Jefferson Lab
Thomas Jefferson National Accelerator Facility

ML on FPGA for real-time particle identification

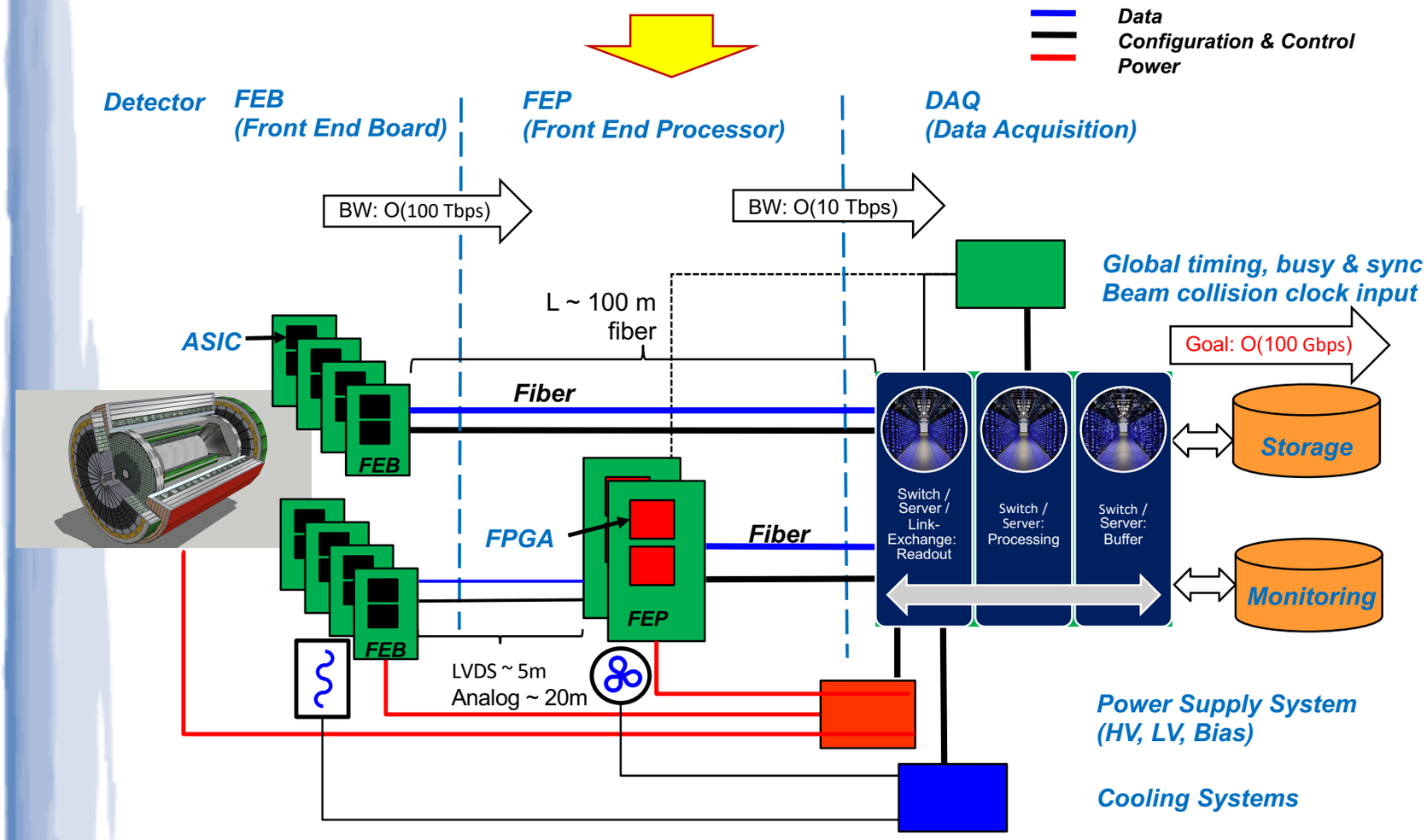
Sergey Furletov
Jefferson Lab

Team :
F. Barbosa, L. Belfore (ODU), C. Dickover, C. Fanelli (MIT), Y. Furletova,
S. Furletov, L. Jokhovets (Jülich Research Centre, Germany),
D. Lawrence, D. Romanov

Workshop on Streaming readout X

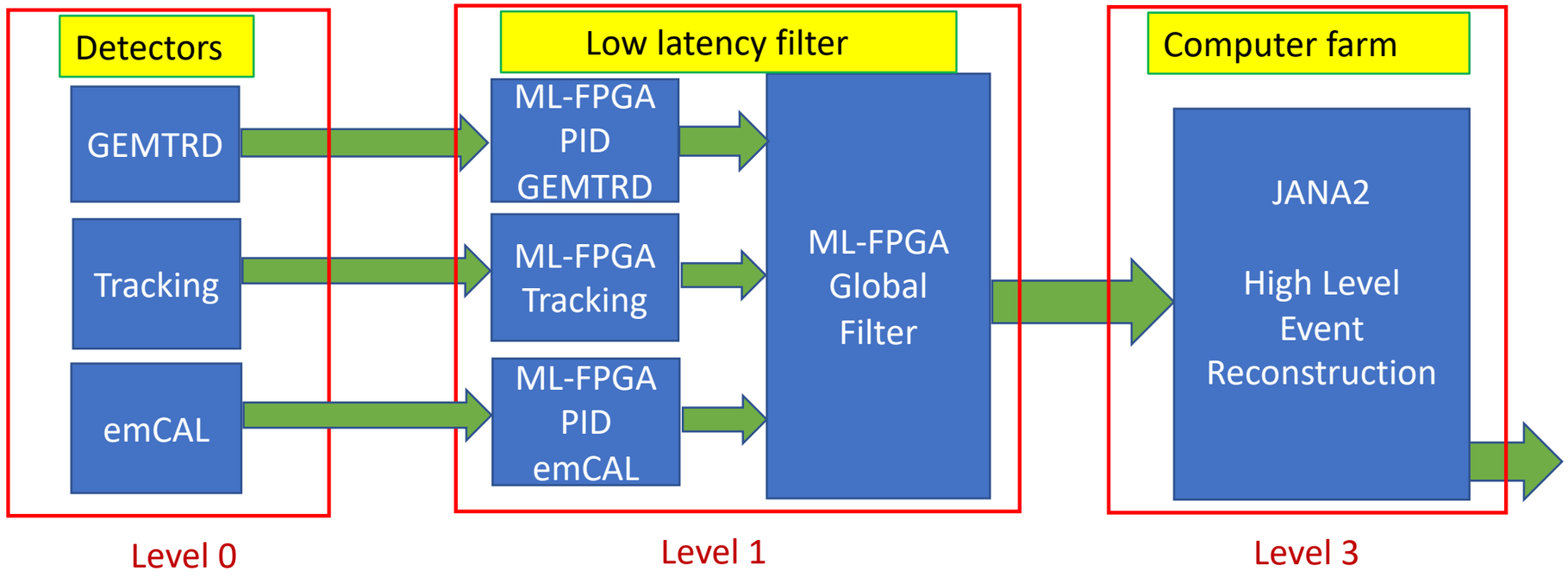
19 May 2022

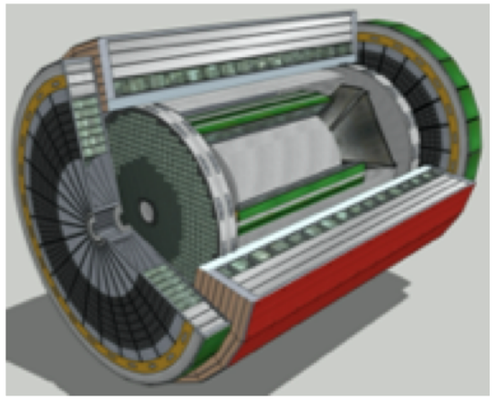
EIC readout as motivation

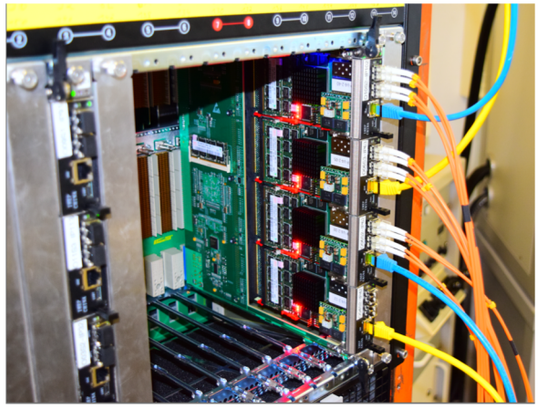


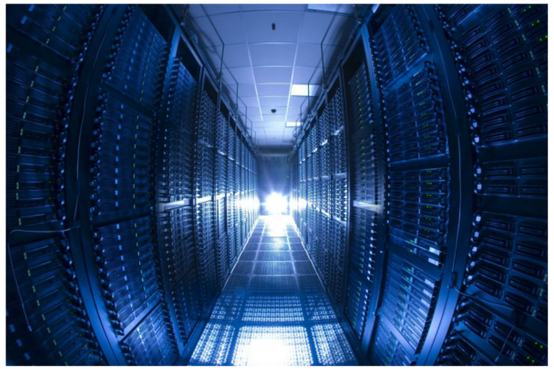
- *The correct location for the ML on the FPGA filter is called "FEP" in this figure.*
- *This gives us a chance to reduce traffic earlier.*
- *Allows us to touch physics: ML brings intelligence to L1.*
- *However, it is now unclear how far we can go with physics at the FPGA.*
- *Initially, we can start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

GEMTRD/emCAL test setup data flow





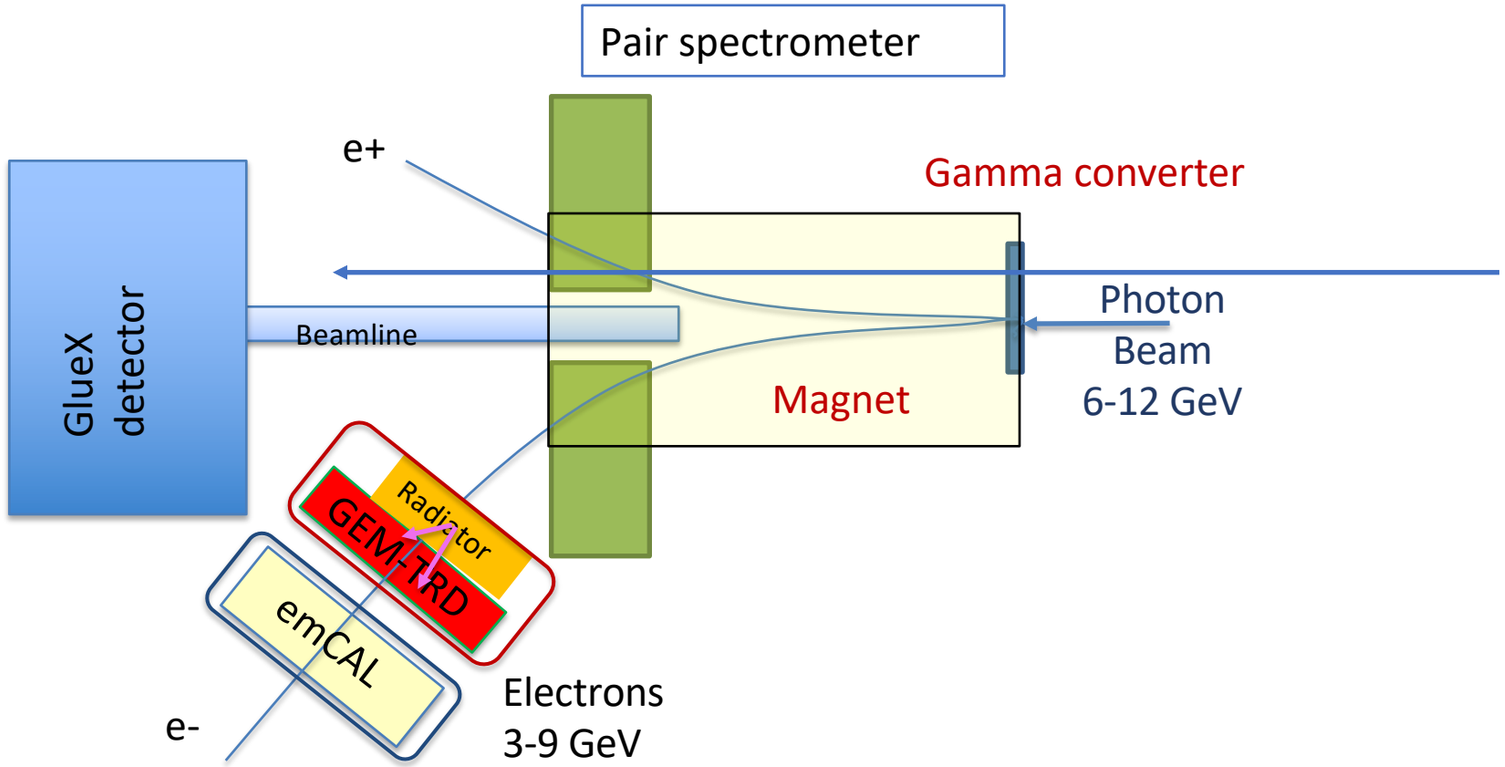




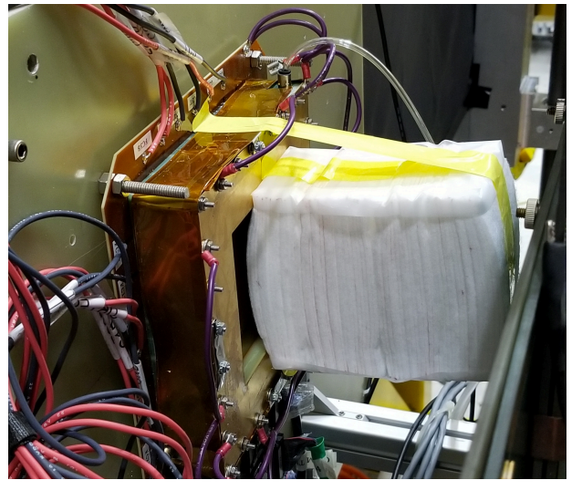
Images from the Internet are for illustration of scale only.

Beam setup at JLab Hall-D

- *Tests were carried out using electrons with an energy of 3-6 GeV, produced in the converter of a pair spectrometer at the upstream of GlueX detector.*



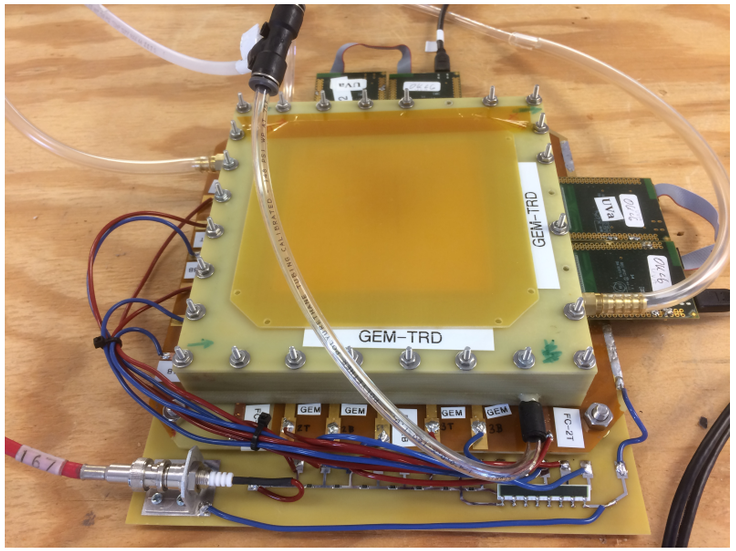
GEMTRD prototype

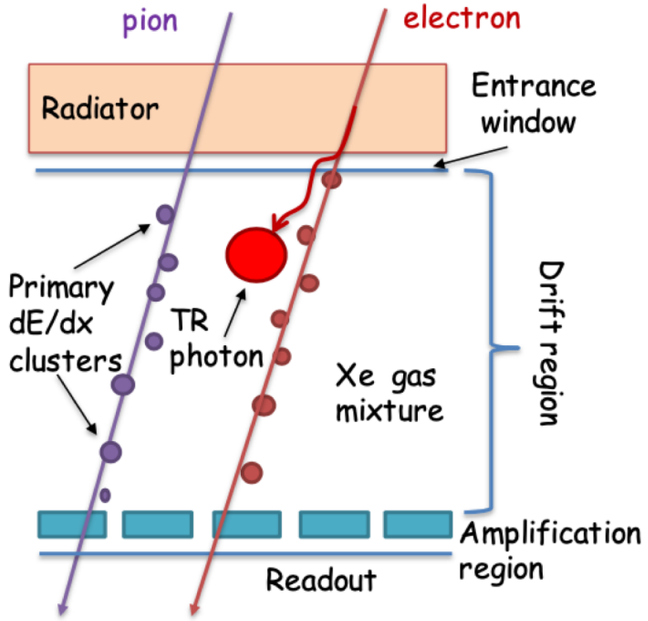


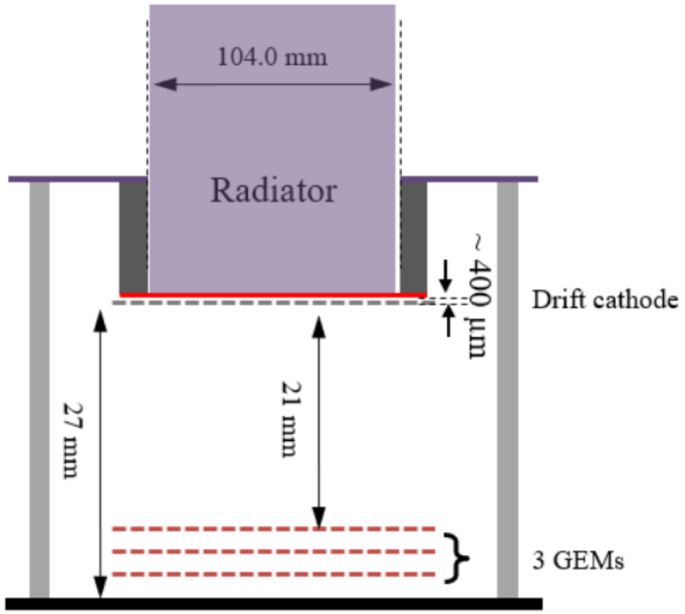
GEM-TRD prototype

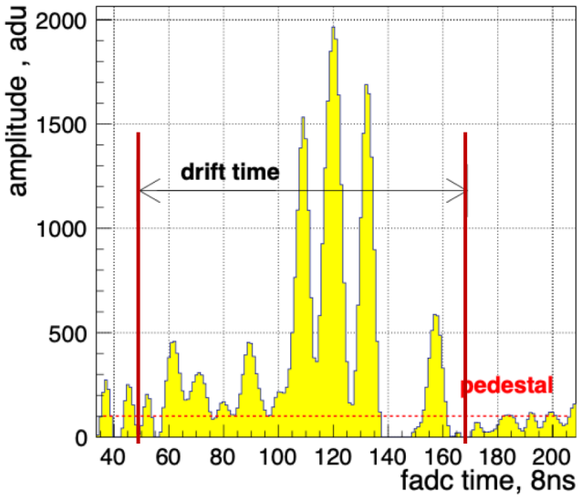
- A test module was built at the University of Virginia
- The prototype of GEMTRD/T module has a size of 10 cm × 10 cm with a corresponding to a total of 512 channels for X/Y coordinates.
- The readout is based on flash ADC system developed at JLAB (fADC125) @125 MHz sampling.

- GEM-TRD provides e/hadron separation and tracking



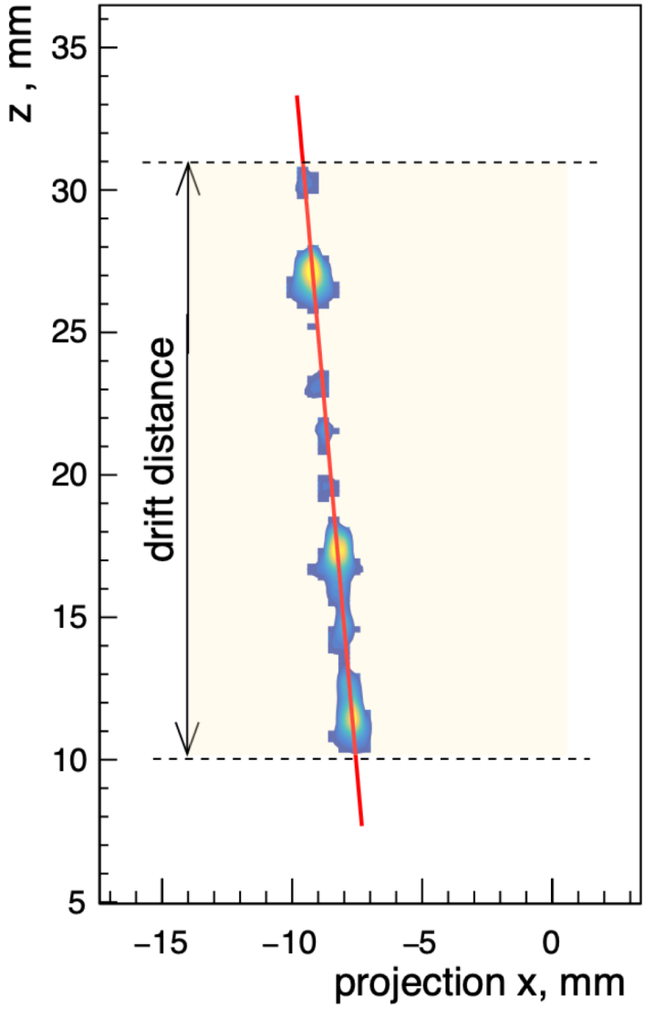


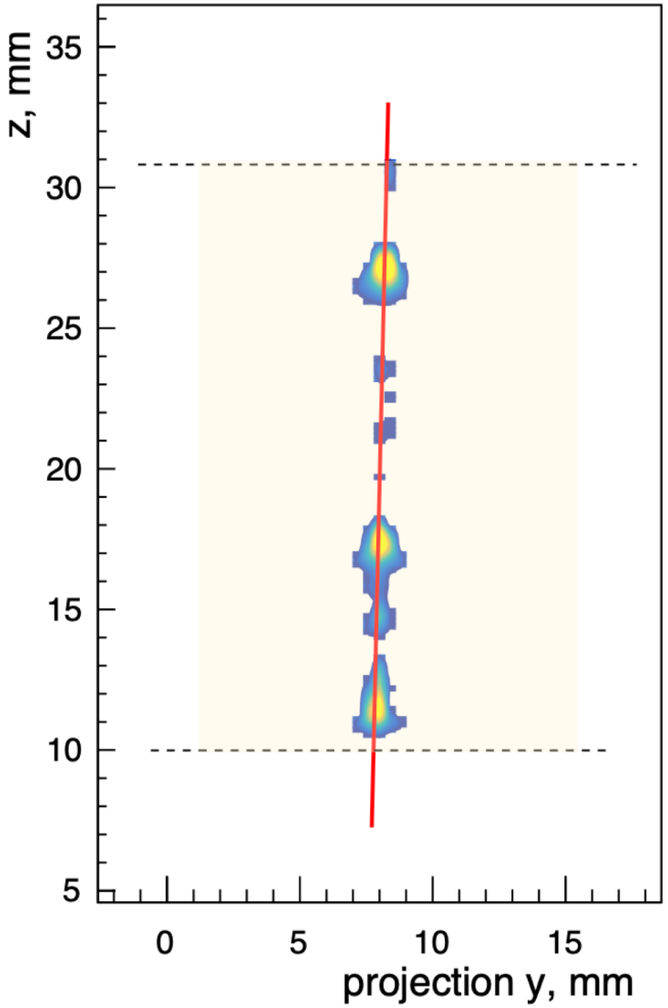




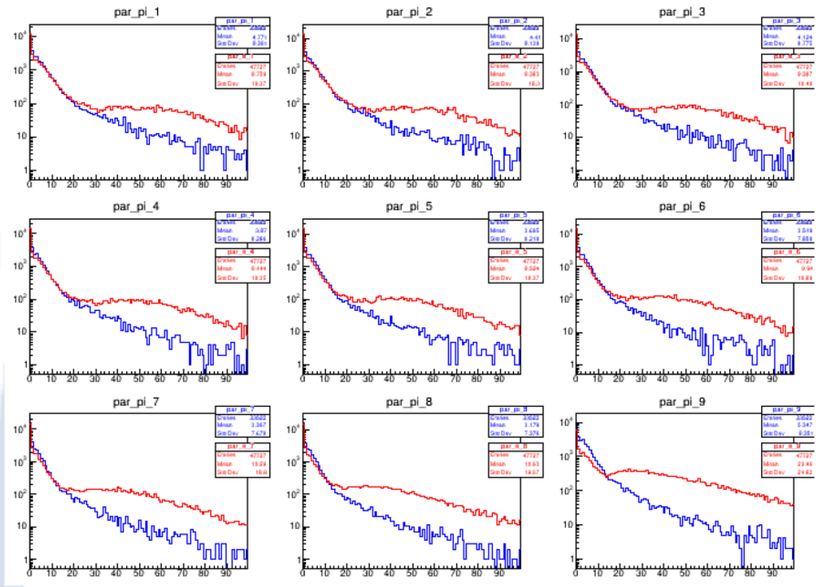
GEMTRD clusters on the track

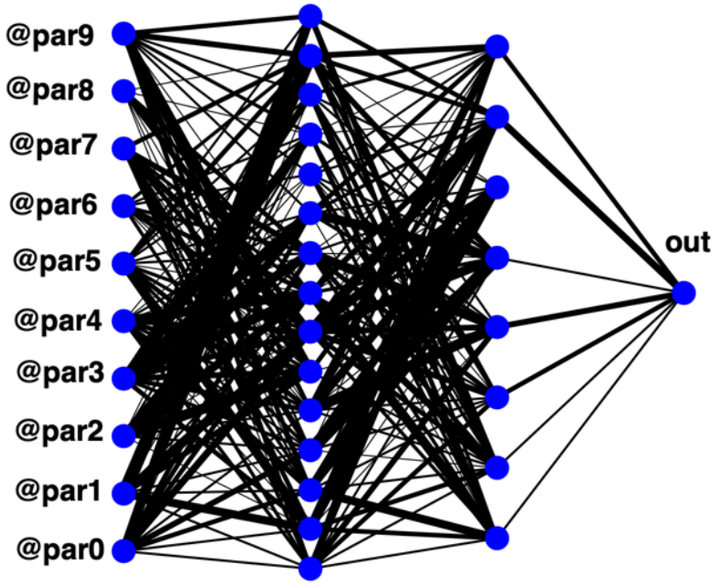
GEM-TRD can work as micro TPC, providing 3D track segments

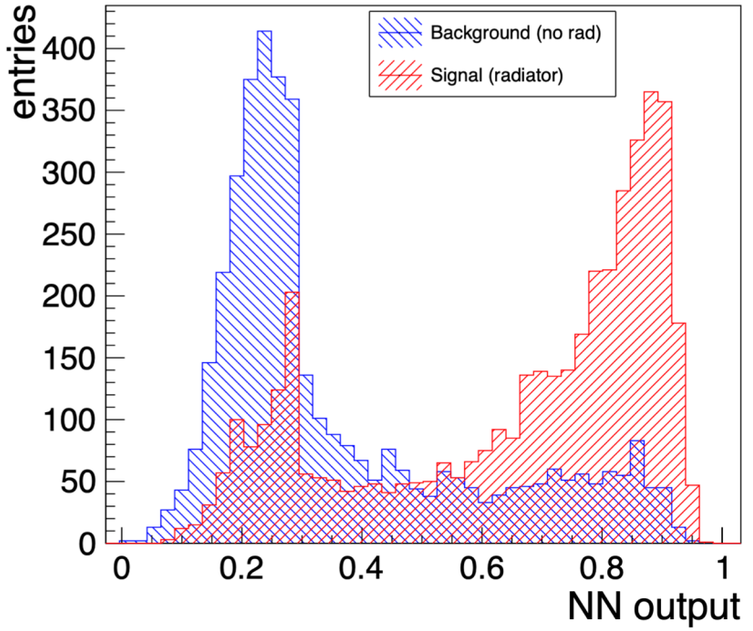




GEMTRD offline analysis







- *For data analysis we used a neural network library provided by root /TMVA package : MultiLayerPerceptron (MLP)*
- *All data was divided into 2 samples: training and test samples*
- *Top right plot shows neural network output for single module:*
  - Red - electrons with radiator
  - Blue – electrons without radiator

DNN in FPGA for GEMTRD as PID

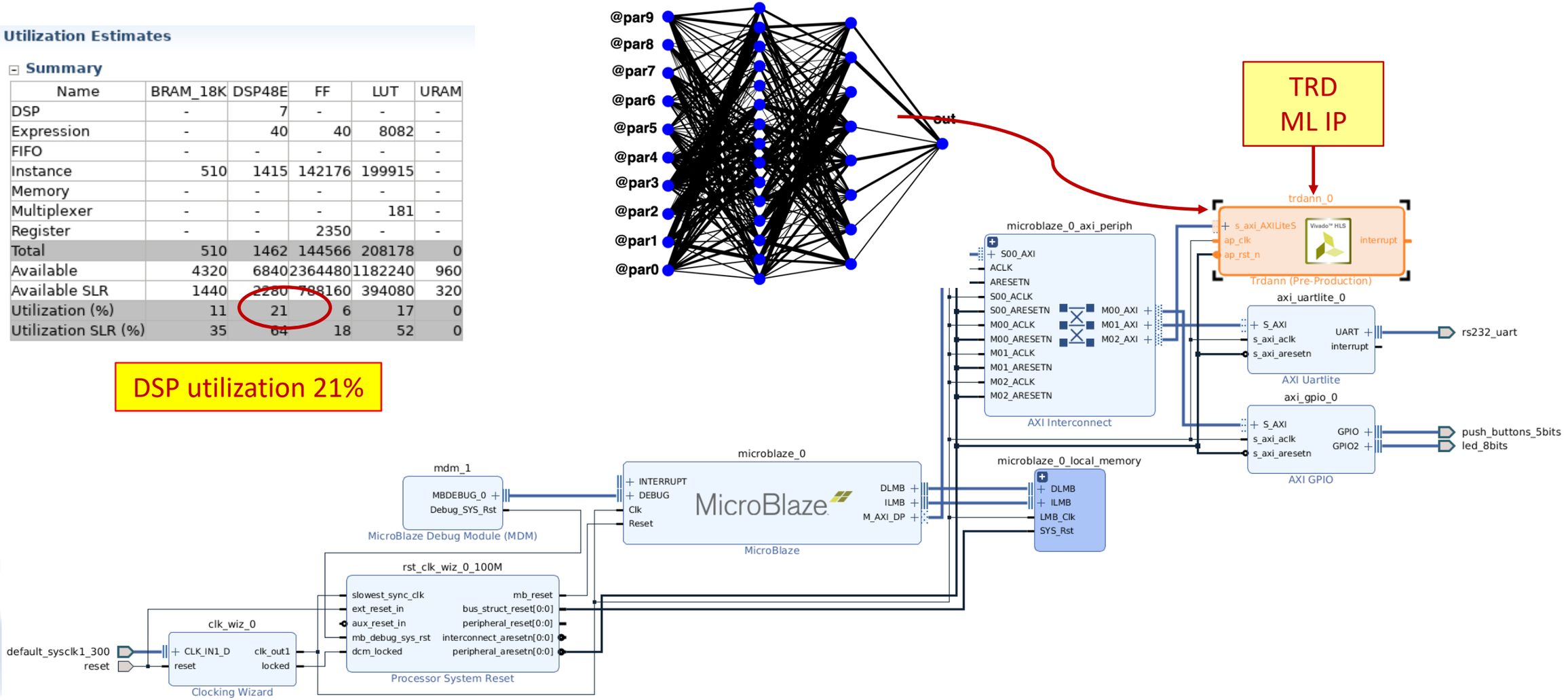
- *Using HLS significantly decreases development time. (at the cost of lower efficiency of use of FPGA resources)*

Utilization Estimates

Summary

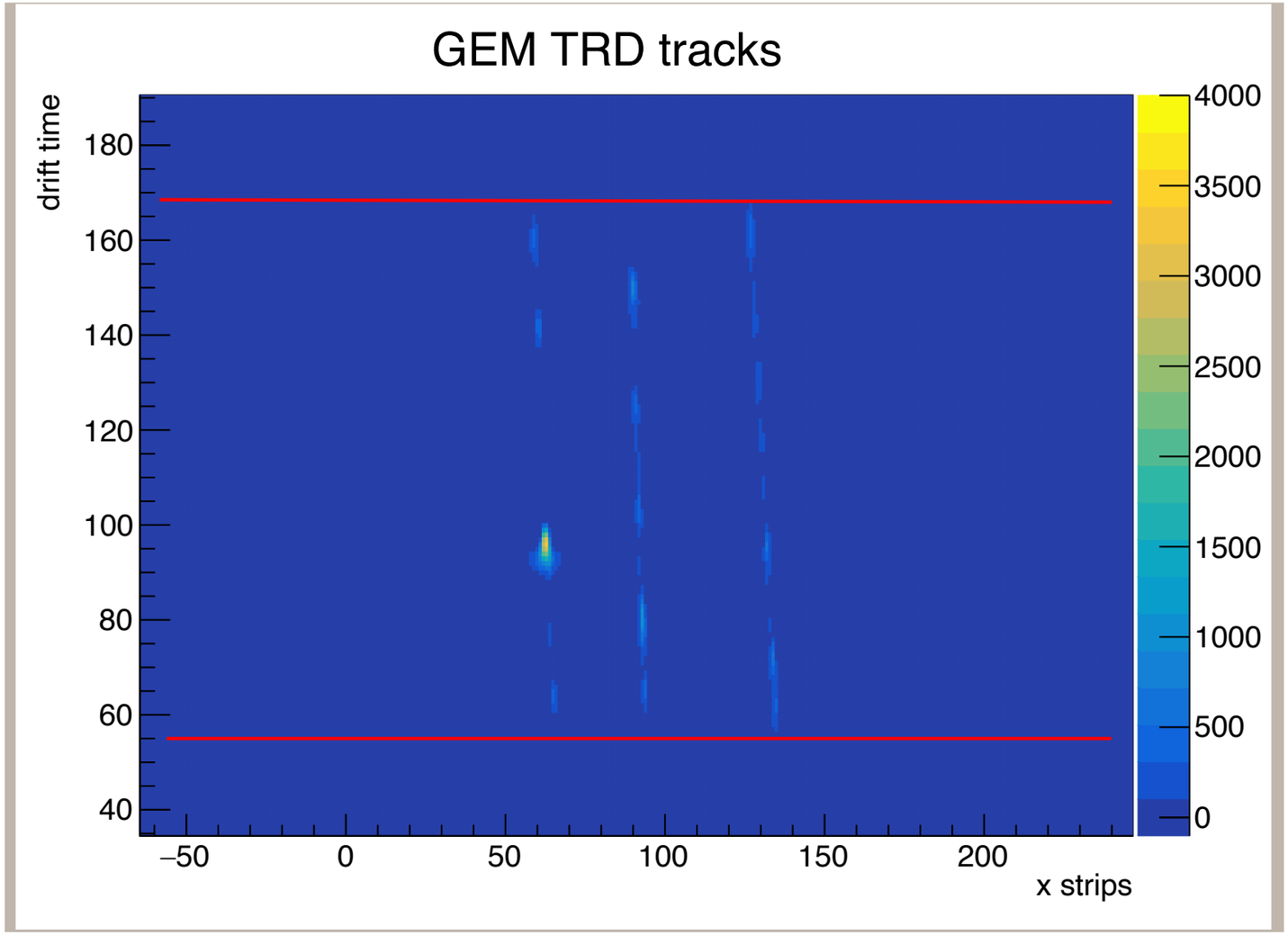
Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	7	-	-	-
Expression	-	40	40	8082	-
FIFO	-	-	-	-	-
Instance	510	1415	142176	199915	-
Memory	-	-	-	-	-
Multiplexer	-	-	-	181	-
Register	-	-	2350	-	-
Total	510	1462	144566	208178	0
Available	4320	6840	2364480	1182240	960
Available SLR	1440	2280	788160	394080	320
Utilization (%)	11	21	6	17	0
Utilization SLR (%)	35	64	18	52	0

DSP utilization 21%



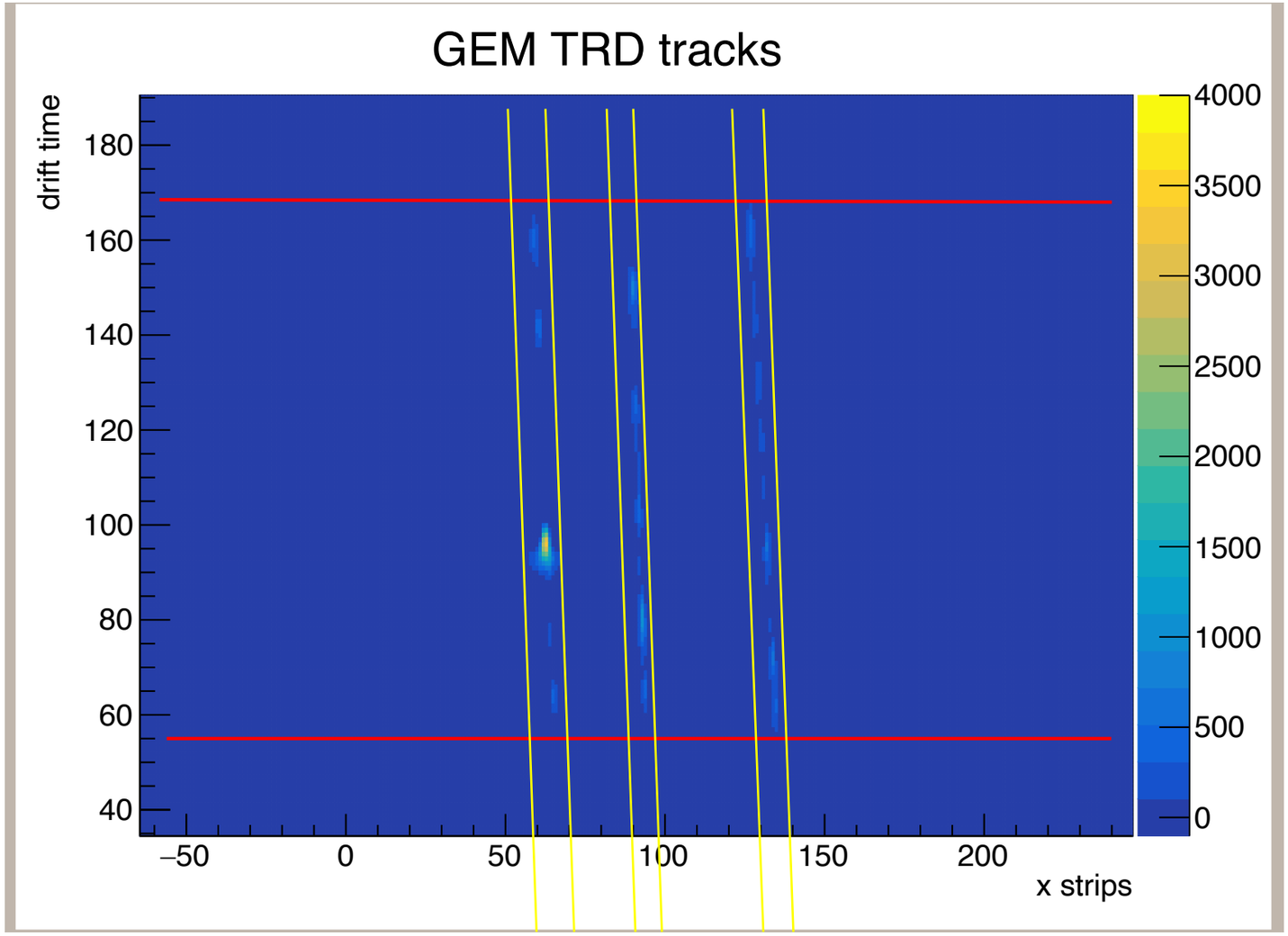
GEMTRD tracks

- In a real experiment, GEMTRD will have multiple tracks.
- So we also need an pattern recognition algorithm in the FPGA
- As well as track fitting.



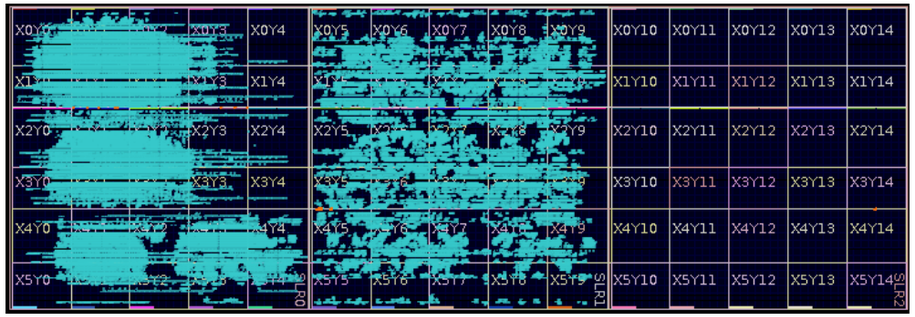
GEMTRD tracks

- In a real experiment, GEMTRD will have multiple tracks.
- So we also need an pattern recognition algorithm in the FPGA
- As well as track fitting.



GEMTRD FPGA processing test board

Jefferson Lab
Thomas Jefferson National Accelerator Facility



DNN PID module

Pattern recognition

DNN track fit

mdm_1
MBDEBUG_0
Debug_SYS_Rst
MicroBlaze Debug Module (MDM)

rst_clk_wiz_0_100M
slowest_sync_clk
ext_reset_in
aux_reset_in
mb_debug_sys_rst
dcm_locked
mb_reset
bus_struct_reset[0:0]
peripheral_reset[0:0]
interconnect_aresetn[0:0]
peripheral_aresetn[0:0]
Processor System Reset

microblaze_0_local_memory
DLMB
ILMB
LMB_Clk
SYS_Rst

clk_wiz_0
CLK_IN1_D
reset
clk_out1
locked
Clocking Wizard

microblaze_0
INTERRUPT
DEBUG
Clk
Reset
MicroBlaze
DLMB
ILMB
M_AXI_DP
MicroBlaze

microblaze_0_axi_periph
S00_AXI
ACLK
ARESETN
S00_ACLK
S00_ARESETN
M00_ACLK
M00_ARESETN
M01_ACLK
M01_ARESETN
M02_ACLK
M02_ARESETN
M03_ACLK
M03_ARESETN
M04_ACLK
M04_ARESETN
M00_AXI
M01_AXI
M02_AXI
M03_AXI
M04_AXI
AXI Interconnect

trd_fit_0
s_axi_AXILiteS
ap_clk
ap_rst_n
interrupt
Trd_fit (Pre-Production)

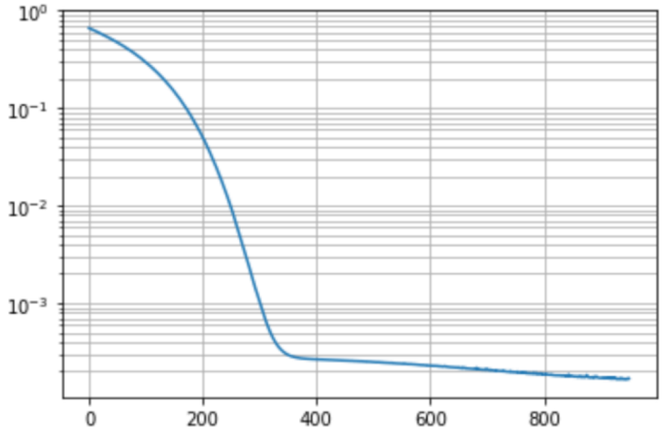
trd_trk_0
s_axi_AXILiteS
ap_clk
ap_rst_n
interrupt
Trd_trk (Pre-Production)

trdann_0
s_axi_AXILiteS
ap_clk
ap_rst_n
interrupt
Trdann (Pre-Production)

axi_gpio_0
S_AXI
s_axi_aclk
s_axi_aresetn
GPIO
GPIO2
AXI GPIO

axi_uartlite_0
S_AXI
s_axi_aclk
s_axi_aresetn
UART
interrupt
AXI Uartlite

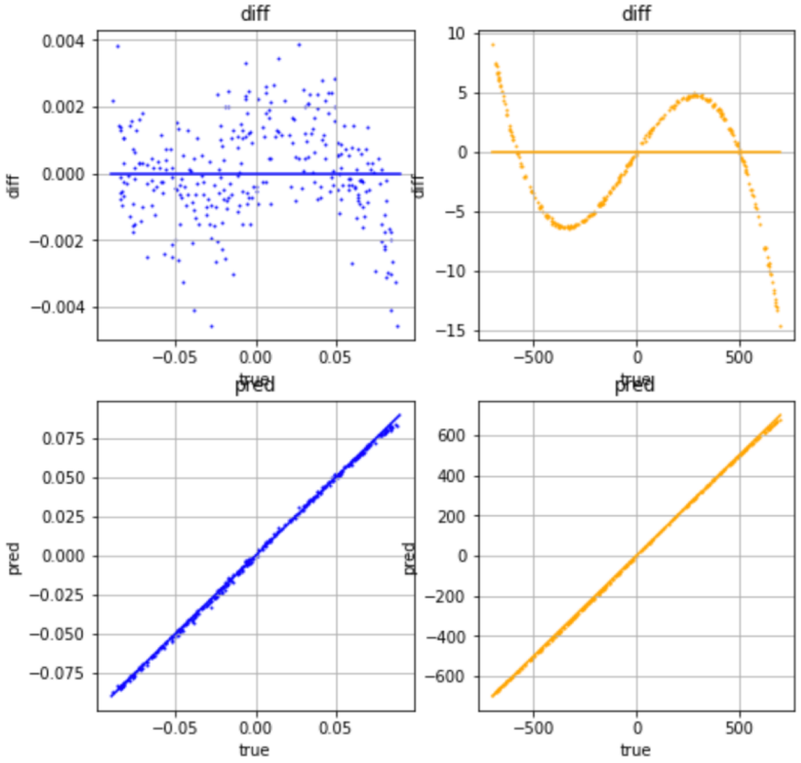
DNN for GEMTRD Track fit



- Tested solutions based on DNN and LSTM for track fitting.
- LSTM shows better performance but is not yet fully supported in HLS4ML.
- We are currently using a conventional clustering and pattern recognition algorithms. They work slowly and the execution time depends on the number of hits.
- We are currently working on an alternative pattern recognition solution based on a Graph Neural Network (GNN).

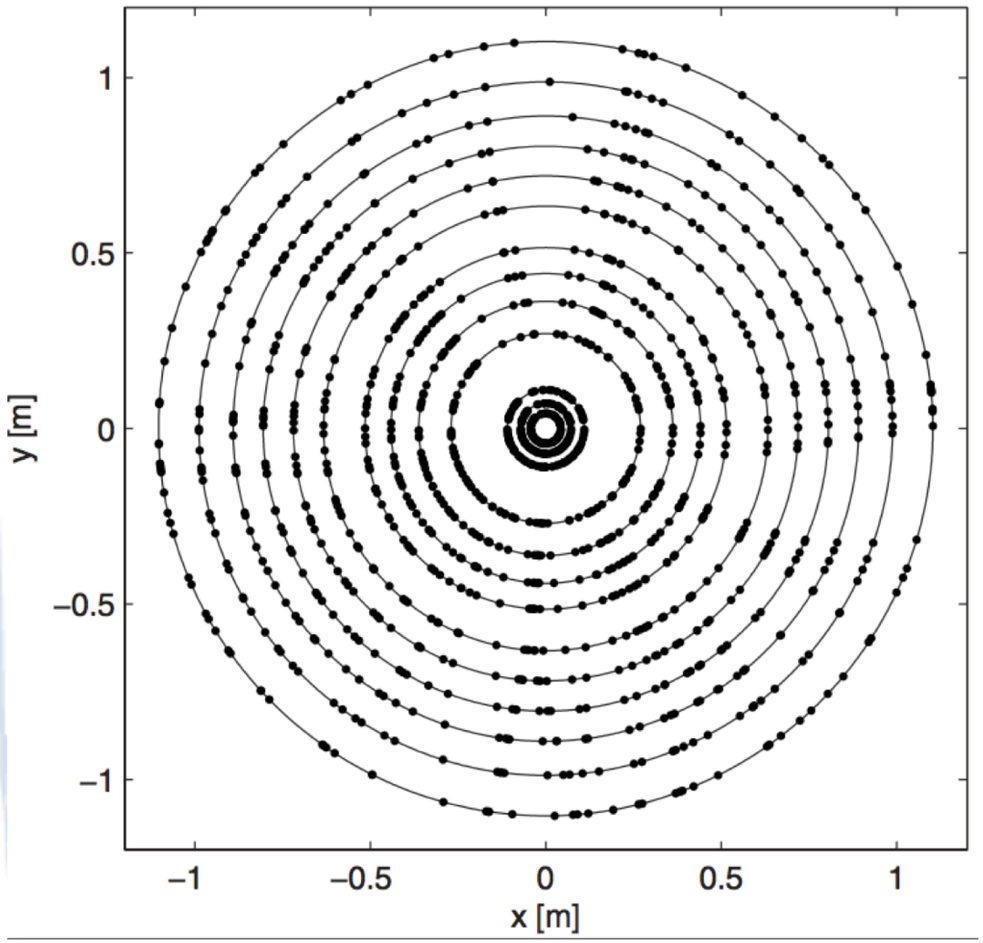
```
================================================
== Performance Estimates
================================================
+ Timing (ns):
    * Summary:
    +--------+-------+----------+------------+
    |  Clock | Target| Estimated| Uncertainty|
    +--------+-------+----------+------------+
    |ap_clk  |   5.00|     4.350|        0.62|
    +--------+-------+----------+------------+

+ Latency (clock cycles):
    * Summary:
    +-----+-----+-----+-----+----------+
    |  Latency  |  Interval | Pipeline |
    | min | max | min | max |   Type   |
    +-----+-----+-----+-----+----------+
    |   18|   18|    4|    4| function |
    +-----+-----+-----+-----+----------+
```

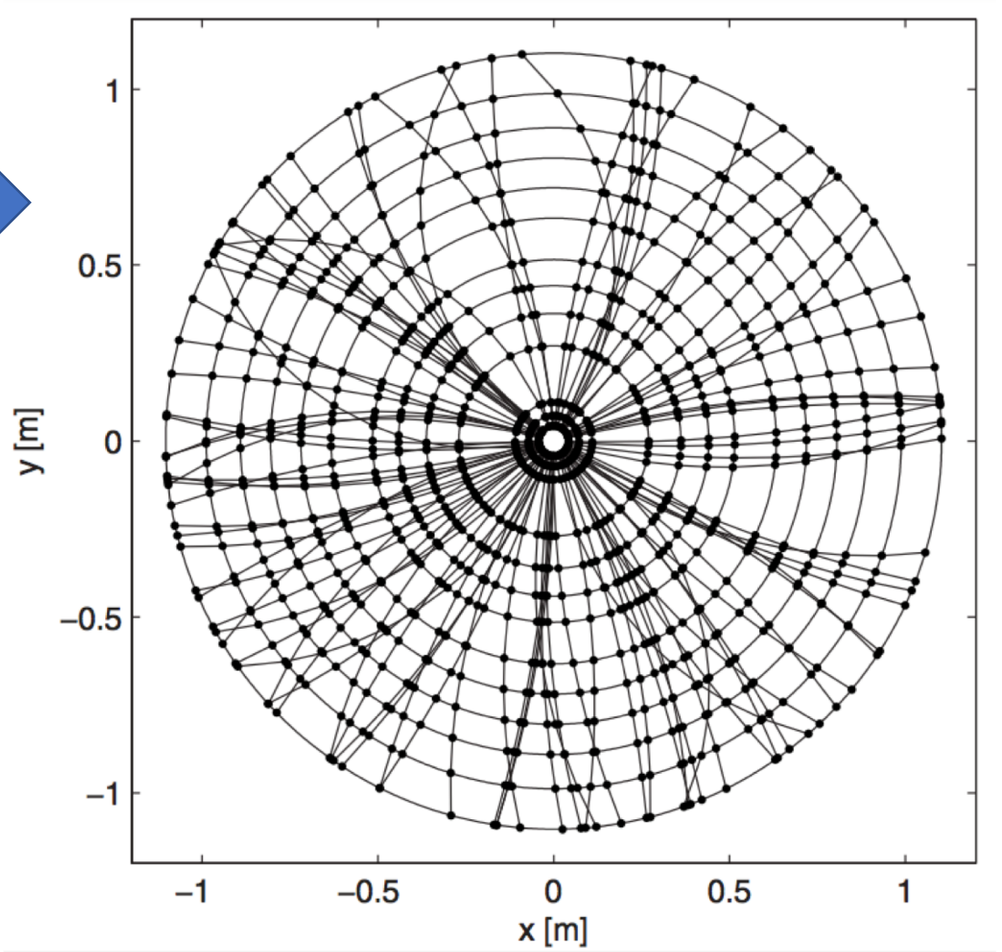
Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	6	-
FIFO	-	-	-	-	-
Instance	48	1418	26905	182279	-
Memory	-	-	-	-	-
Multiplexer	-	-	-	81	-
Register	-	-	4239	-	-
Total	48	1418	31144	182366	0
Available SLR	1440	2280	788160	394080	320
Utilization SLR (%)	3	62	3	46	0
Available	4320	6840	2364480	1182240	960
Utilization (%)	1	20	1	15	0



Track reconstruction example

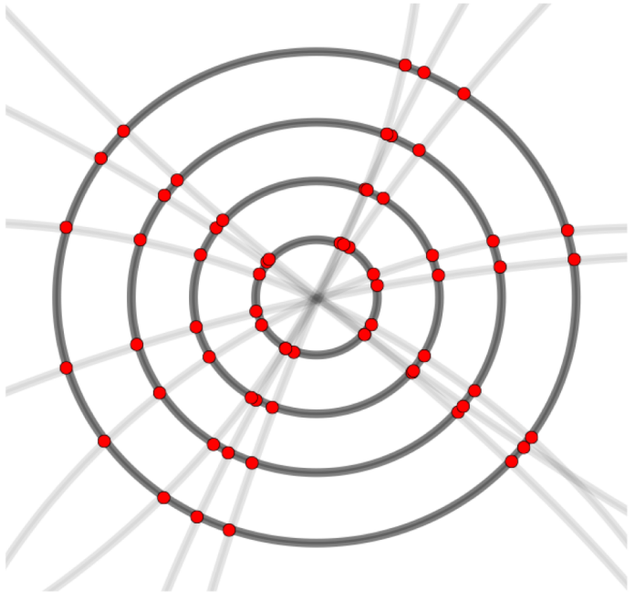
JAVIER DUARTE, OCTOBER 21, 2020 IRIS-HEP MEETING

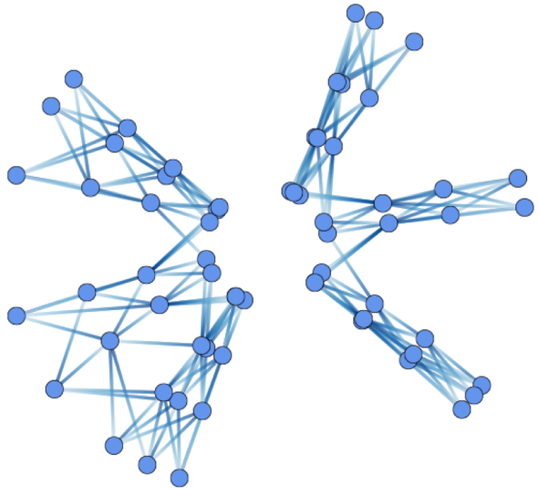




Graph construction.

- 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.



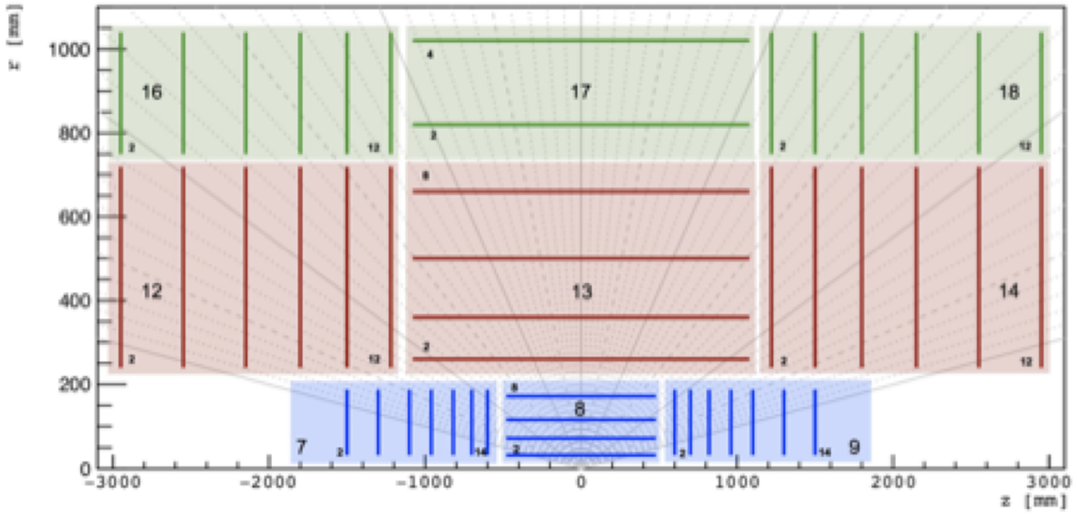


Javier Duarte arXiv:2012.01249v2 [hep-ph] 7 Dec 2020

Existing GNN tracking projects

Jefferson Lab
Thomas Jefferson National Accelerator Facility

- **TrackML Dataset**
  Public dataset hosted on Kaggle for particle tracking:
  https://www.kaggle.com/c/trackml-particle-identification



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

- *A working GNN implementation available here:*
  - https://github.com/vesal-rm/hls4ml/tree/graph_pipeline/example-prjs/graph/gnn_simple/
  - Vesal Razavimaleki , IRIS-HEP Fellows Presentation September 28, 20

So we decided to start by evaluating a working solution.

GNN FPGA Synthesis

Here is the result of the synthesis. :

General Information

Date: Sun May 15 20:48:02 2022
Version: 2019.1 (Build 2552052 on Fri May 24 15:28:33 MDT 2019)
Project: gnn2pl_prj
Solution: solution_rf1
Product family: virtexuplus
Target device: xcvu9p-flga2104-2L-e

Performance Estimates

Timing (ns)

Summary

Clock	Target	Estimated	Uncertainty
ap_clk	5.00	4.864	0.62

Latency (clock cycles)

Summary

Latency		Interval		
min	max	min	max	Type
1006	1006	1006	1006	none

Detail

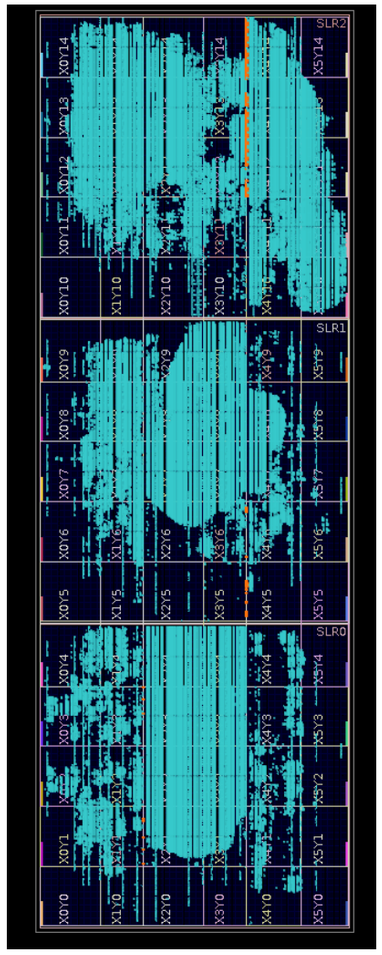
Instance

Loop

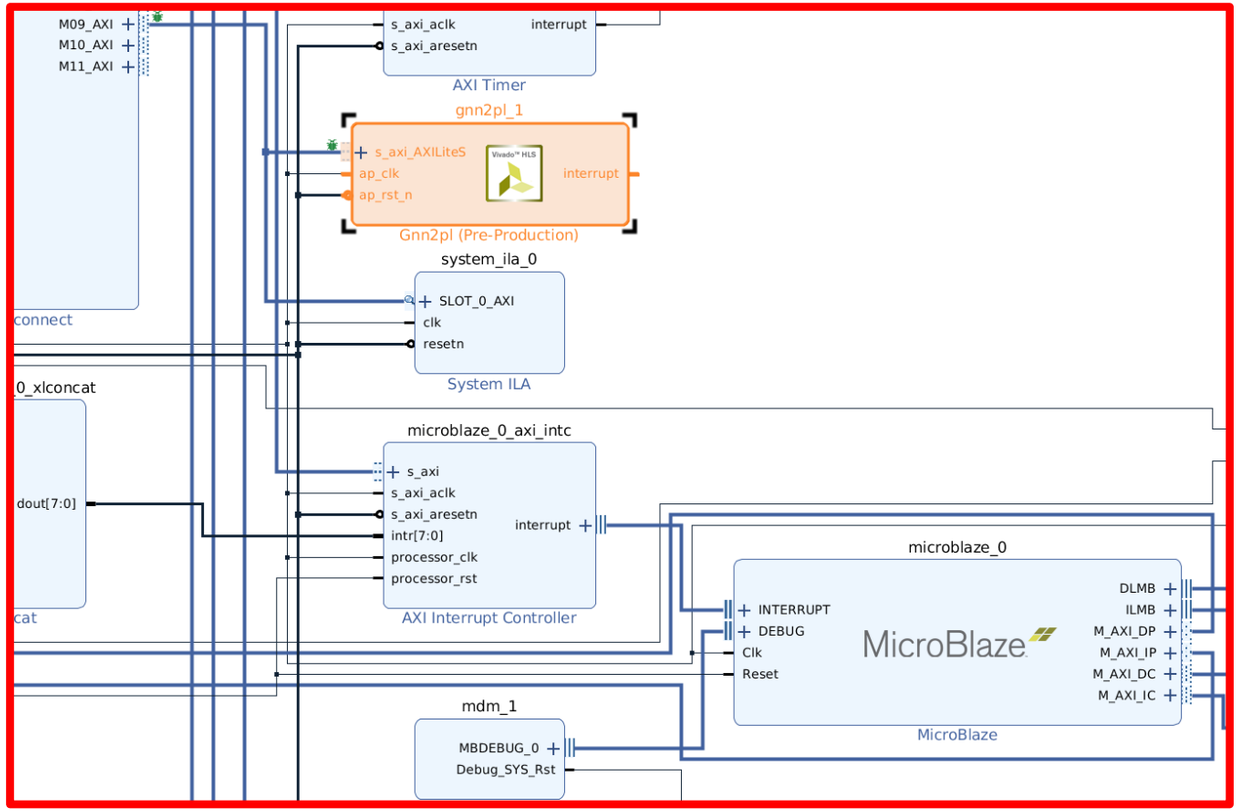
Utilization Estimates

Summary

Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	2	-
FIFO	-	-	-	-	-
Instance	148	4654	552257	363876	-
Memory	48	-	0	0	0
Multiplexer	-	-	-	4245	-
Register	-	-	22339	-	-
Total	196	4654	574596	368123	0
Available	4320	6840	2364480	1182240	960
Available SLR	1440	2280	788160	394080	320
Utilization (%)	4	68	24	31	0
Utilization SLR (%)	13	204	72	93	0



- 122 nodes x 3 features
- 148 edges x 4 features
- ~ 5 µs latency
- Still working on the interface

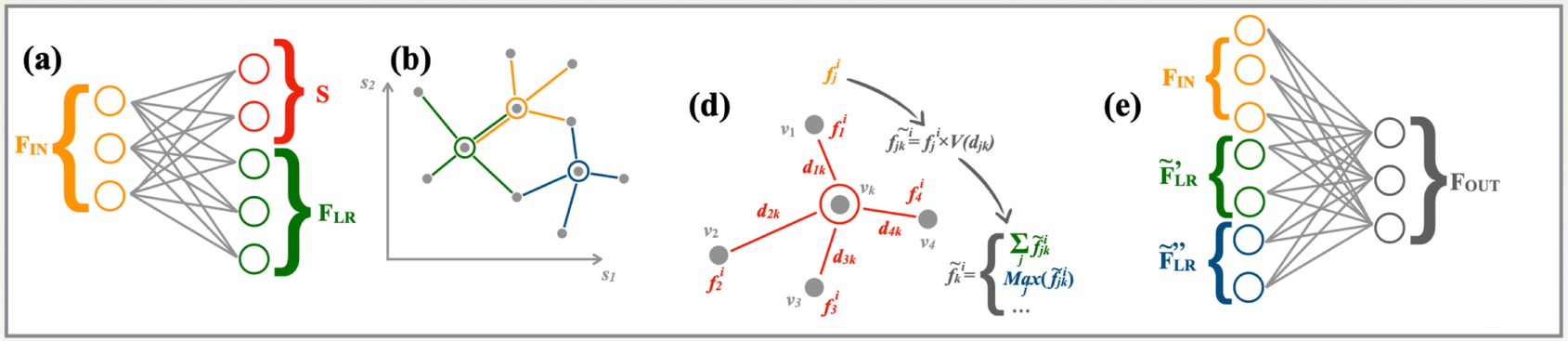


GNN for tracking and high-granularity calorimeters

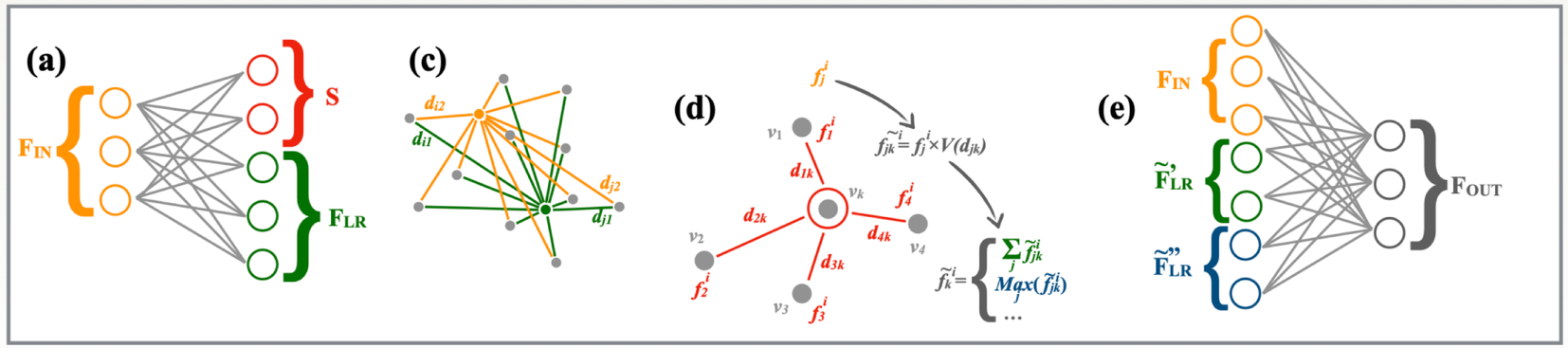
Jefferson Lab
Thomas Jefferson National Accelerator Facility

"Learning representations of irregular particle-detector geometry with distance-weighted graph networks"
arXiv:1902.07987v2 [physics.data-an] 24 Jul 2019

- GravNet



- GarNet



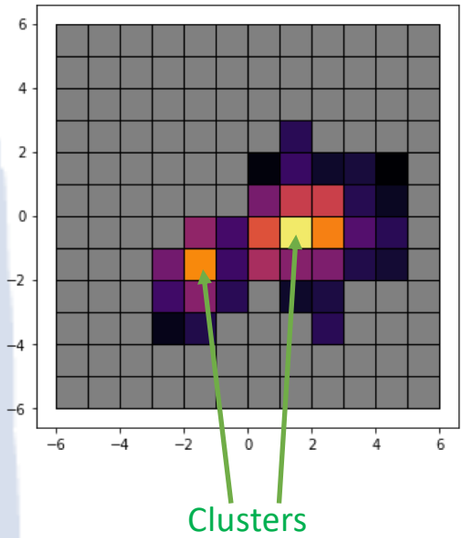
S.R. Qasim, J.K, Y. Iiyama, M Pierini arXiv:1902.07987, EPJC

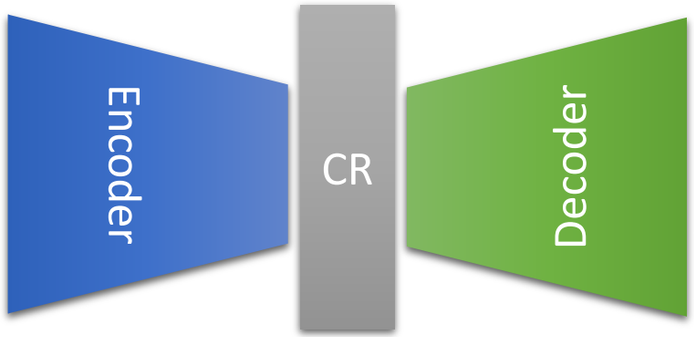
Calorimeter parameters reconstruction

Jefferson Lab
Thomas Jefferson National Accelerator Facility

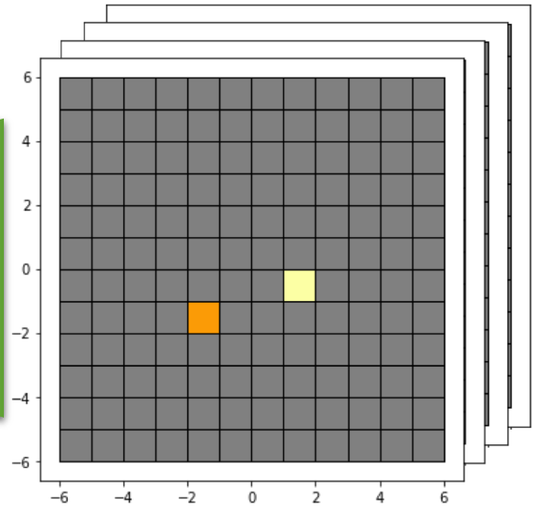
By Dmitry Romanov

Modules deposits





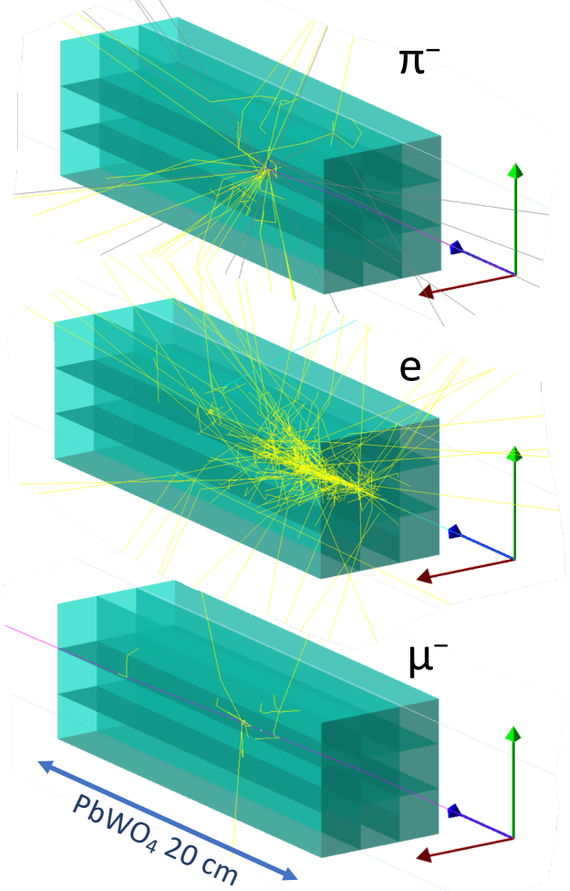
Convolutional variational autoencoder



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

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

Geant 4 simulation

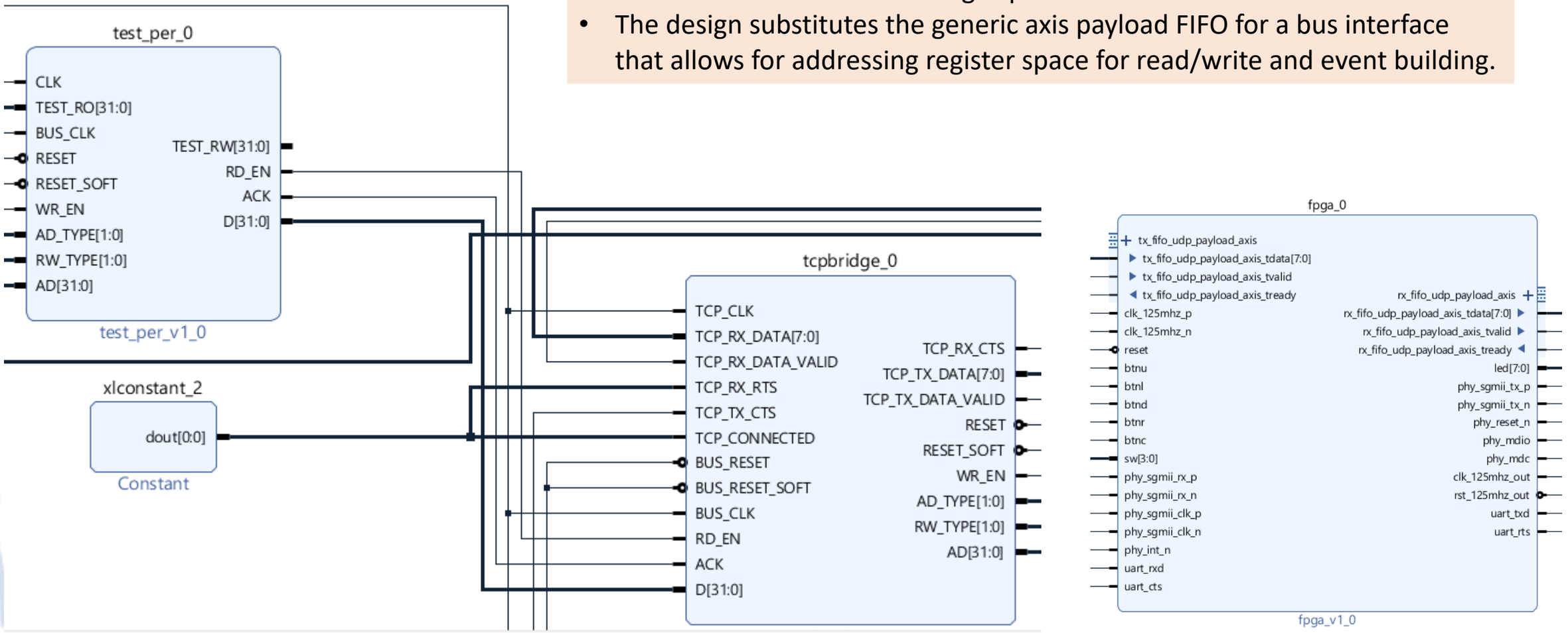


Examples of events with e and π^- showers and μ^- passing through.

Developing ethernet interface

By Cody Dickover

- Currently we using Microblaze setup for tests.
- For the beam test we need high speed interface to FADC.
- The design substitutes the generic axis payload FIFO for a bus interface that allows for addressing register space for read/write and event building.



GEMTRD proposal for the GlueX experiment

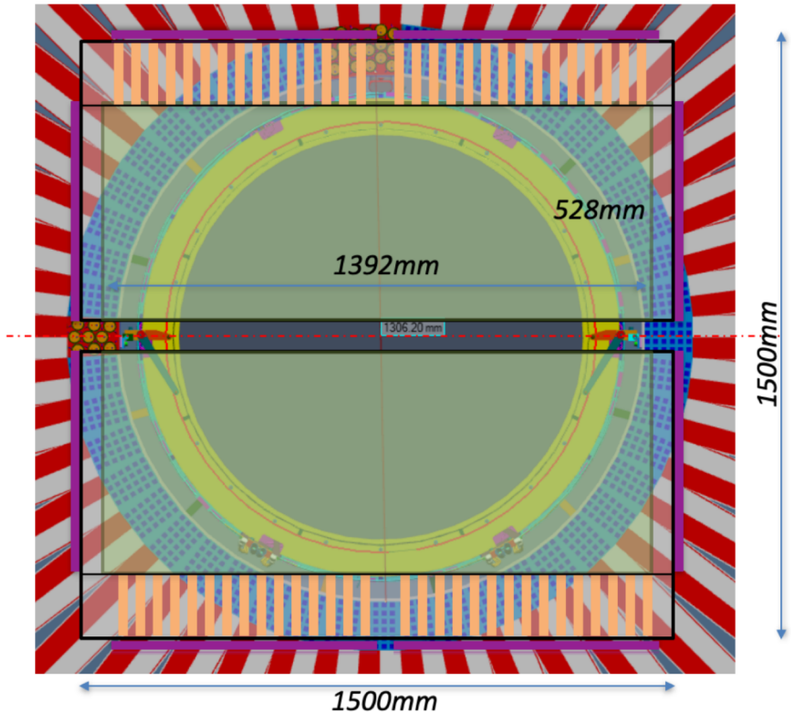


FIG. 2: Front view of the GEM-TRD detector placed at the face of the solenoid magnet.

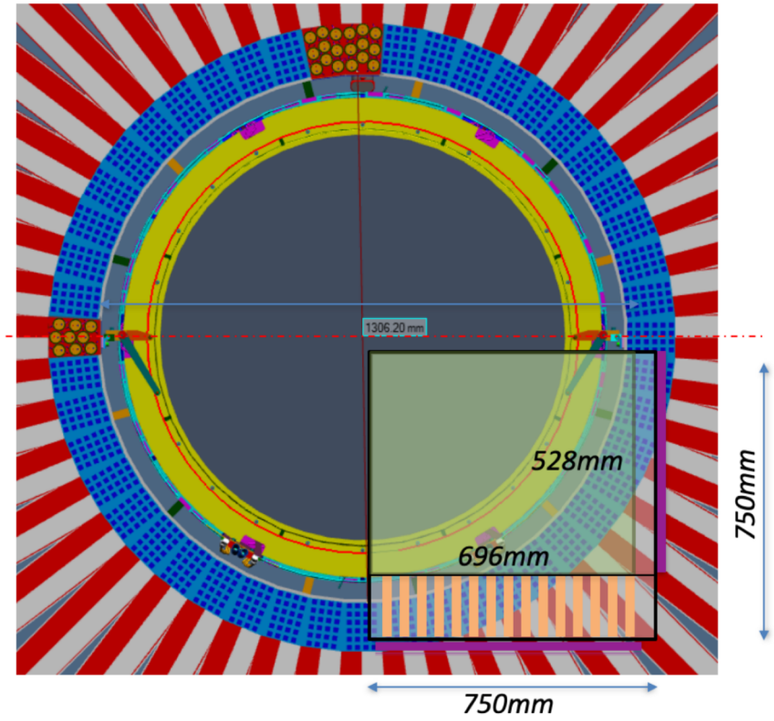


FIG. 16: Front view of the GEM-TRD large-scale prototype that has 696×528 mm^2 sensitive area.

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 on the order of 100 nanoseconds.*
- *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.*

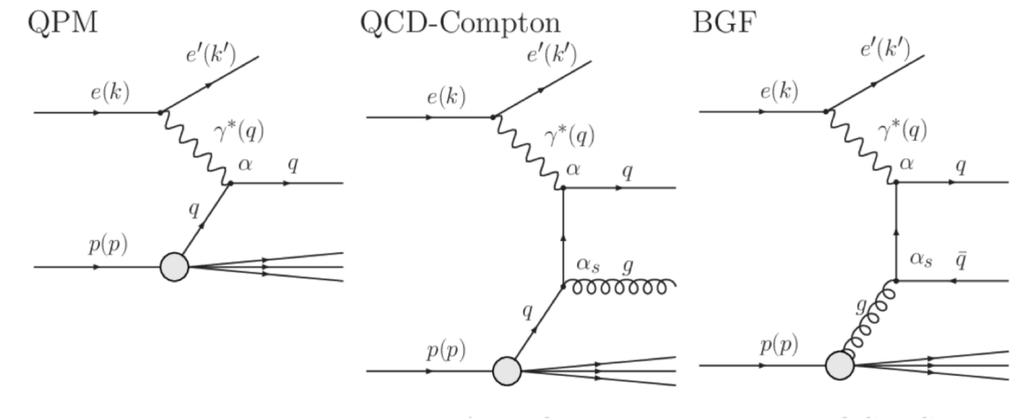
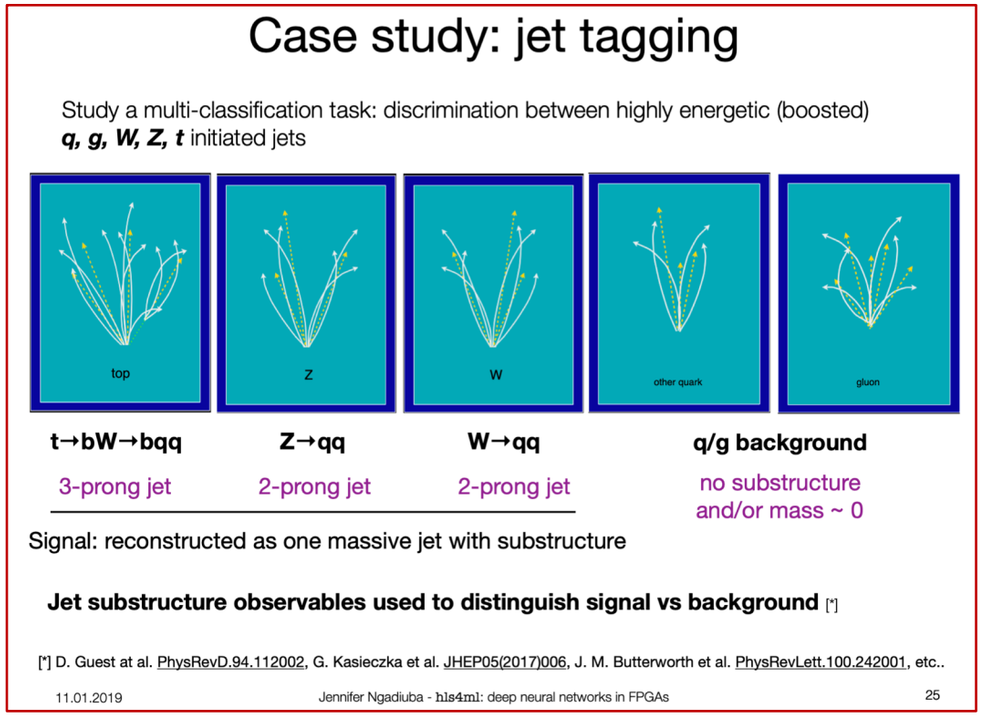


Figure 2.1: Feynman diagrams of the Quark Parton Model, QCD-Compton and Boson Gluon Fusion processes in NC DIS.

Published in 2007

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

T. Gosau



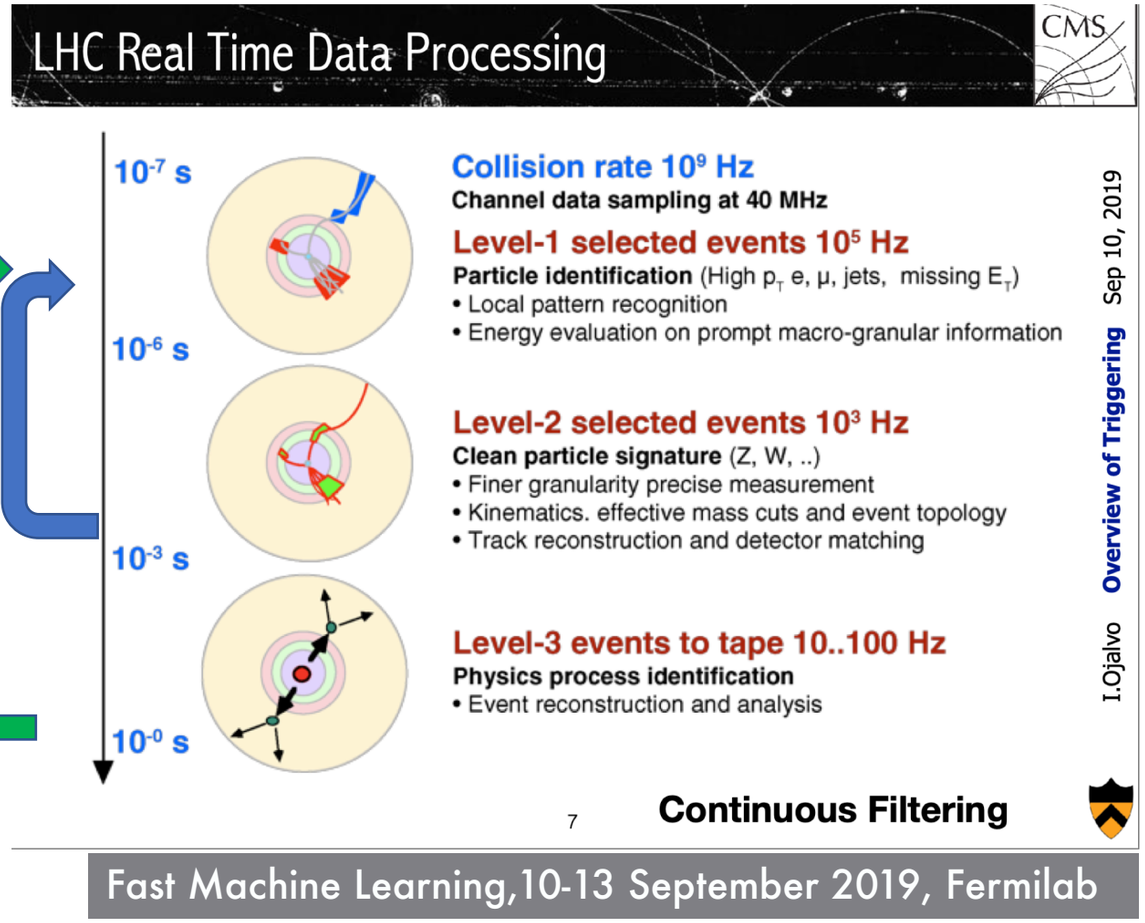
Backup

Motivation

- *The growing computational power of modern FPGA boards allows us to add more sophisticated algorithms for real time data processing.*
- *Many tasks could be solved using modern Machine Learning (ML) algorithms which are naturally suited for FPGA architectures.*

Level 1 works with Regional and sub-detector Trigger primitives

Using ML on FPGA many tasks from Level 2 and/or Level 3 can be performed at Level 1



ML in physics

Jefferson Lab
Thomas Jefferson National Accelerator Facility

- Machine learning methods are widely used and have proven to be very powerful in particle physics.
- Although the methods of machine learning and artificial intelligence are developed by many groups and have a lot in common, nevertheless, the hardware used and performance is different:

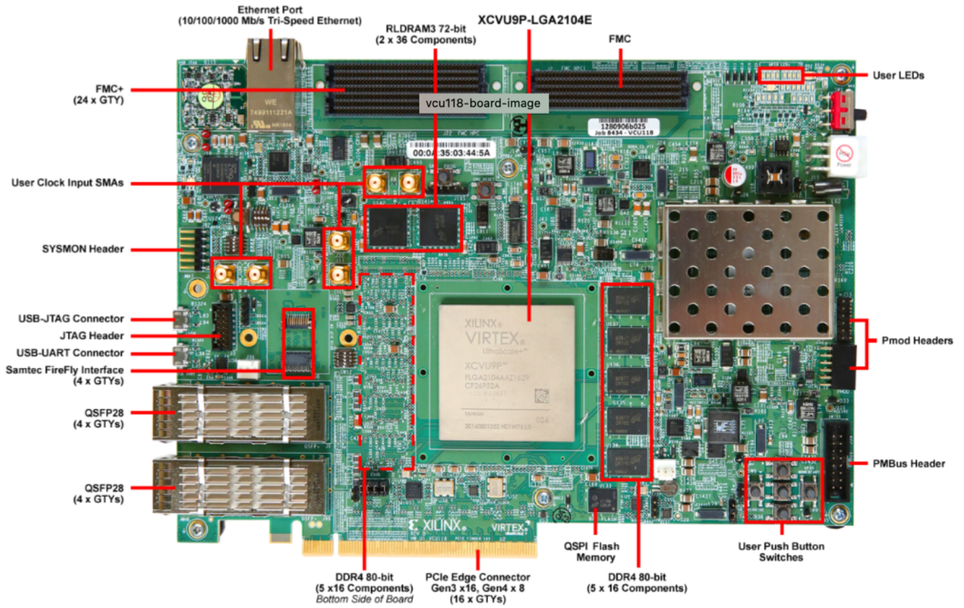
  1) CPU only (farm)
  2) CPU and GPU accelerator
  3) CPU and FPGA accelerator
  4) pure FPGA

- While the large numerical processing capability of GPUs is attractive, these technologies are optimized for high throughput, not low latency.
- FPGA-based trigger and data acquisition systems have extremely low, sub-microsecond latency requirements that are unique to particle physics.
- Definitely FPGA can work on a computer farm as an ML accelerator, but the internal FPGA performance will be degraded due to slow I/O through the computer and the PCIe bus. Not to mention the latency, which will increase by 2-3 orders of magnitude.
- Therefore, the most effective would be the use of ML-FPGA directly between the front-end stream and a computer farm, on which it is already more efficient to use the CPU and GPU for ML/AI.

FPGA test board for ML

- 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+™

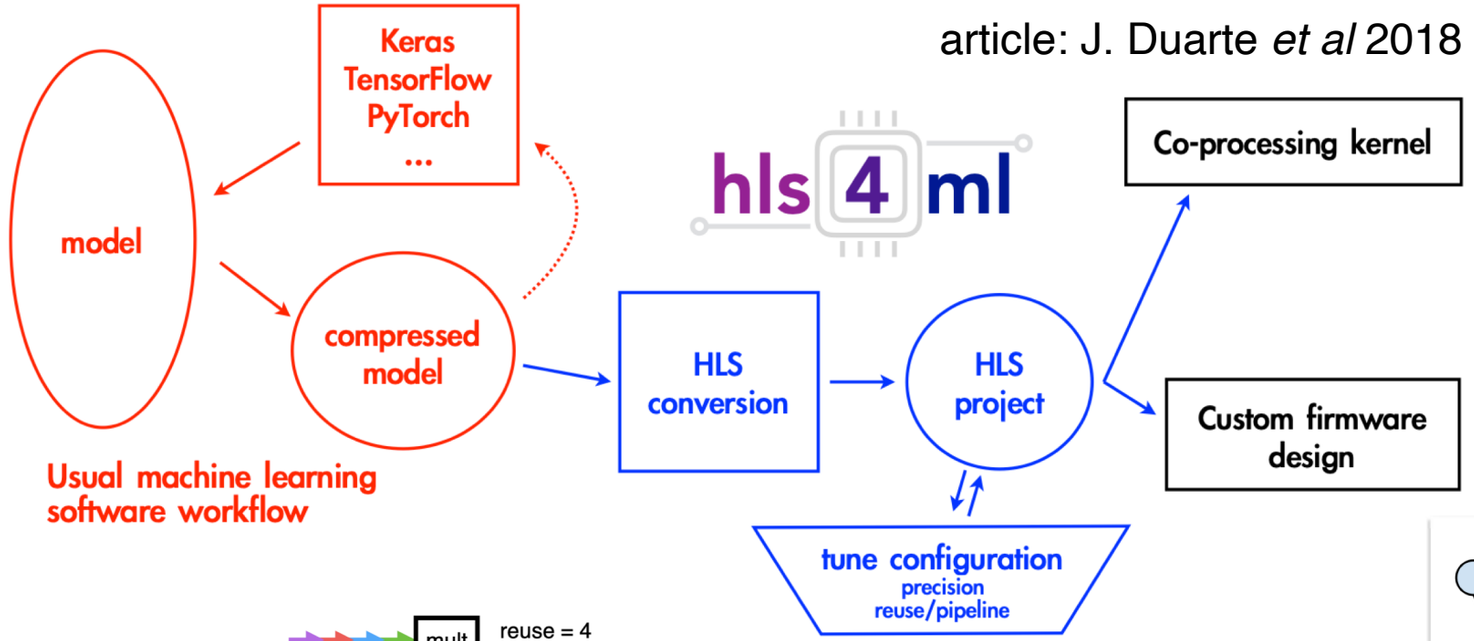


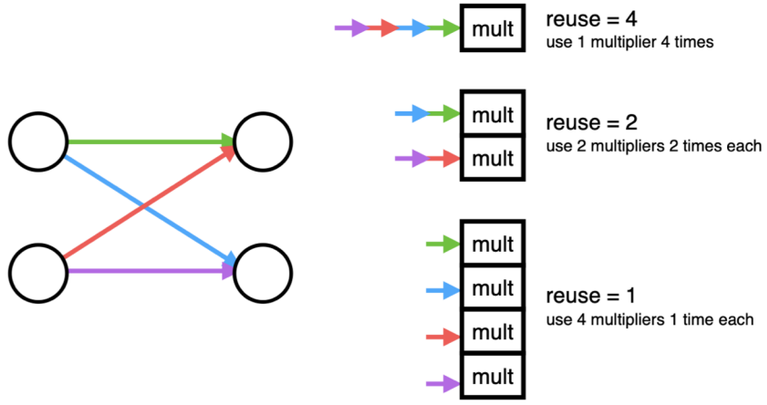
Optimization with hls4ml package

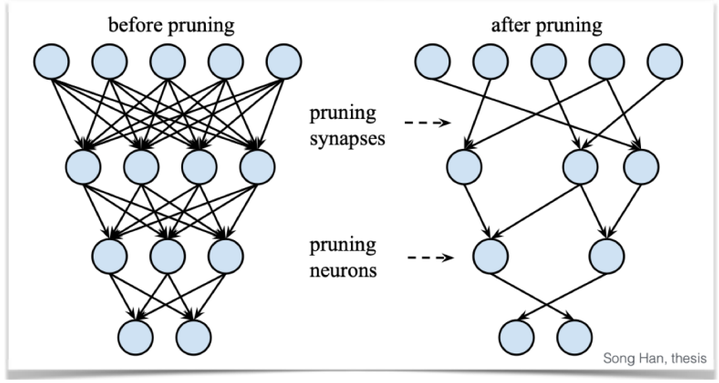
Jefferson Lab
Thomas Jefferson National Accelerator Facility

- 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







https://fastmachinelearning.org/hls4ml/

GEMTRD PID network optimization

Jefferson Lab
Thomas Jefferson National Accelerator Facility

Full size neural network, accuracy-optimized.

```
+ 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   |
    +-----+-----+-----+-----+----------+
    |   15|   15|    1|    1| function |
    +-----+-----+-----+-----+----------+
```

Latency = 75ns

Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	2	-	-	-
Expression	-	-	0	24	-
FIFO	-	-	-	-	-
Instance	19	692	3737	16446	-
Memory	2	-	0	0	-
Multiplexer	-	-	-	36	-
Register	-	-	1532	-	-
Total	21	694	5269	16506	0
Available SLR	1440	2280	788160	394080	320
Utilization SLR (%)	1	30	~0	4	0
Available	4320	6840	2364480	1182240	960
Utilization (%)	~0	10	~0	1	0

DSP utilization 10%

Size-optimized neural network

```
+ Timing (ns):
    * Summary:
    +--------+-------+----------+------------+
    |  Clock | Target| Estimated| Uncertainty|
    +--------+-------+----------+------------+
    |ap_clk  |   5.00|     3.883|        0.62|
    +--------+-------+----------+------------+

+ Latency (clock cycles):
    * Summary:
    +-----+-----+-----+-----+----------+
    |  Latency  |  Interval | Pipeline |
    | min | max | min | max |   Type   |
    +-----+-----+-----+-----+----------+
    |   17|   17|    3|    3| function |
    +-----+-----+-----+-----+----------+
```

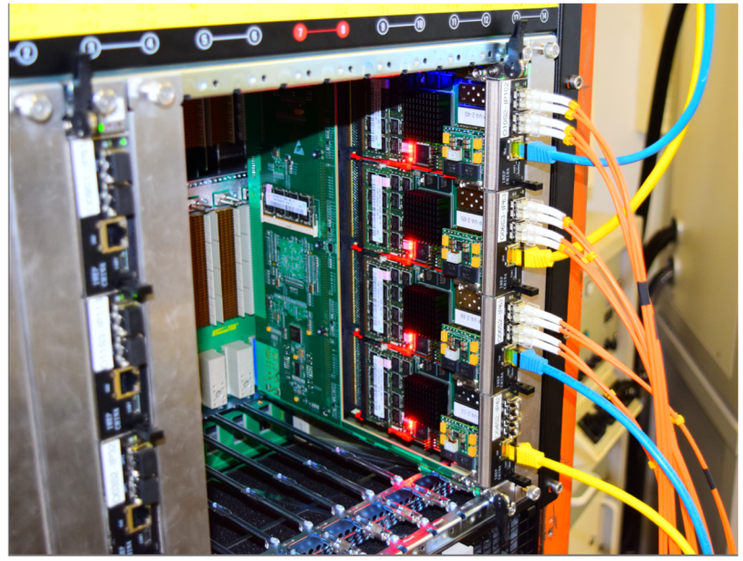
Latency = 85ns

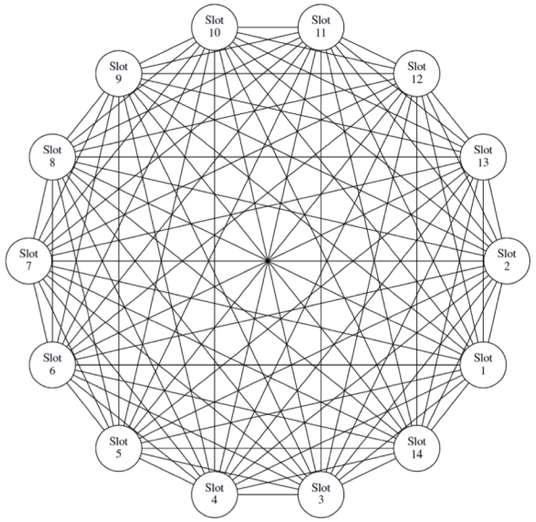
Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	2	-	-	-
Expression	-	-	0	24	-
FIFO	-	-	-	-	-
Instance	-	177	3132	10696	-
Memory	2	-	0	0	-
Multiplexer	-	-	-	81	-
Register	-	-	1423	-	-
Total	2	179	4555	10801	0
Available SLR	1440	2280	788160	394080	320
Utilization SLR (%)	~0	7	~0	2	0
Available	4320	6840	2364480	1182240	960
Utilization (%)	~0	2	~0	~0	0

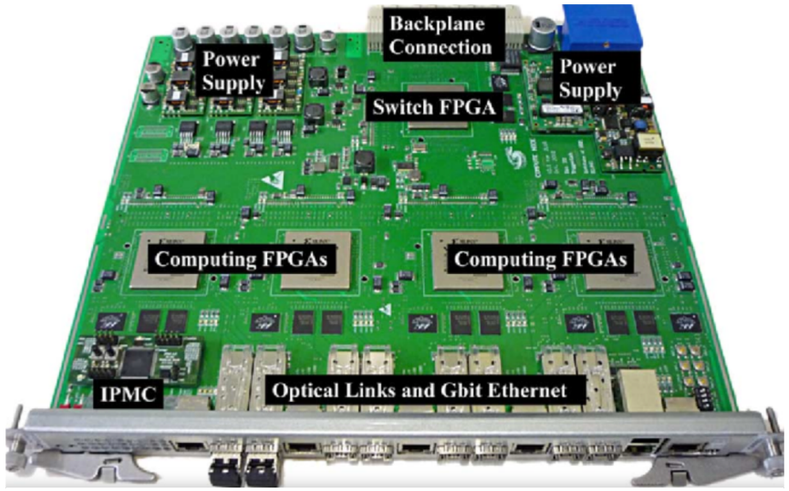
DSP utilization 2%

Compute Node (PXD,Belle II)

- The pixel detector of Belle II with its ~ 8 million channels will deliver data at rate of 22 Gbytes/s for a trigger rate of 30 kHz
- A hardware platform capable of processing this amount of data is the ATCA based Compute Node. (Advanced Telecommunications Computing Architecture).
- A single ATCA crate can host up to 14 boards interconnected via a full mesh backplane.
- Each AMC board is equipped with 4 Xilinx Virtex-5 FX70T FPGA.







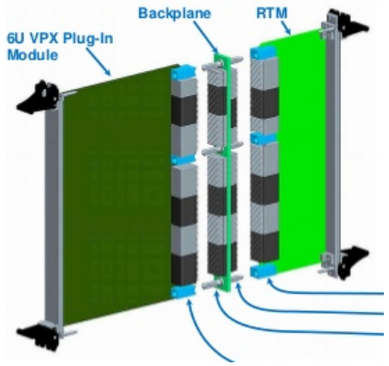
ADC based DAQ for PANDA STT

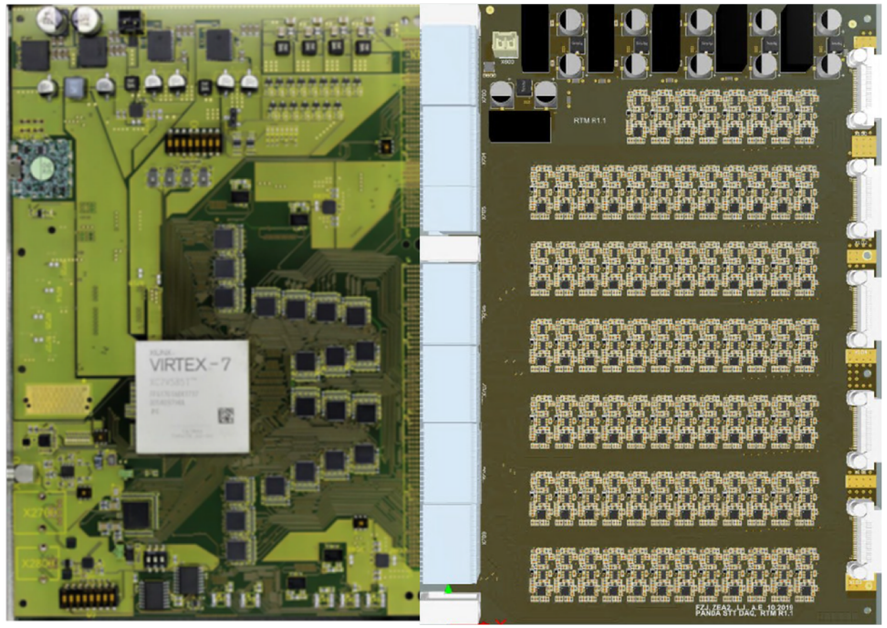
Jefferson Lab
Thomas Jefferson National Accelerator Facility

Level 0 Open VPX Crate

ADC based DAQ for PANDA STT (one of approaches):

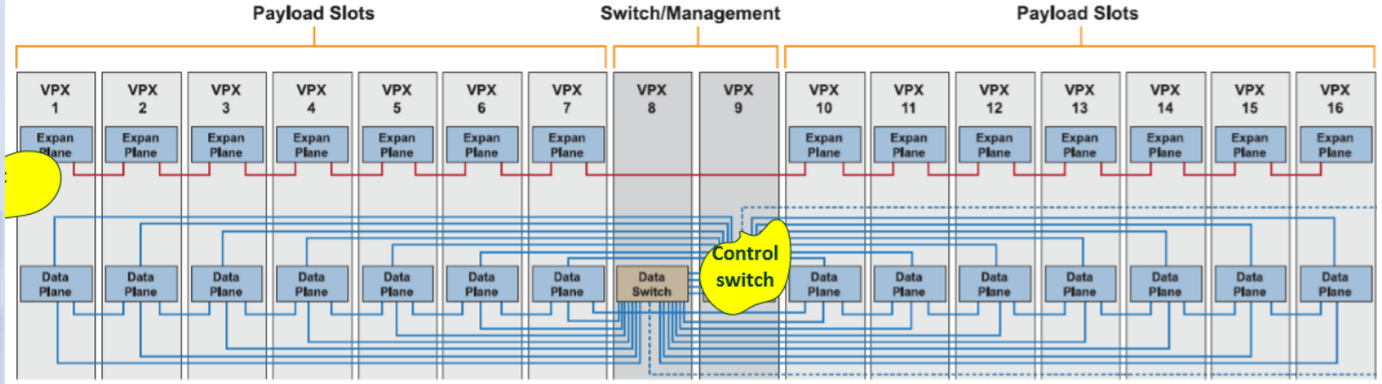
- 160 channels (shaping, sampling and processing) per payload slot, 14 payload slots+2 controllers;
- **totally 2200 channels per crate**;
- time sorted output data stream (arrival time, energy,...)
- noise rejection, pile up resolution, base line correction, ..





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

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



Powerful Backplane up to 670 GBs

- *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.*

L. Jokhovets, P Kulessa ..

JÜLICH Forschungszentrum

Unified hardware solution (ATCA or OpenVPX)