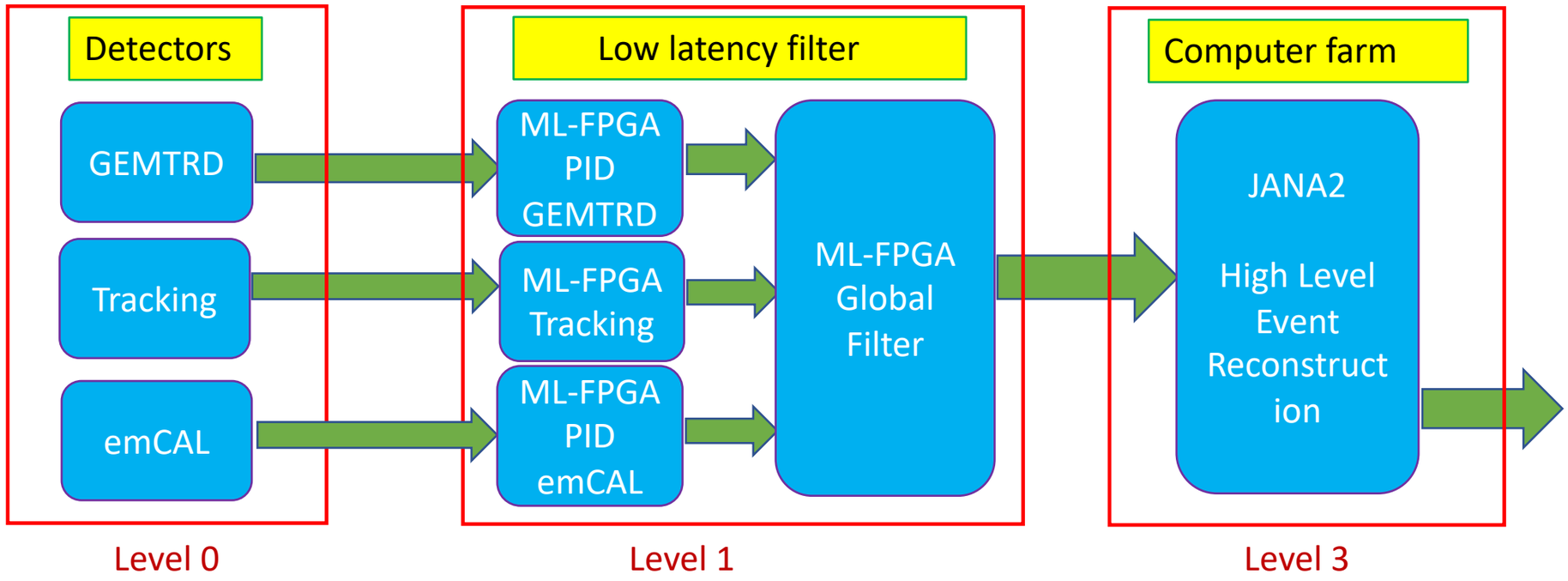
https://doi.org/10.1088/1748-0221/17/04/C04022
2022_JINST_17_C04022

L. Jokhovets, P Kulessa ..

JÜLICH Forschungszentrum

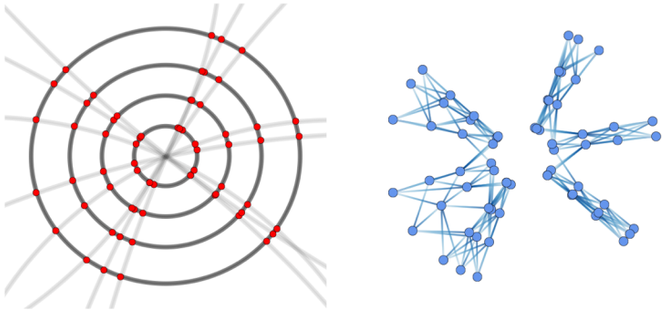
Data processing ML-(on)-FPGA

- *The work was initially funded under the EIC Generic R&D program.*
- *The goal was to build a demonstrator that can operate under beam test conditions in real-time.*
- *The setup consists of several PID and tracking detectors: emCAL, GEMTRD, GEM tracker.*
- *Preprocessed data from detectors including decision on the particle type is 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.*



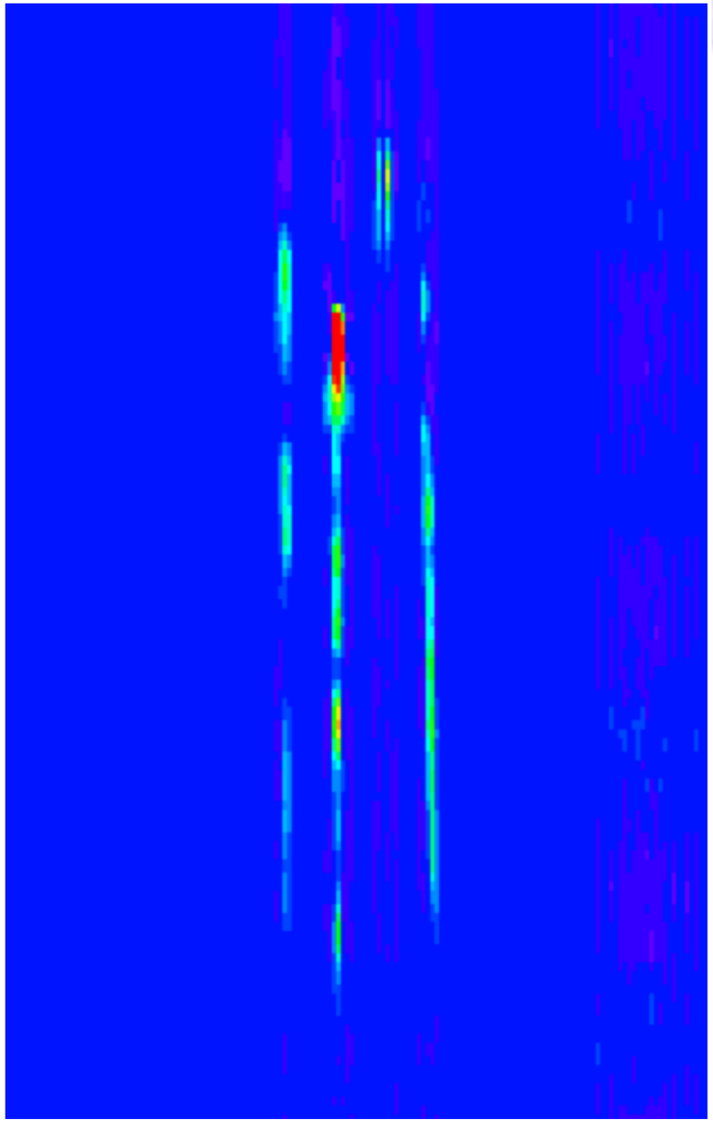
GEMTRD tracks

- *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.*



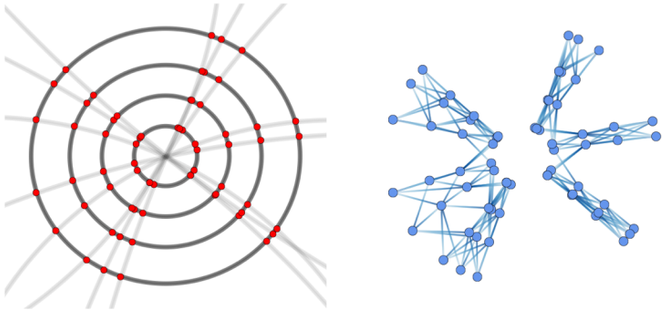
Javier Duarte
arXiv:2012.01249v2 [hep-ph] 7 Dec 2020

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



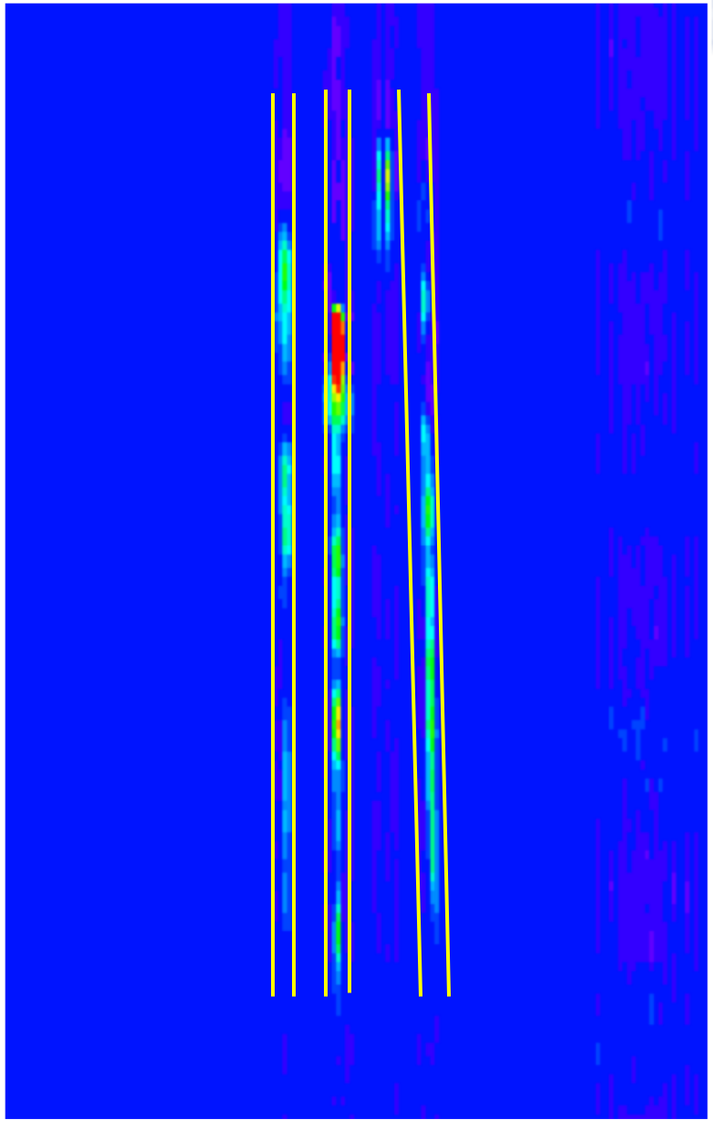
GEMTRD tracks

- *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.*



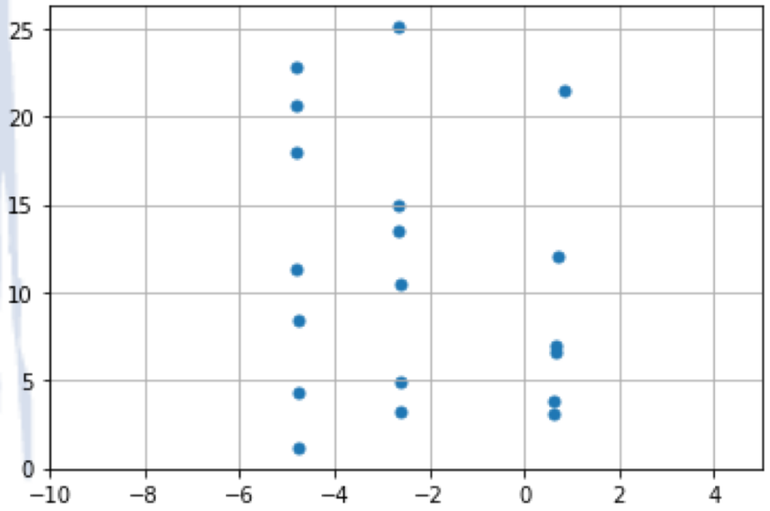
Javier Duarte
arXiv:2012.01249v2 [hep-ph] 7 Dec 2020

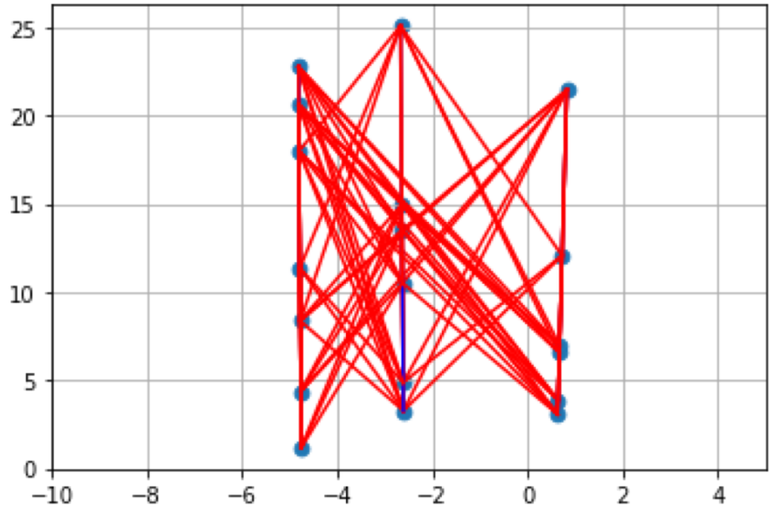
- HEP advanced tracking algorithms at the exascale **(Project Exa.TrkX)**
- https://exatrkx.github.io/

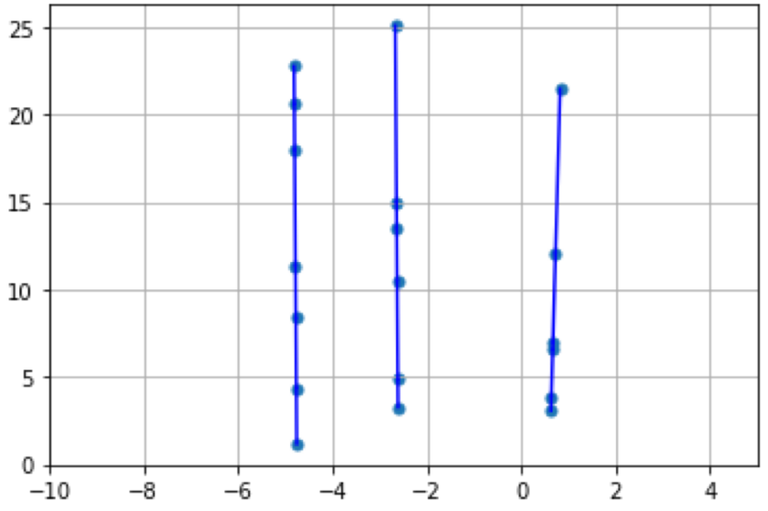


GNN for pattern recognition

- *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.*
- *So, 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*

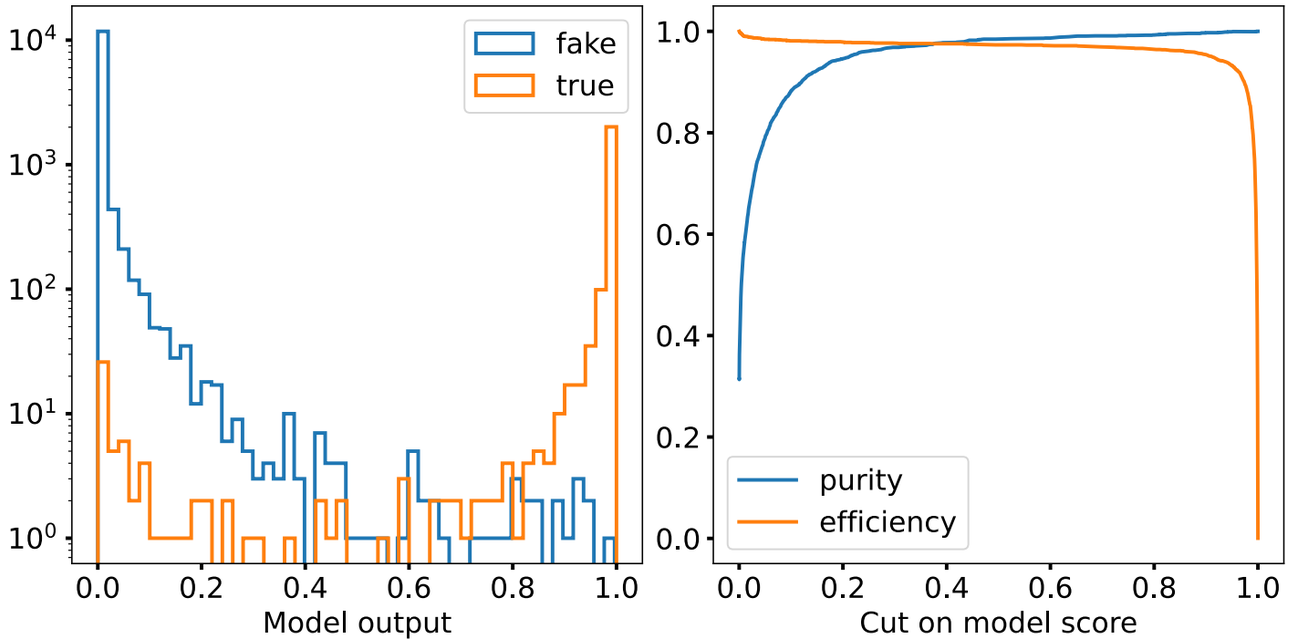






GNN performance

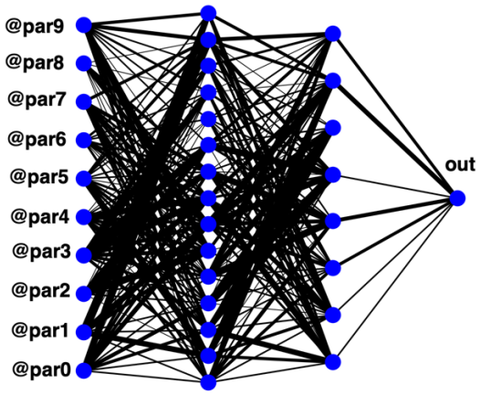
- *This type of graph neural network is not yet supported in HLS4ML.*
- *So we did a manual conversion first to C++ and then to Verilog using Vitis_HLS.*
- *This neural network has not been optimized/pruned, so it consumes a lot of resources - 70% of DSPs, (4651 of 6840).*
  - Network use precision ap fixed < 16,9 >
  - 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).*

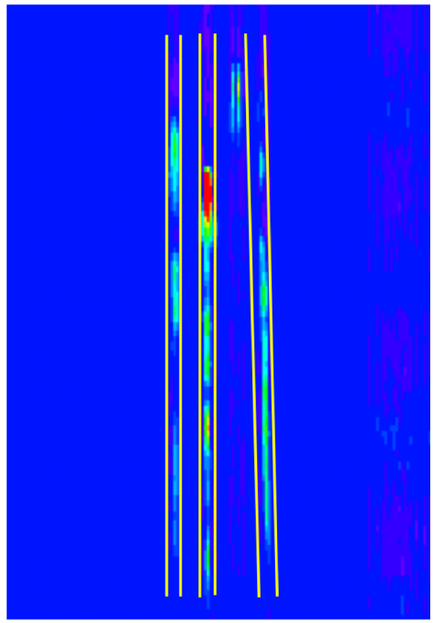


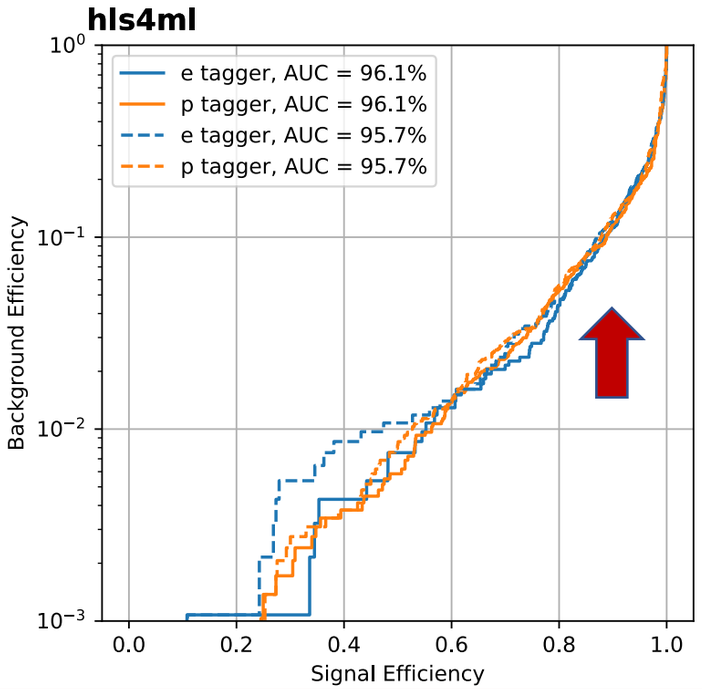
Modules & Loops	Issue Type	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
gnn2dfs2		-	589	2.945E3	-	590	-	no	42	4424	394036	2519454	0
toGraph		-	499	2.495E3	-	497	-	dataflow	42	4424	391308	2515320	0
fromGraph		-	331	1.655E3	-	1	-	yes	0	0	197686	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	-	-	-	-	-

MLP neural network for PID

- *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|
+-----------------+---------+-------+--------+--------+-----+
```

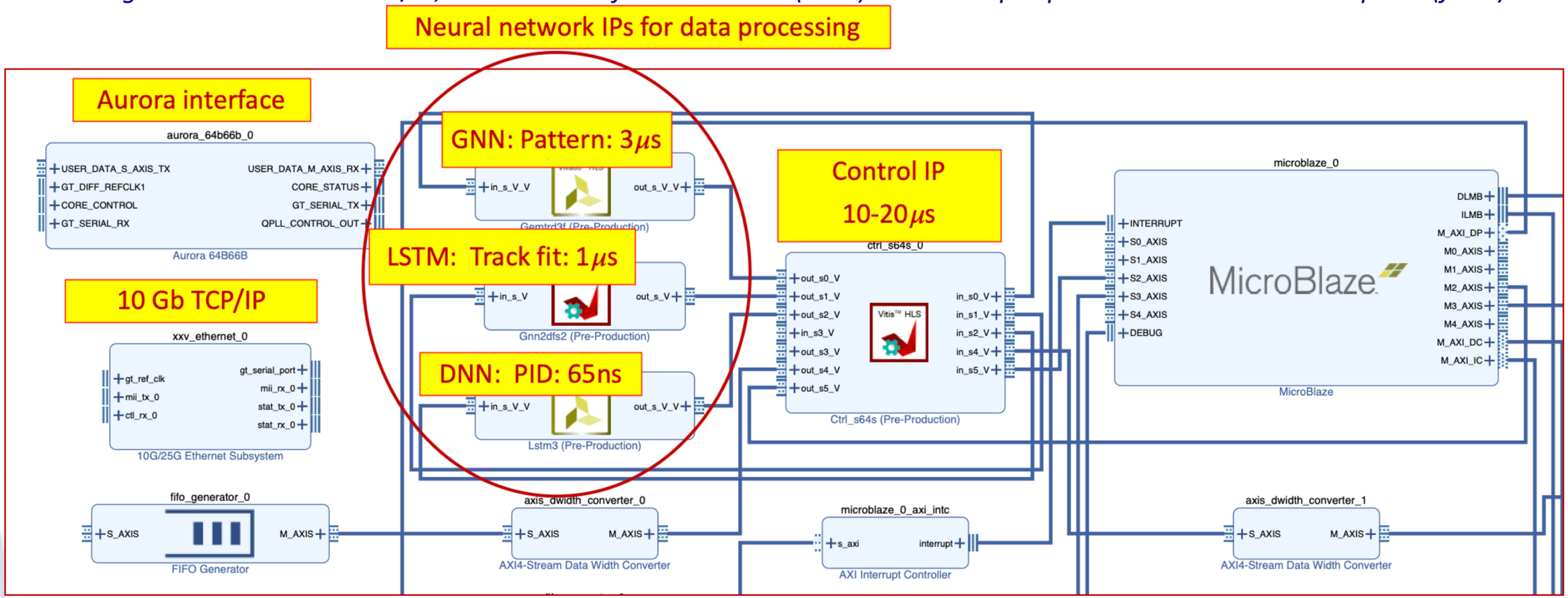
Latency = 65ns

II = 5ns

DSP utilization 3%

Board design

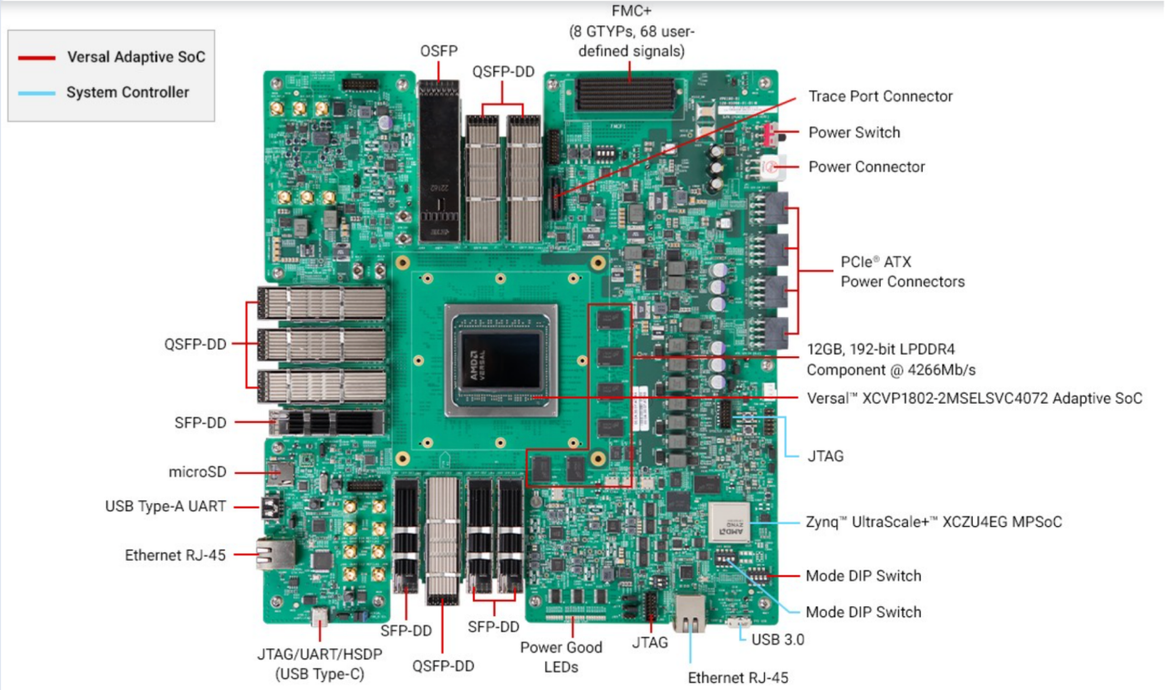
- *All data I/O operations are performed by Control IP*
- *MicroBlaze is only used to configure the board and monitor data processing.*
- *Aurora interface provides communication with a second FPGA board that processes the calorimeter data (CNN).*
- *10 Gigabit Ethernet uses TCP/IP, receives data from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



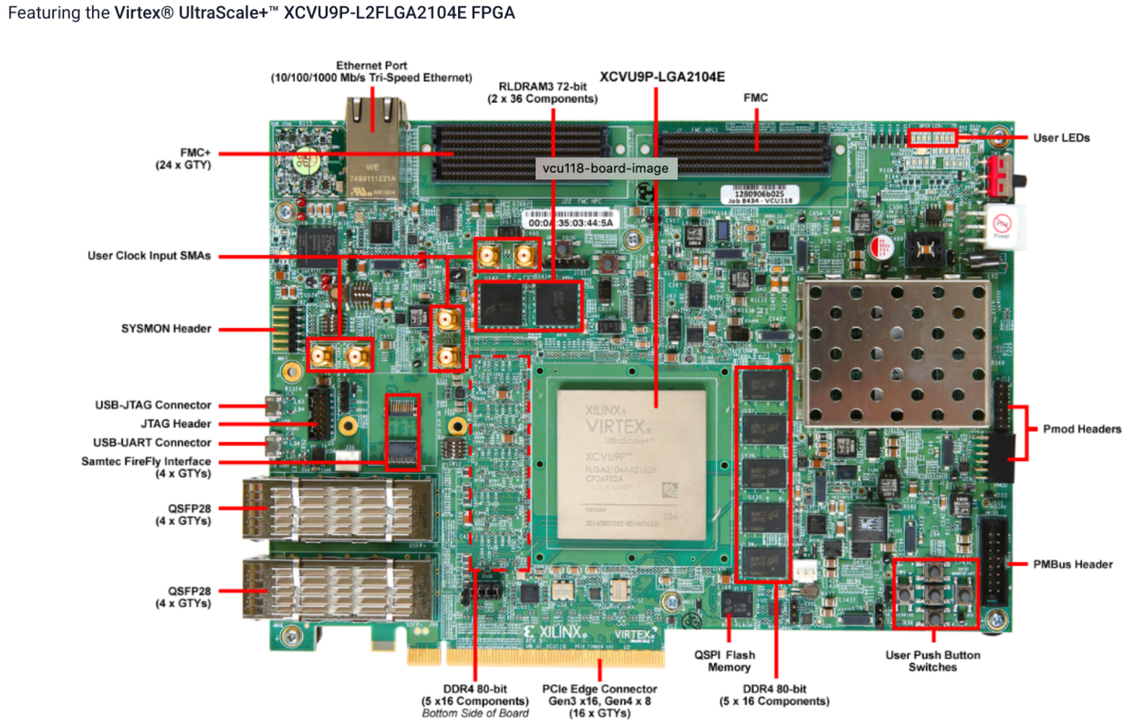
FPGA boards for ML development

- For development ML algorithms on FPGA , we use a standard Xilinx evaluation boards.
- These boards have functions and interfaces sufficient for proof of principle of ML-FPGA.

AMD Versal™ Premium
VPK180 14,352 DSP

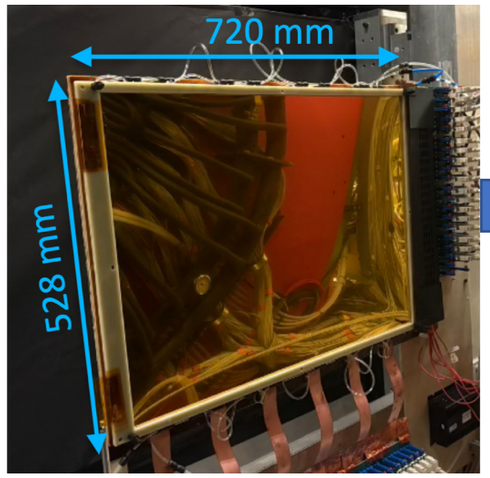


Xilinx Virtex® UltraScale+™
VCU118 6,840 DSP

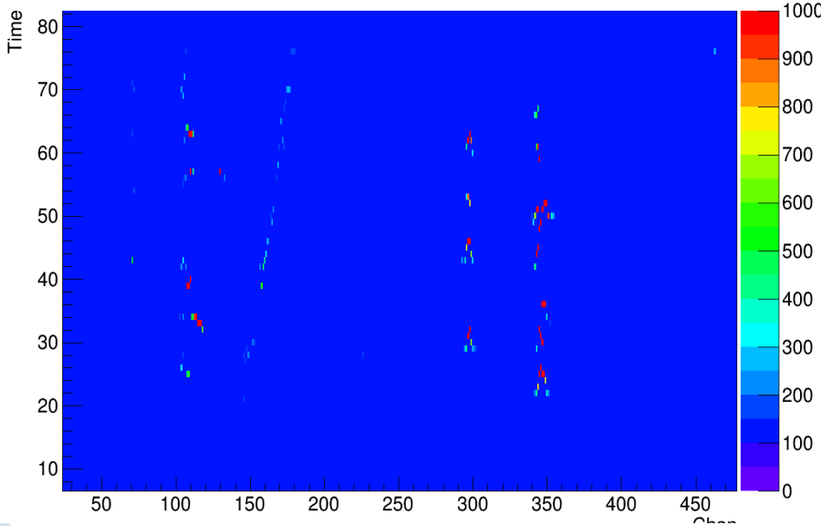


Data flow diagram

GEMTRD detector

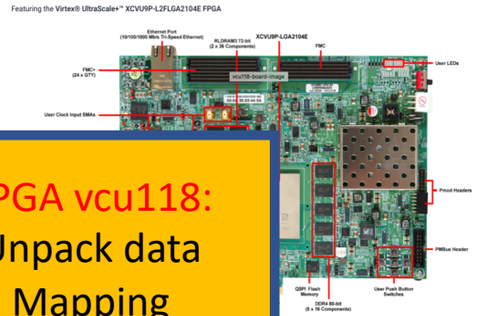


roctrd2:F125_display_x



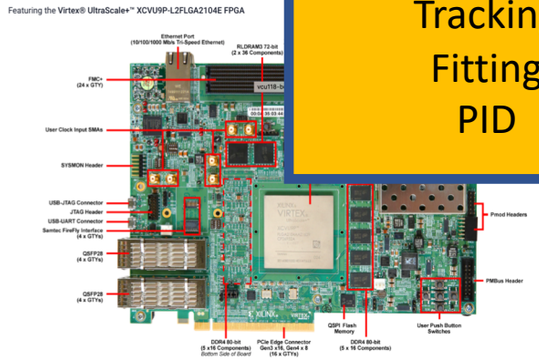
FADC125/
VXS Crate

ETH0
tcp/ip



FPGA vcu118:
Unpack data
Mapping
Clustering

Aurora

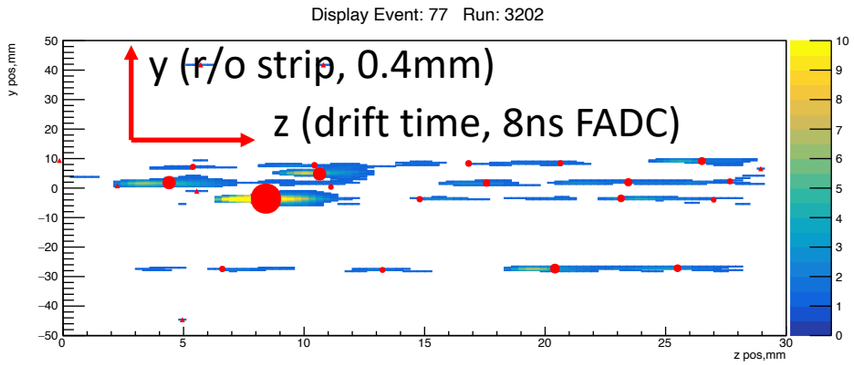


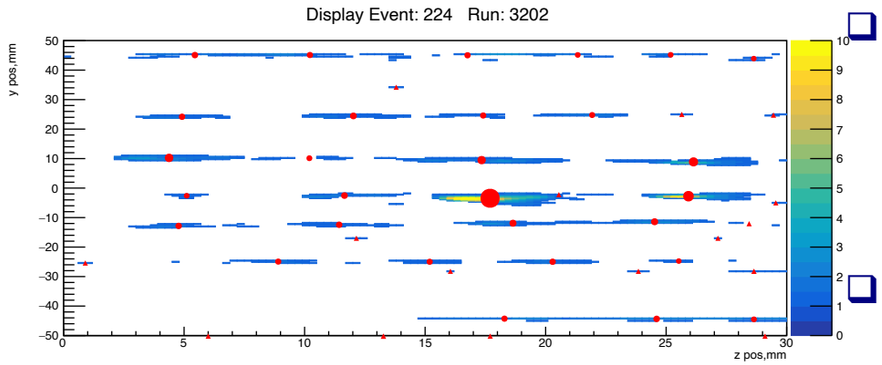
FPGA vcu118
GNN, LSTM, DNN
Tracking
Fitting
PID

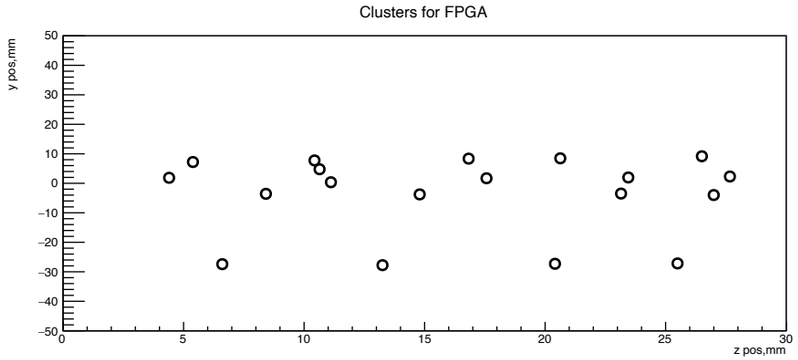
ETH1
tcp/ip

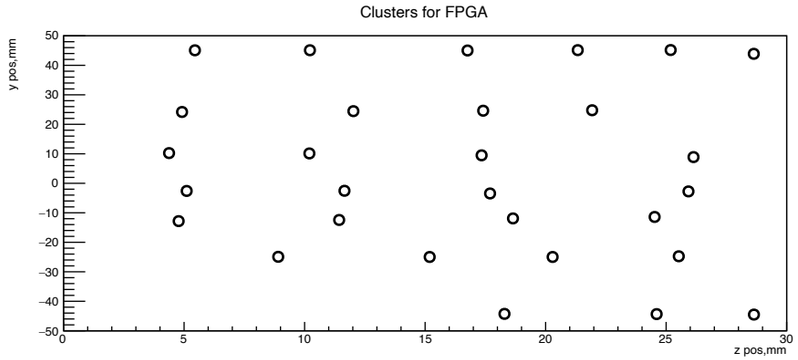
DAQ
Farm
Jana 2

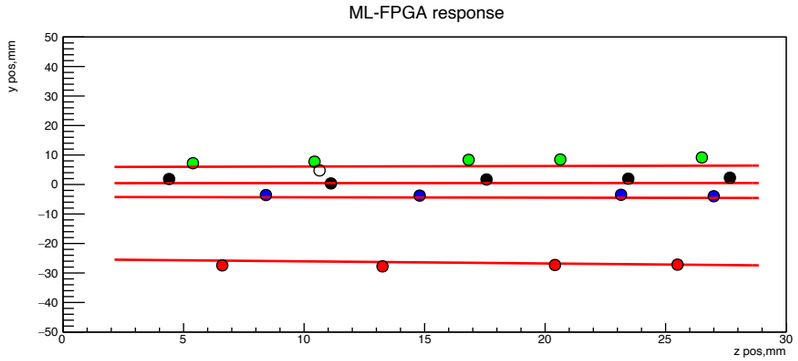
Tracking performance

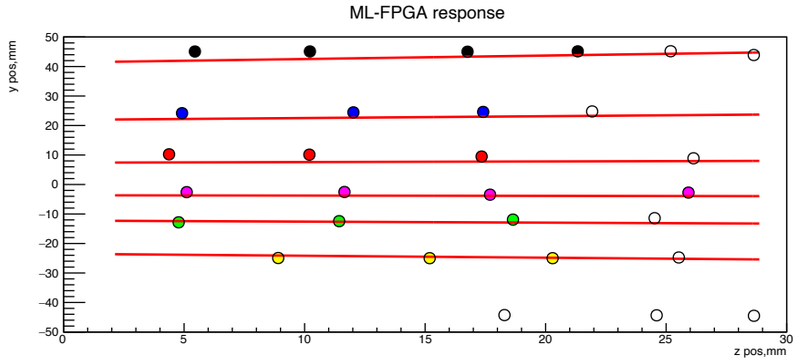










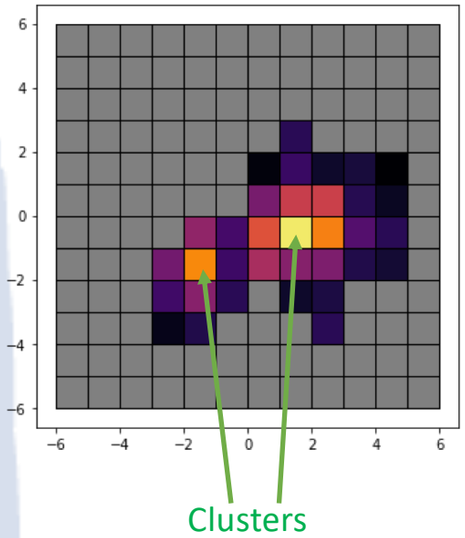


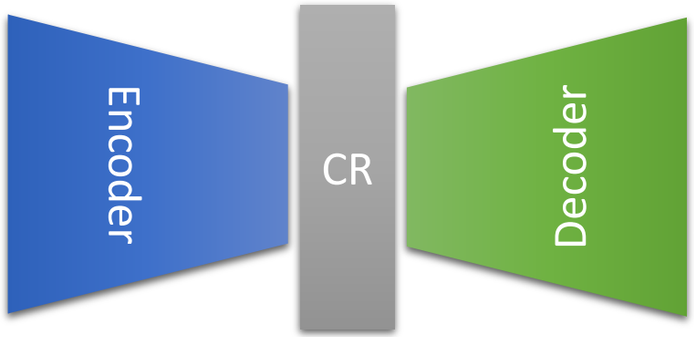
- *Top rows: show ionization along the track in GEMTRD detector.*
  - Red circles are reconstructed clusters using some dE/dx threshold. The size is proportional to energy.
- *Middle rows: after filtering out the noisy clusters, the coordinates of the clusters are sent to the FPGA/GNN for pattern recognition.*
- *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.*

Calorimeter parameters reconstruction

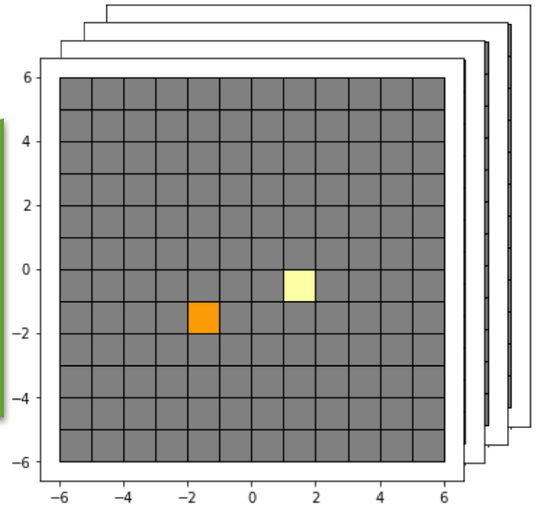
By Dmitry Romanov

Modules deposits





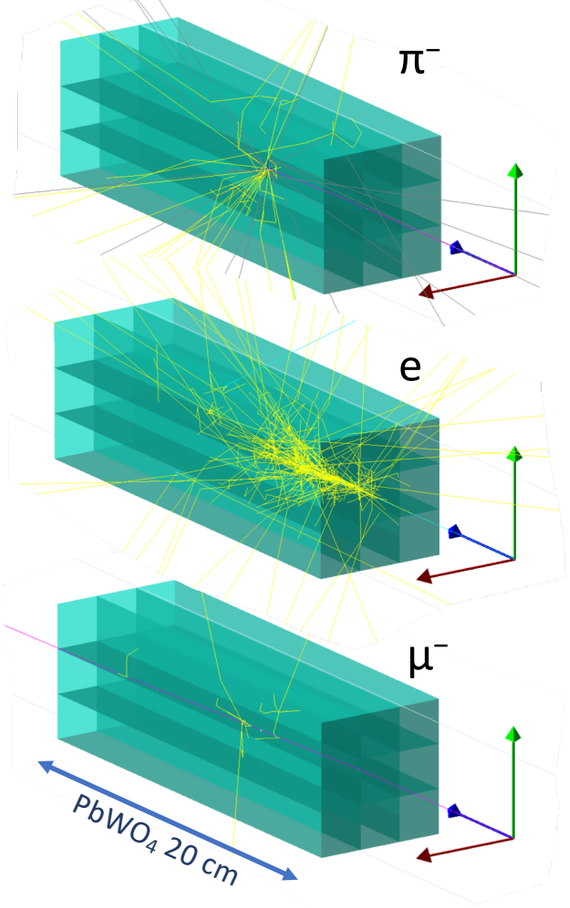
Convolutional variational autoencoder



Per cluster values: energy, x, y, features

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

Geant 4 simulation



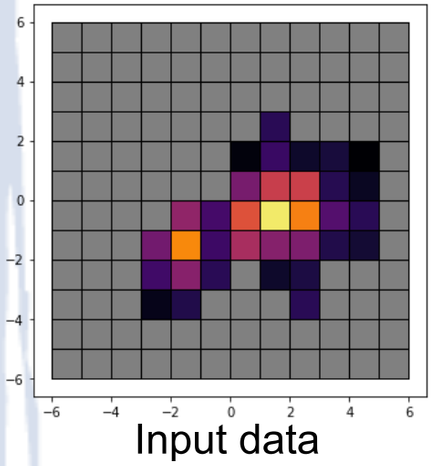
Examples of events with e and π^- showers and μ^- passing through.

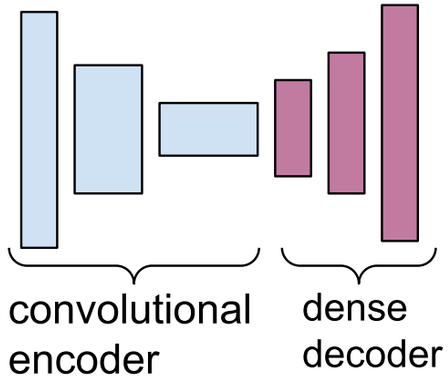
CNN for calorimeter reconstruction

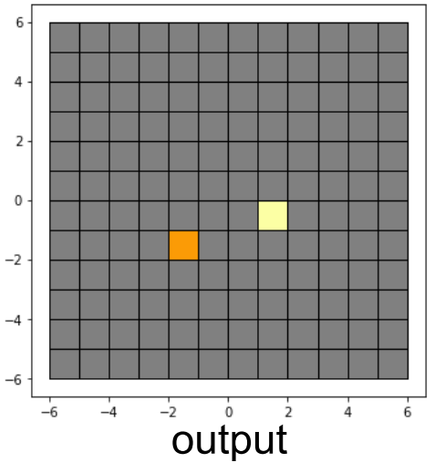
- *In this work we used a convolutional encoder with a decoder consisting of dense layers, which provide e-π separation scores as the output.*
- *Synthesized with HLS4ML, for calorimeter 11x11 cells.*
- *This was done to minimize a network size in FPGA and due to current limitation of HSL4ML of supported network layer types.*
- *FPGA synthesis with reuse factor of 1 has a latency of 0.7μs and an interval of 125 clocks. It uses 74% of DPS resources*
- *Network use precision ap fixed < 20,10 >*

```
+--------+-------+----------+------------+
|  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 |
  +-----+-----+-----+-----+----------+
```





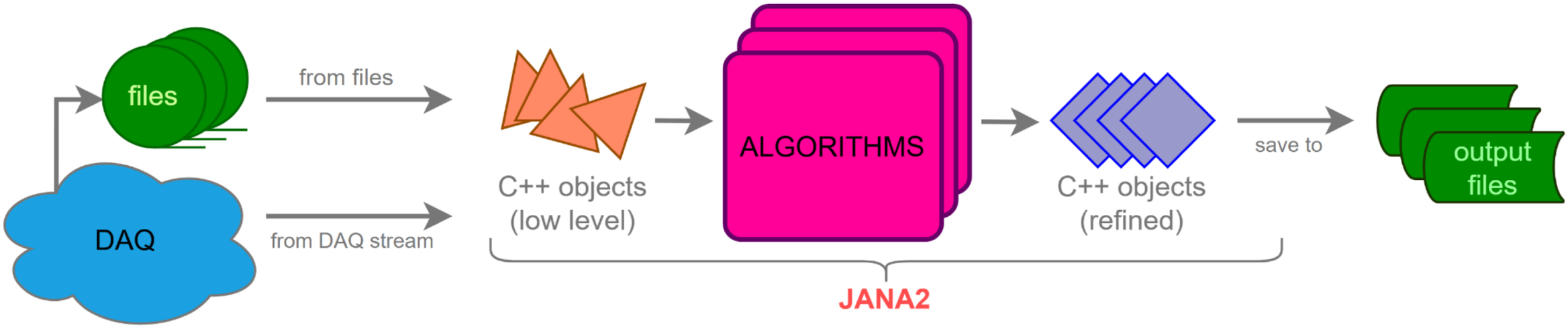


Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	2	-
FIFO	404	-	8999	15698	-
Instance	61	5124	55854	243846	-
Memory	-	-	-	-	-
Multiplexer	-	-	-	-	-
Register	-	-	-	-	-
Total	465	5124	64853	259546	0
Available SLR	1440	2280	788160	394080	320
Utilization SLR (%)	32	224	8	65	0
Available	4320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

JANA2 for ML on FPGA

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

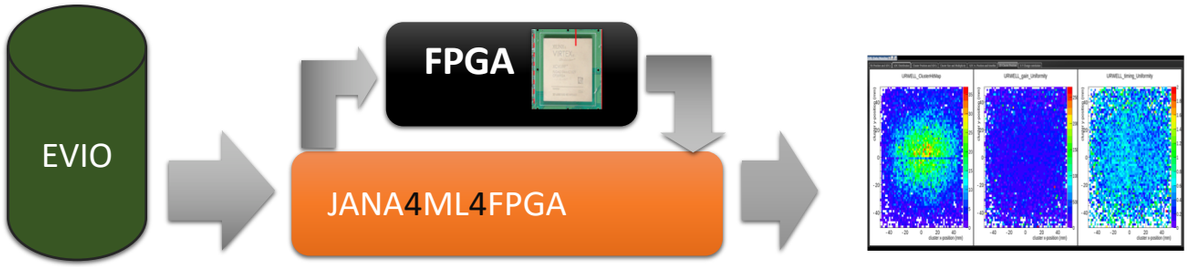
JANA2 Modern C++ framework for HENP data processing



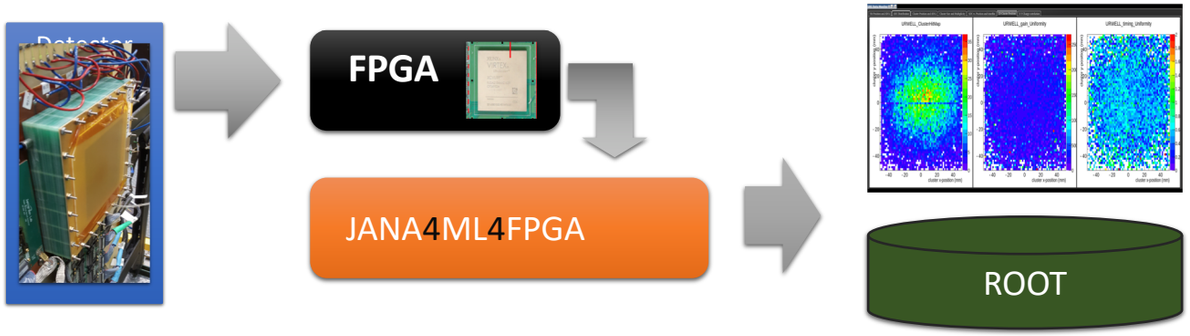
- Very efficient multi-threading (thread-level parallelism).
- Flexible data flow configuration:
  - Lazy parallel algorithms execution
  - Disentangling: data blocks -> stream of chunks
  - Sub-structure parallelism: can to split data and process e.g. on GPU, FPGA, do AI/ML inference
- Modern C++. Organized to scale. Big collaboration projects (e.g. EIC and GlueX)

JANA4ML4FPGA

Specialized JANA2 based software toolkit, that provides full featured software support for the project.



Scenario I: offline data stream simulation



Scenario II: online use

Used:

- FPGA communication
- Online data aggregation
- Offline hardware simulation
- Monitoring
- Data persistence

Outlook

- *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.*

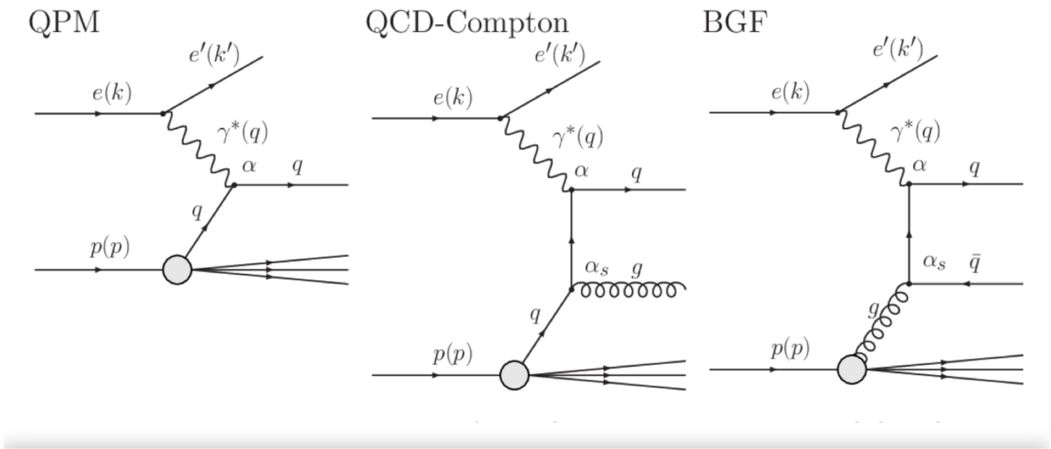


Figure 2.1: Feynman diagrams of the Quark Parton Model, QCD-Compton and Boson Gluon Fusion processes in NC DIS.

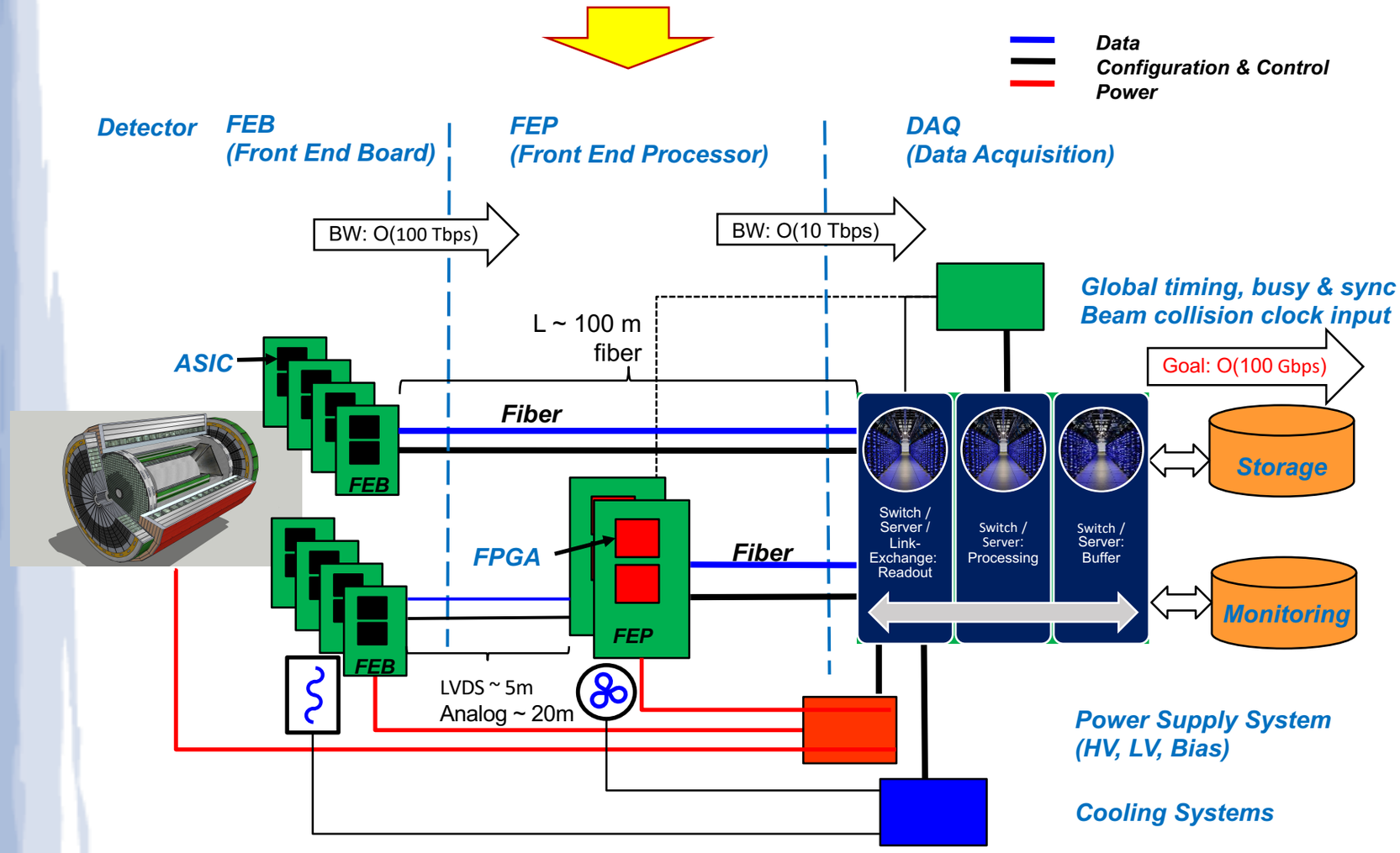
Published in 2007

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

T. Gosau

Backup

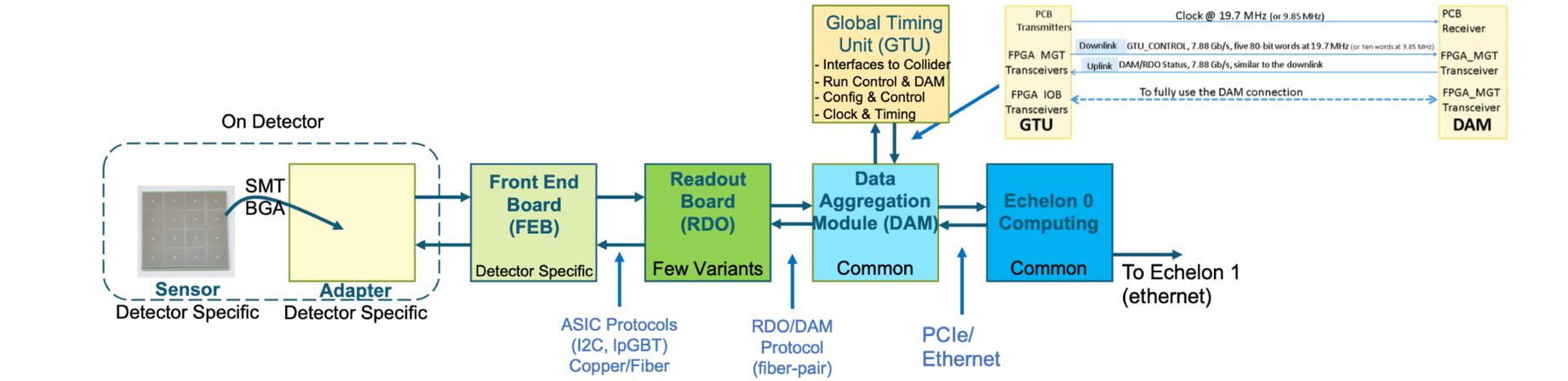
EIC streaming readout as motivation for ML-FPGA

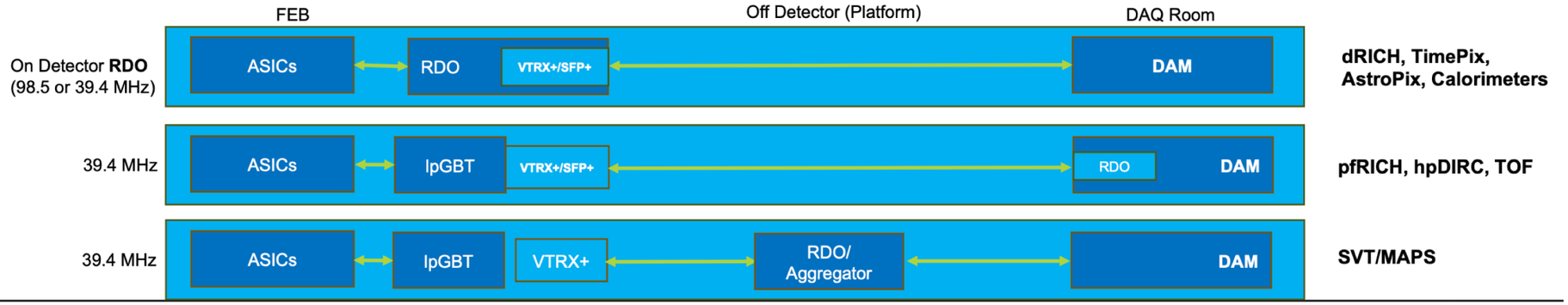


- *The correct location for the ML on the FPGA filter is called "FEP" in this figure.*
- *This gives us a chance to reduce traffic earlier.*
- *Allows us to touch physics: ML brings intelligence to L1.*
- *However, it is now unclear how far we can go with physics at the FPGA.*
- *Initially, we can start 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.*

Updated EIC streaming readout

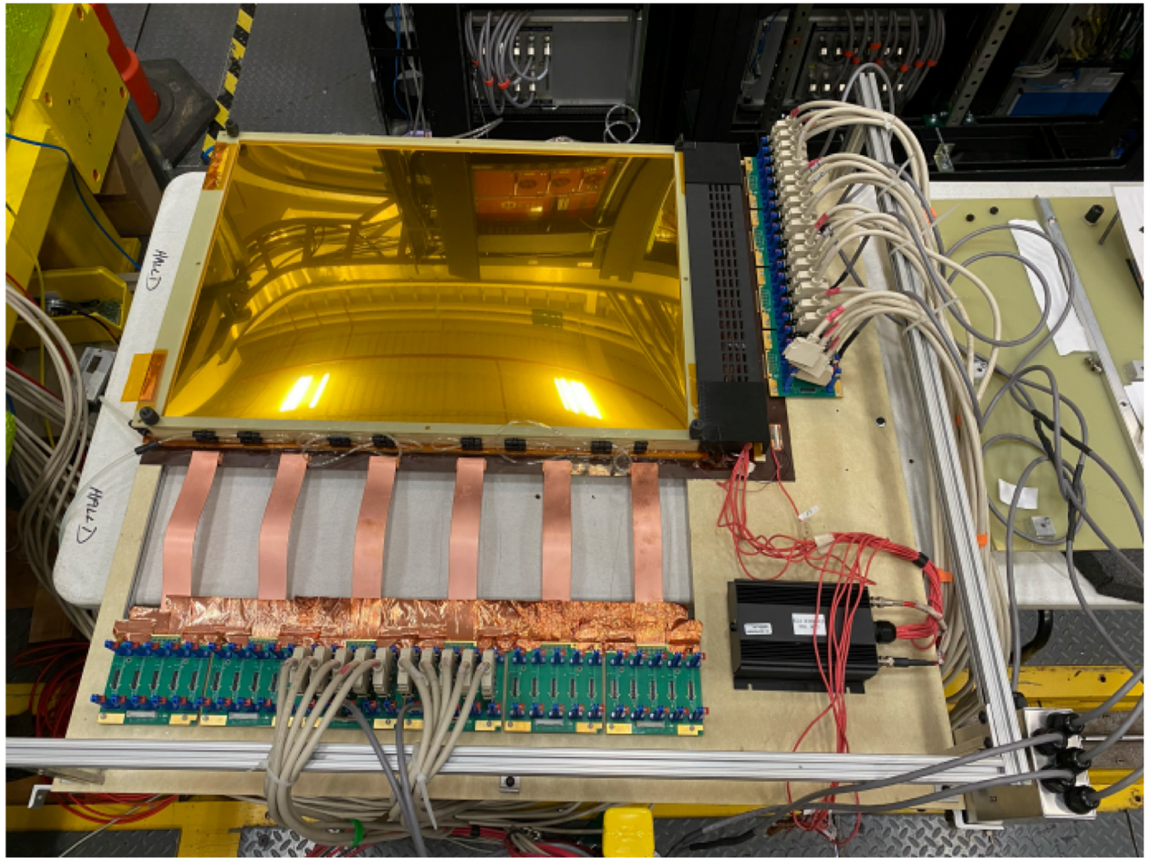
ePIC Readout Chain





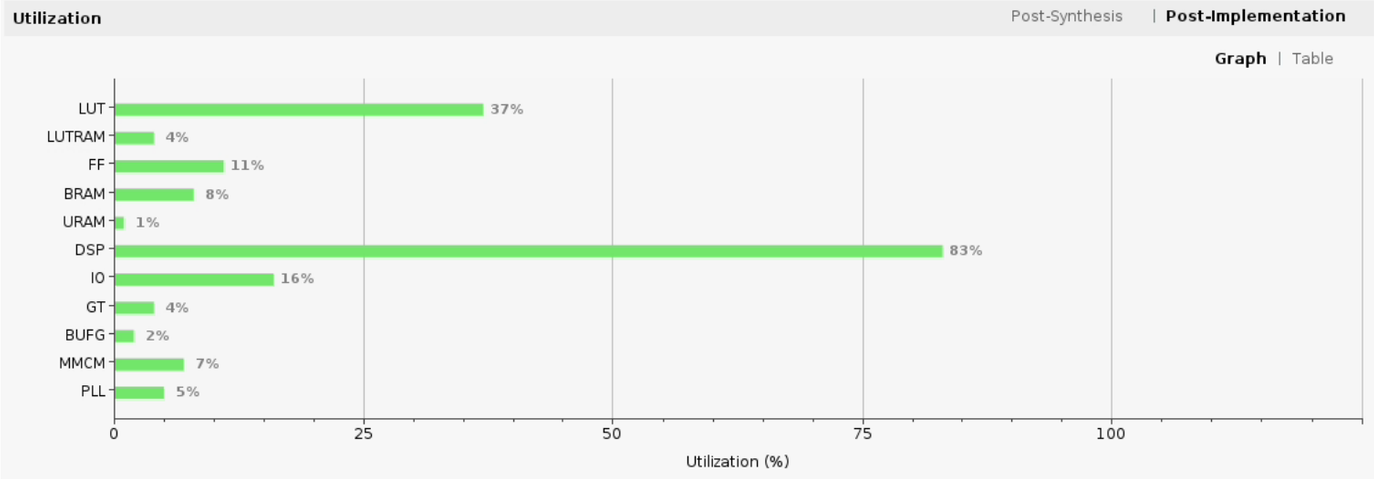
Electron-Ion Collider

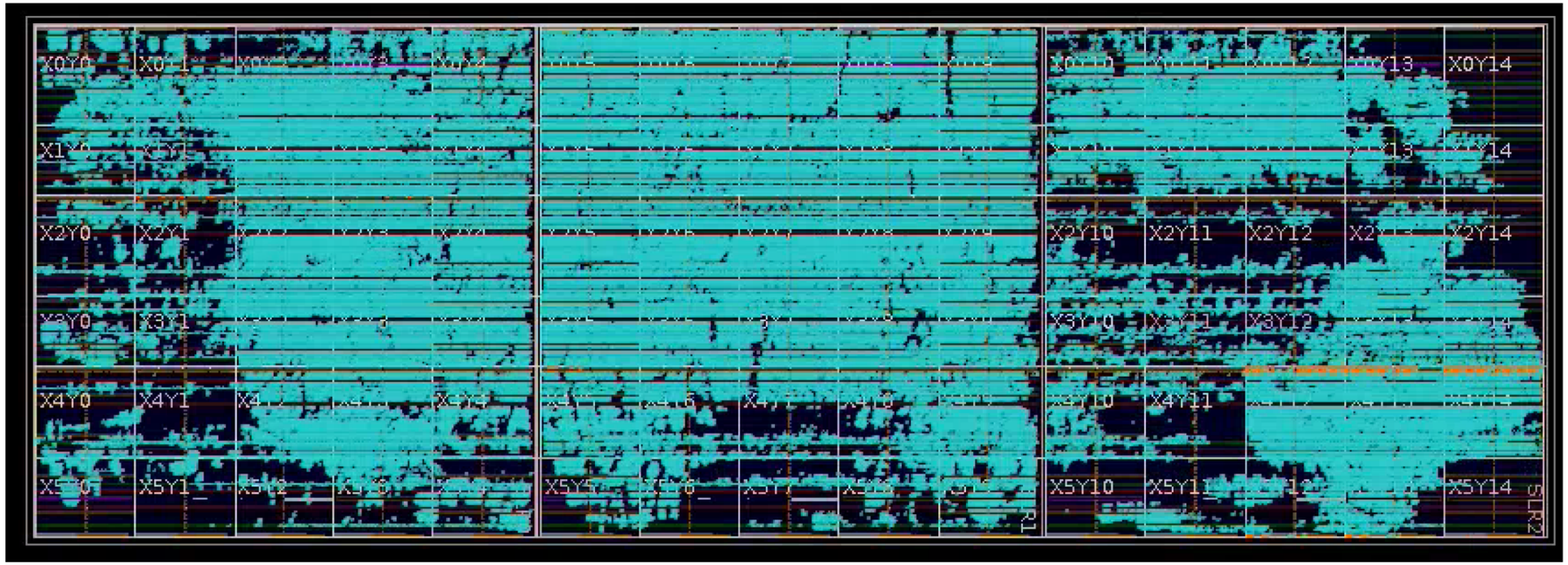
ePIC Collaboration Meeting, July 13-17, 2025 D. Abbott 4



FPGA board resources for GEMTRD

- *Neural networks use a lot of FPGA resources.*
- *Therefore, one VCU118 board can only process data from GEMTRD.*



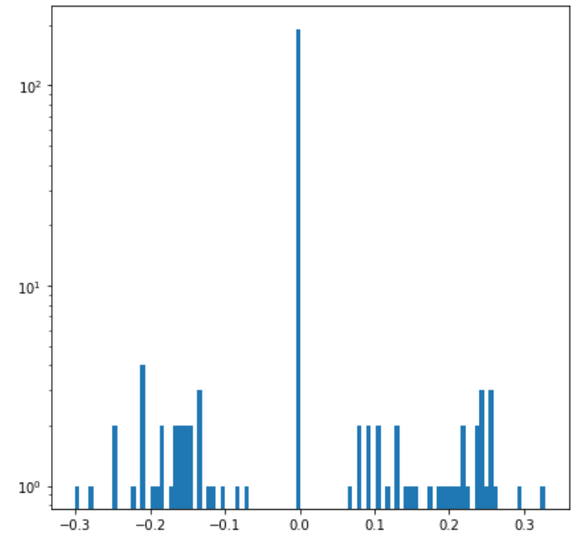


RNN/LSTM for track fit

- *The hits sorted by tracks from the pattern recognition GNN are fed into another neural network trained to fit the tracks.*
- *We use RNN/LSTM neural networks. (thanks to Dylan Rankin for help)*
  - Network use precision ap fixed < 24,11 >
  - The input layer is designed for 26 hits.
  - 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  |  Interval | Pipeline |
    | min | max | min | max |   Type   |
    +-----+-----+-----+-----+----------+
    |  213|  213|  208|  208| function |
    +-----+-----+-----+-----+----------+
```

```
% of zeros = 0.75
```



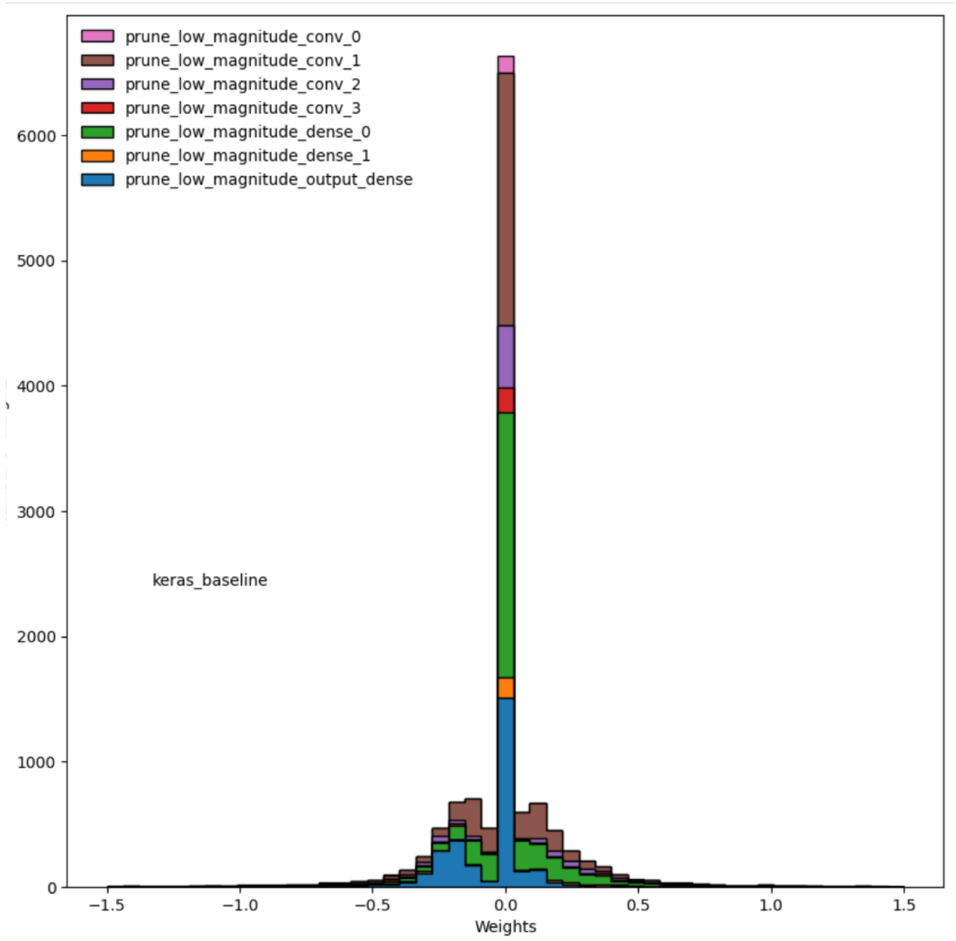

```
================================================================
== Utilization Estimates
================================================================
* Summary:
+---------------------+---------+-------+---------+---------+-----+
|         Name        | BRAM_18K| DSP48E|    FF   |   LUT   | URAM|
+---------------------+---------+-------+---------+---------+-----+
|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|
+---------------------+---------+-------+---------+---------+-----+
```

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
```



Optimized GNN IP

- *The GlueX trigger rate is up to 70 kHz, so on average we have ~14 µs to process events.*
- *We optimized the GNN to have a latency of ~10 µs, which allows it to operate at 70 kHz.*
- *On the other hand, the neural network fits in an FPGA and supports 150 nodes and 256 edges.*
- *Next we plan to test it on hardware.*

MODULES & LOOPS	IS TY	SLACK	LATENCY(CYCLES)	LATENCY(NS)	ITERATION LATENCY	INTERVAL	TRIP COUNT	PIPELINED	BRAM(%	DSP(%)	FF(%)	LUT(%)
runGraphNetwork (6)		-0.24	1991	9.955E3	-	1992	-	no	~0	11	10	25
edge_network (1)		-	271	1.355E3	-	271	-	no	~0	3	~0	3
node_runner (1)	⚠	-0.24	363	1.815E3	-	363	-	no	0	5	7	21
runGraphNetwork_Pipeline_INPUT_HIT_LOOP (1)		-	159	795.000	-	159	-	no	0	2	1	~0
runGraphNetwork_Pipeline_VITIS_LOOP_42_1 (1)		-	302	1.510E3	-	302	-	no	0	0	~0	~0
runGraphNetwork_Pipeline_VITIS_LOOP_72_2 (1)		-	514	2.570E3	-	514	-	no	0	0	~0	~0
runGraphNetwork_Pipeline_VITIS_LOOP_136_3 (1)		-	258	1.290E3	-	258	-	no	0	0	~0	~0

Event display, single track