Jefferson Lab
Thomas Jefferson National Accelerator Facility

AI/ML on FPGA

Sergey Furletov
Jefferson Lab

JLab AI Town Hall

26 July 2021

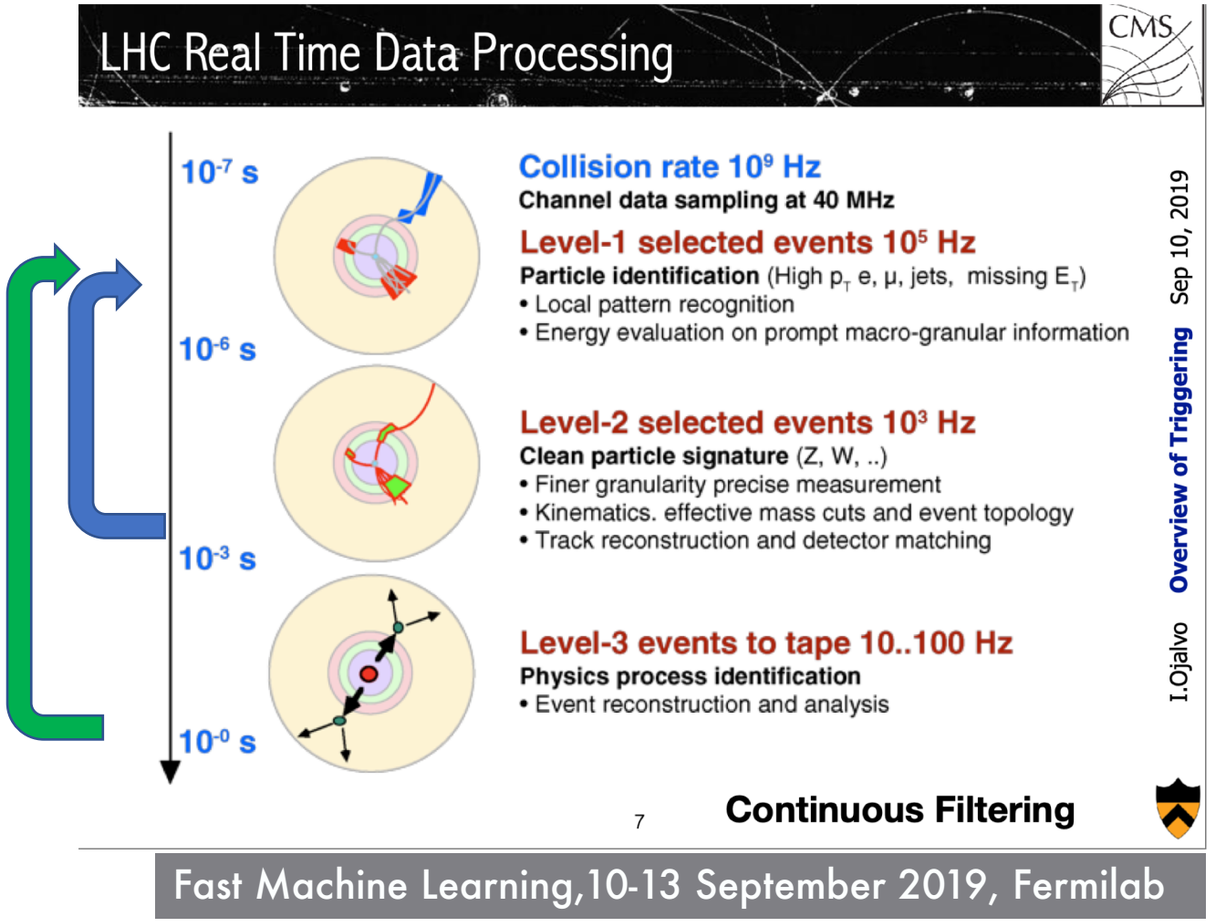
Motivation

- *Concepts of trigger-less readout and data streaming will produce large data volumes being read from the detectors.*
- *Many tasks could be solved using modern Machine Learning (ML) algorithms which are naturally suited for FPGA architectures.*
- *The growing computational power of modern FPGA boards allows us to add more sophisticated algorithms for real time data processing.*

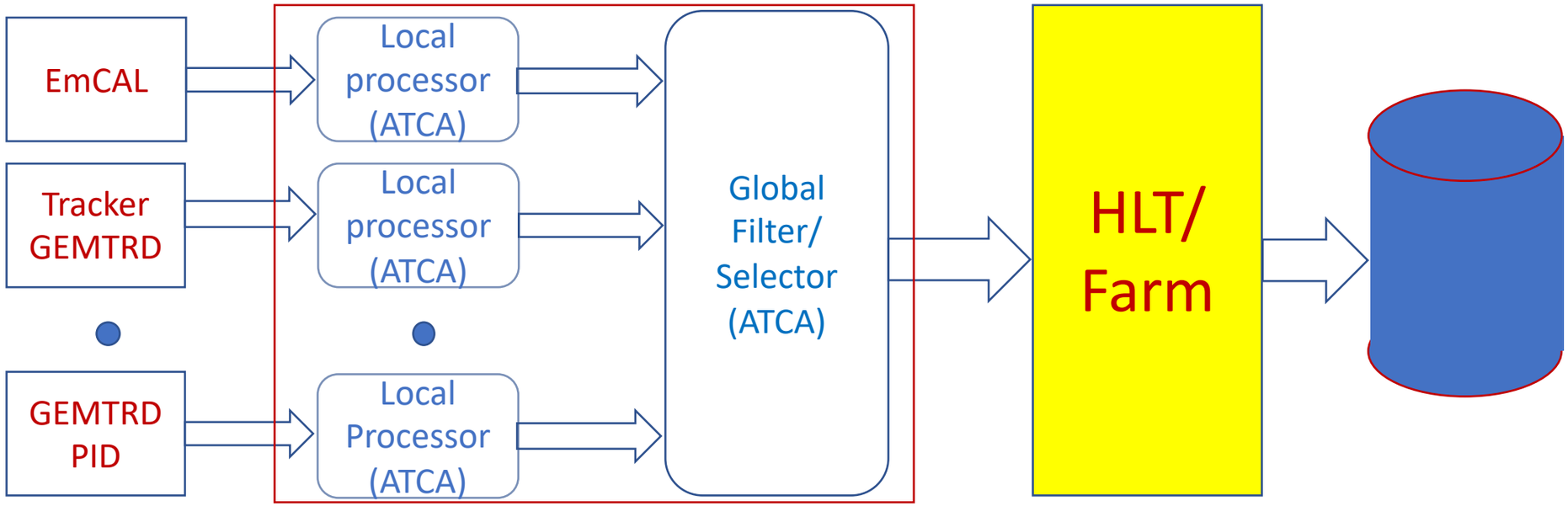
Level 1 works with Regional and sub-detector trigger primitives, typically uses custom hardware with ASICs or FPGAs (decision ~4 μs)

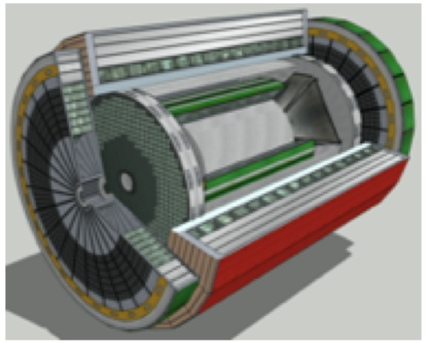
High Level Trigger (HLT), uses commercial CPUs to process the filtered data in software.
(decision ~100 000 μs)

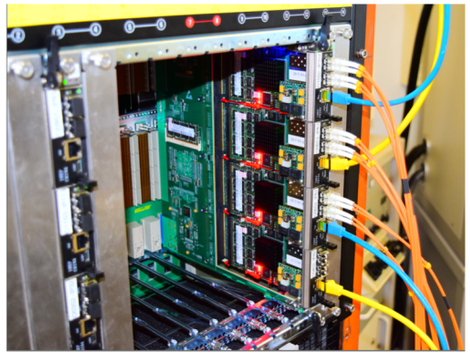
Using ML on FPGA many tasks from HLT can be performed at Level 1



Hall-D ML filter design test setup



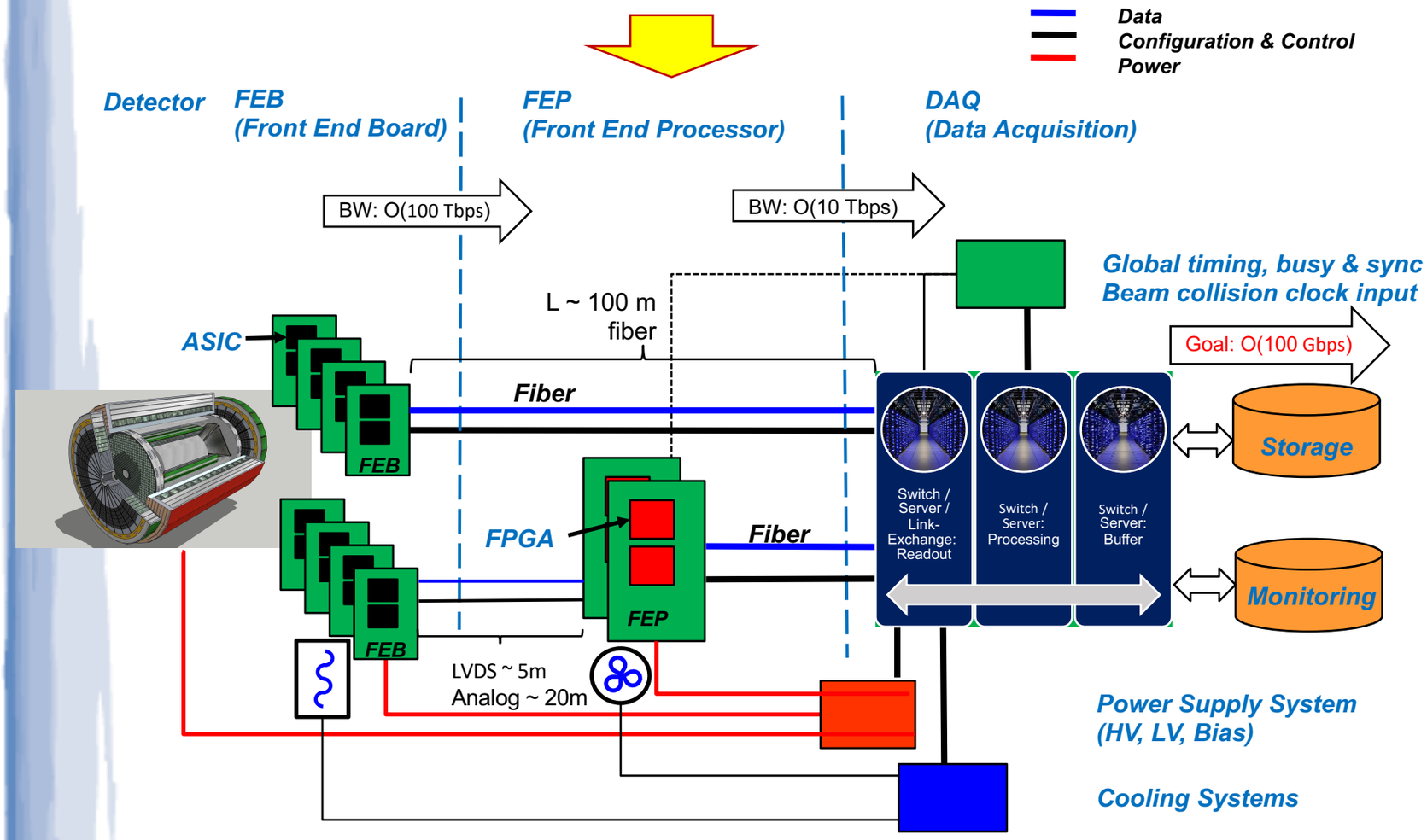




Team :

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

ML in EIC readout



- *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.*

ML FPGA Core for TRD PID

Jefferson Lab
Thomas Jefferson National Accelerator Facility

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

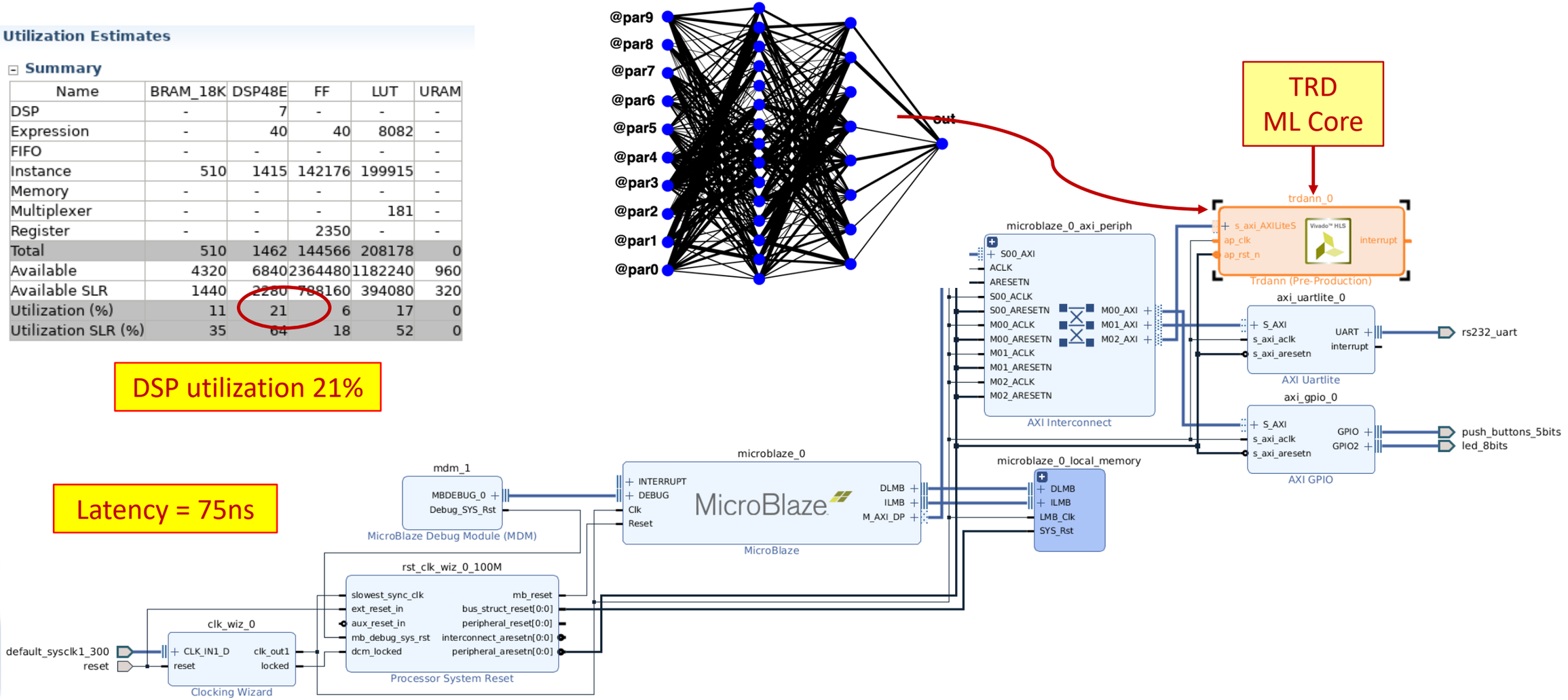
Utilization Estimates

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%

Latency = 75ns

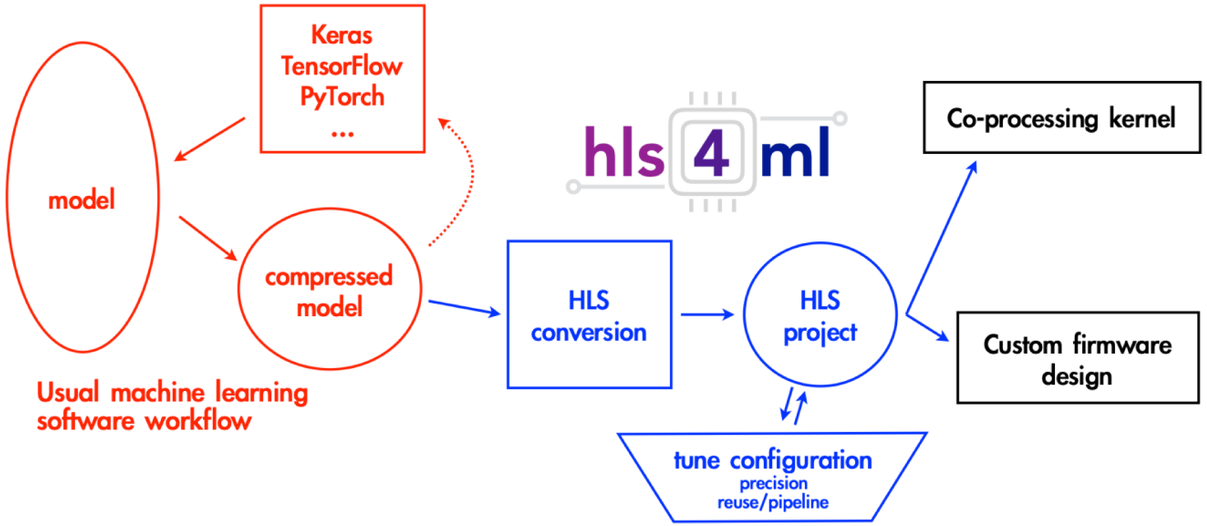


GEMTRD tracking with HLS4ML package

Jefferson Lab
Thomas Jefferson National Accelerator Facility

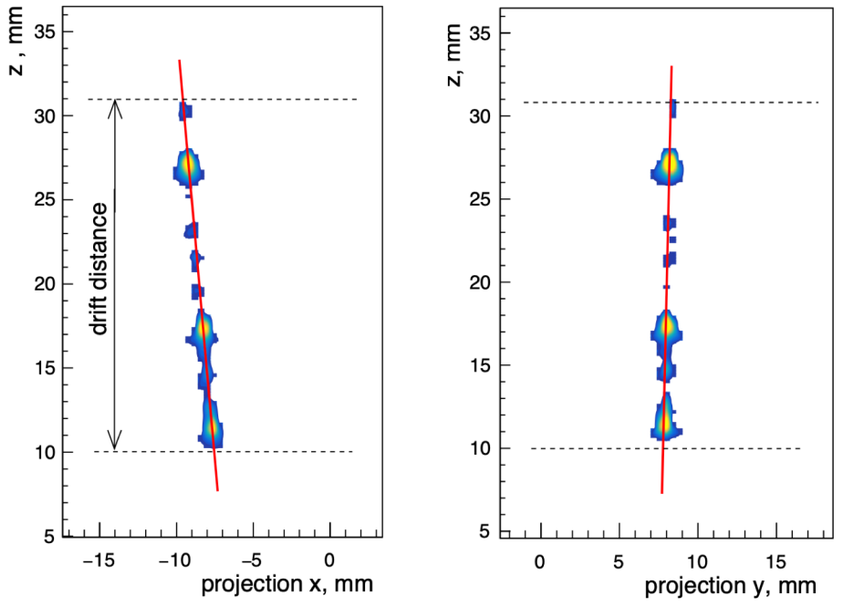
- A package hls4ml is developed based on High-Level Synthesis (HLS) to build machine learning models in FPGAs.

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

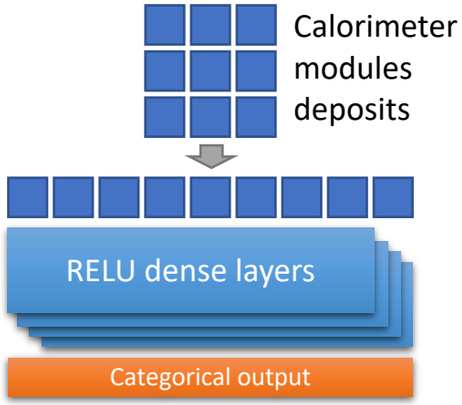


- *Clustering*
- *Pattern recognition*
- *ML Track fitting*

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



ML for Calorimeter e/pi separation

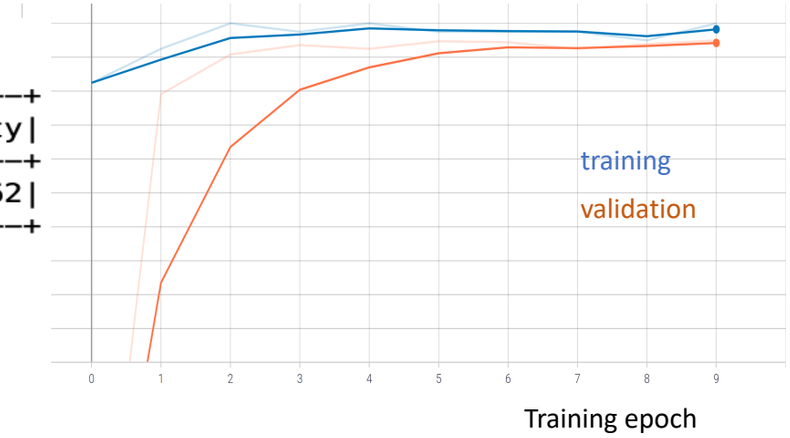


Classification	Last-layer activation	Loss function
single-label	softmax	categorical_crossentropy
multi-label (scores for candidates)	sigmoid	binary_crossentropy

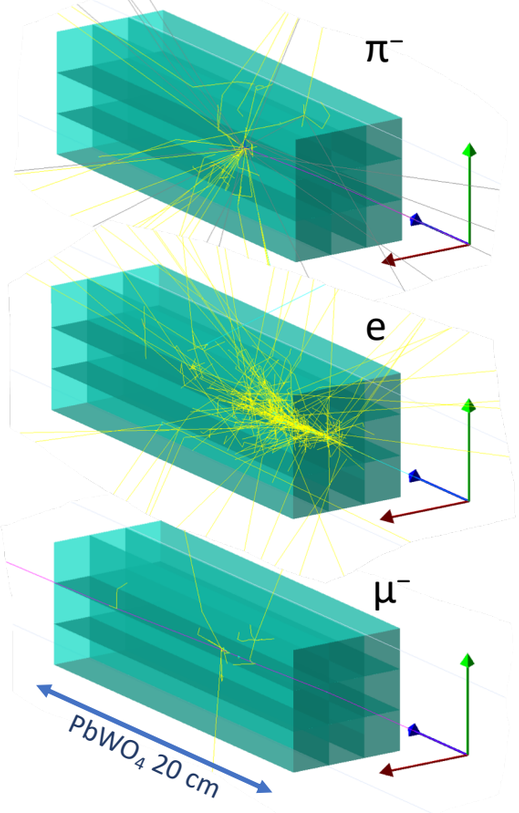
```
+ Timing (ns):
    * Summary:
    +--------+-------+----------+------------+
    |  Clock | Target| Estimated| Uncertainty|
    +--------+-------+----------+------------+
    |ap_clk  |   5.00|     3.883|        0.62|
    +--------+-------+----------+------------+

+ Latency (clock cycles):
    * Summary:
    +-----+-----+-----+-----+----------+
    |  Latency  |  Interval | Pipeline |
    | min | max | min | max |   Type   |
    +-----+-----+-----+-----+----------+
    |   12|   12|    1|    1| function |
    +-----+-----+-----+-----+----------+
```

Latency = 60ns



Geant 4 simulation



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

by D. Romanov

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.*
- *The unified design will make it easy to increase the processing power by adding ML algorithms:*
  - from the pass-through to trigger mode and finally to physics filter mode.

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

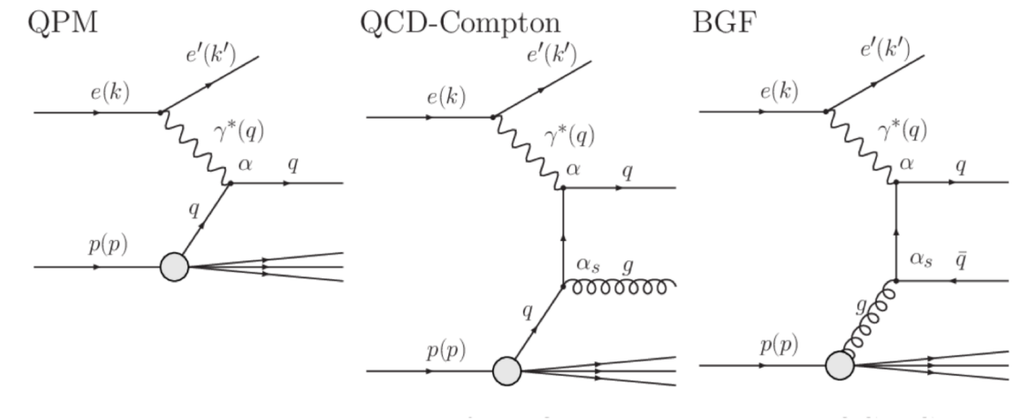


Figure 2.1: Feynman diagrams of the Quark Parton Model, QCD-Compton and Boson Gluon Fusion processes in NC DIS.

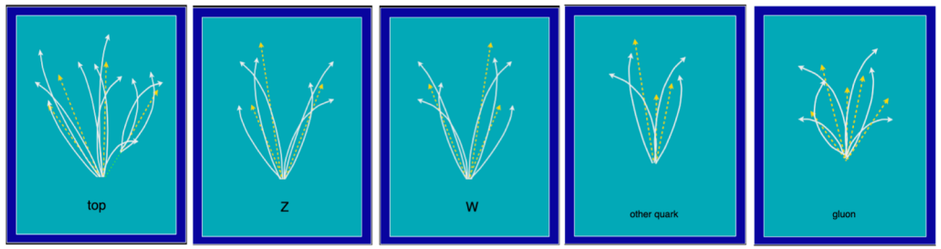
Published in 2007

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

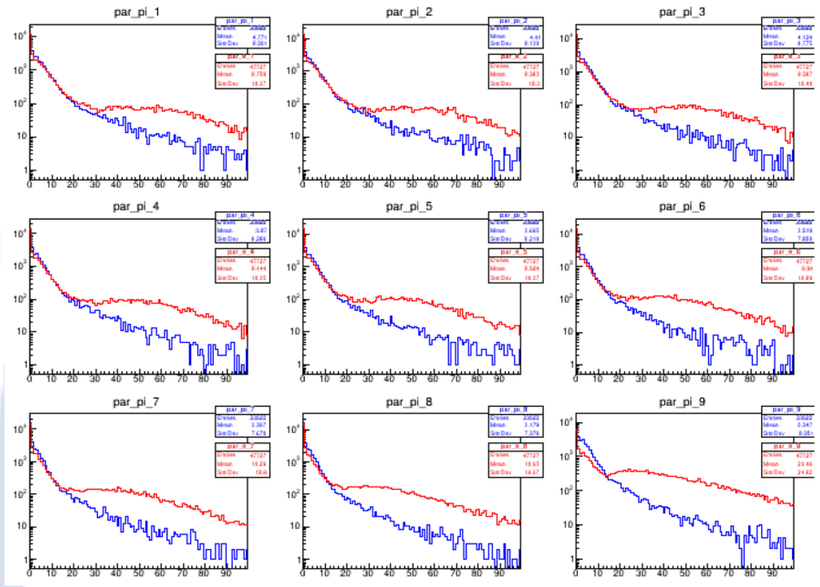
Jet substructure observables used to distinguish signal vs background [*]

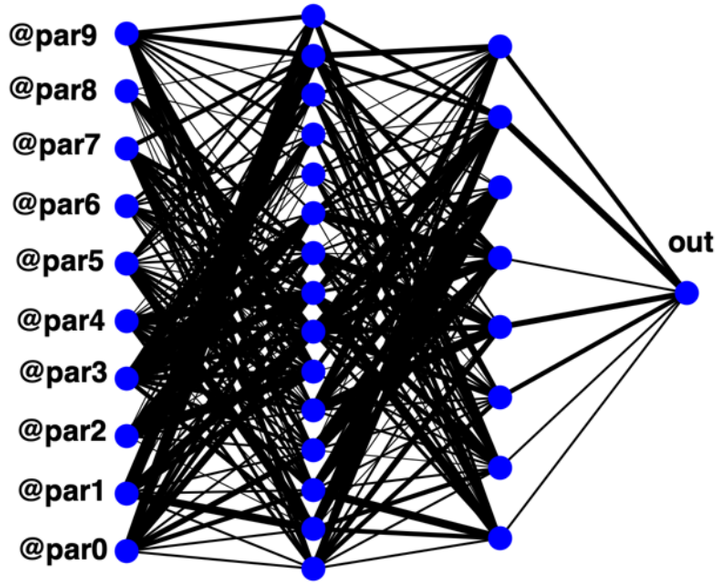
[*] D. Guest at al. PhysRevD.94.112002, G. Kasieczka et al. JHEP05(2017)006, J. M. Butterworth et al. PhysRevLett.100.242001, etc..

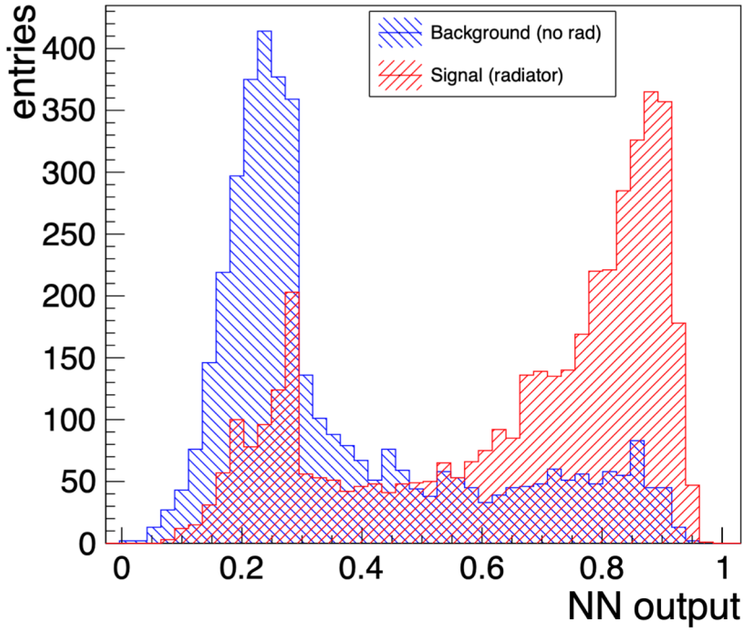
11.01.2019 Jennifer Ngadiuba - hls4ml: deep neural networks in FPGAs 25

Backup

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

GEMTRD neural 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 |
    +-----+-----+-----+-----+----------+
```

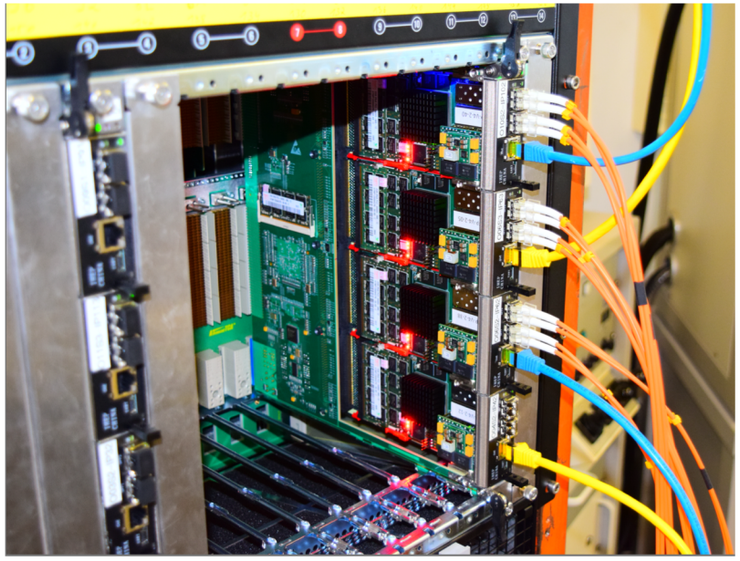
Latency = 85ns

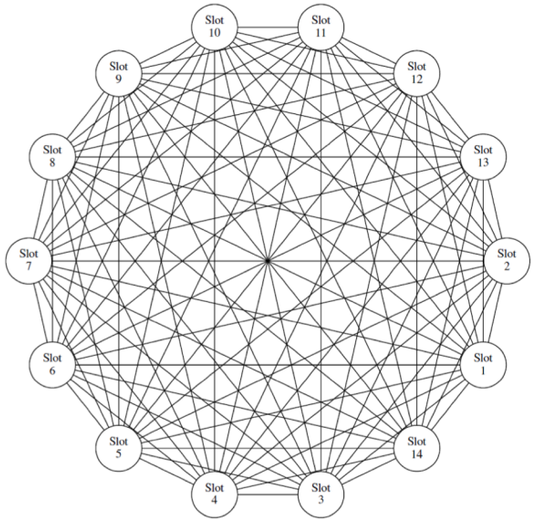
Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	2	-	-	-
Expression	-	-	0	24	-
FIFO	-	-	-	-	-
Instance	-	177	3132	10696	-
Memory	2	-	0	0	-
Multiplexer	-	-	-	81	-
Register	-	-	1423	-	-
Total	2	179	4555	10801	0
Available SLR	1440	2280	788160	394080	320
Utilization SLR (%)	~0	7	~0	2	0
Available	4320	6840	2364480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

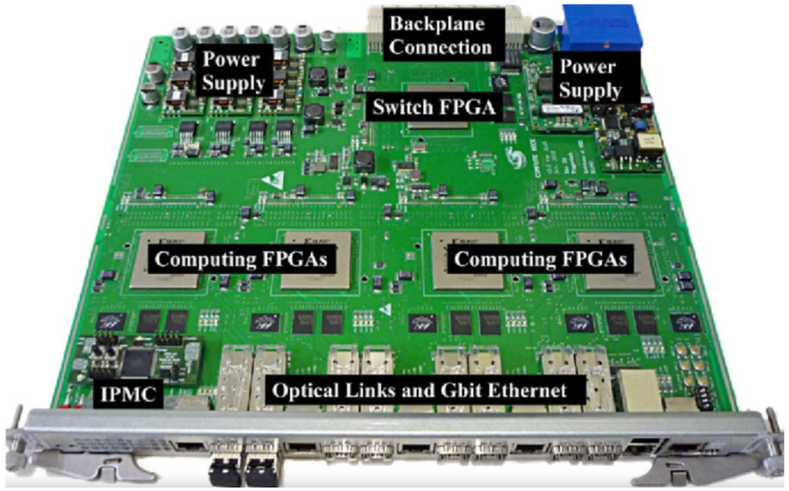
DSP utilization 2%

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.







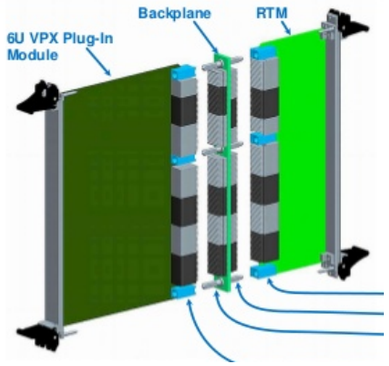
ADC based DAQ for PANDA STT

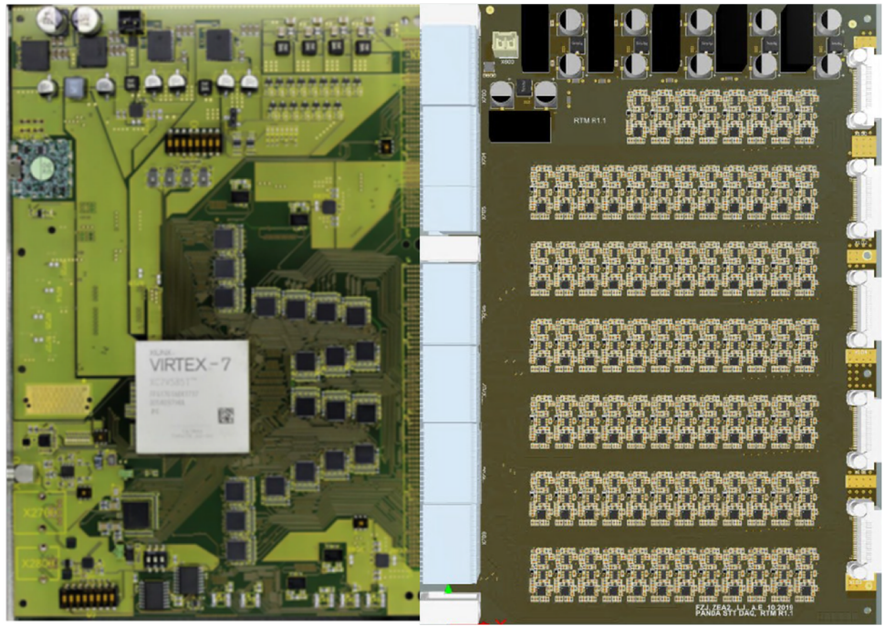
Jefferson Lab
Thomas Jefferson National Accelerator Facility

Level 0 Open VPX Crate

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

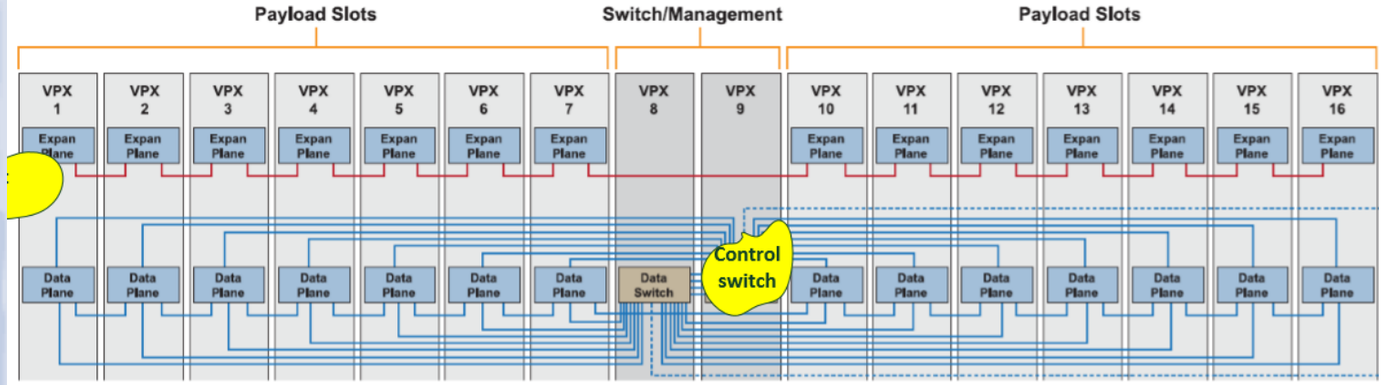
- 160 channels (shaping, sampling and processing) per payload slot, 14 payload slots+2 controllers;
- **totally 2200 channels per crate**;
- time sorted output data stream (arrival time, energy,...)
- noise rejection, pile up resolution, base line correction, ..





- 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.*
- *The design makes it easy to add ML-FPGA to data processing.*

L. Jokhovets, P Kulessa ..

JÜLICH Forschungszentrum