FPGA driven Machine Learning for Heavy Flavour event identification

Jakub Kvapil for the FastML team

2023 RHIC/AGS Annual Users' Meeting

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

LA-UR-23-28744

Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.

Motivation – Heavy Flavour

- Heavy-flavour (HF) events are very rare ~1% probability at RHIC energy
  - RHIC collision rate is several MHz, sPHENIX readout 15 kHz (DAQ - 300 Gb/s)
  - by recording 10 % HF events of all p+p collisions, the HF sample can be increased by a factor of 500x
- The goal is to create a triggering system to select HF production
  1. High Level Trigger
     - If all collision data are collected one can select those events on CPU farm
  2. Hardware trigger
     - This is useful if there is a slow/large data volume detector and the DAQ does not allow full online processing
     - Due to requirement on low latency (~10 us) this triggering system must be implemented on hardware

Motivation – The challenges

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

Motivation – Towards Electron-Ion Collider (EIC)

- The aim is to deploy future system on EIC
  - EIC allows us to constrain initial state effects
  - Precision heavy flavor data from EIC can further constrain nPDFs
  - AI-based electron tagging with streaming readout to identify the (non)interesting Deep-Inelastic-Scattering (DIS) processes in the e+p/A collisions.
    - based on the measured scattering electron energy and direction

- Integrate the AI-based heavy flavor trigger system demonstrator into the sPHENIX experiment for p+p run in 2024

For more information see Xuan talk Today at 11:50 AM

The proposal

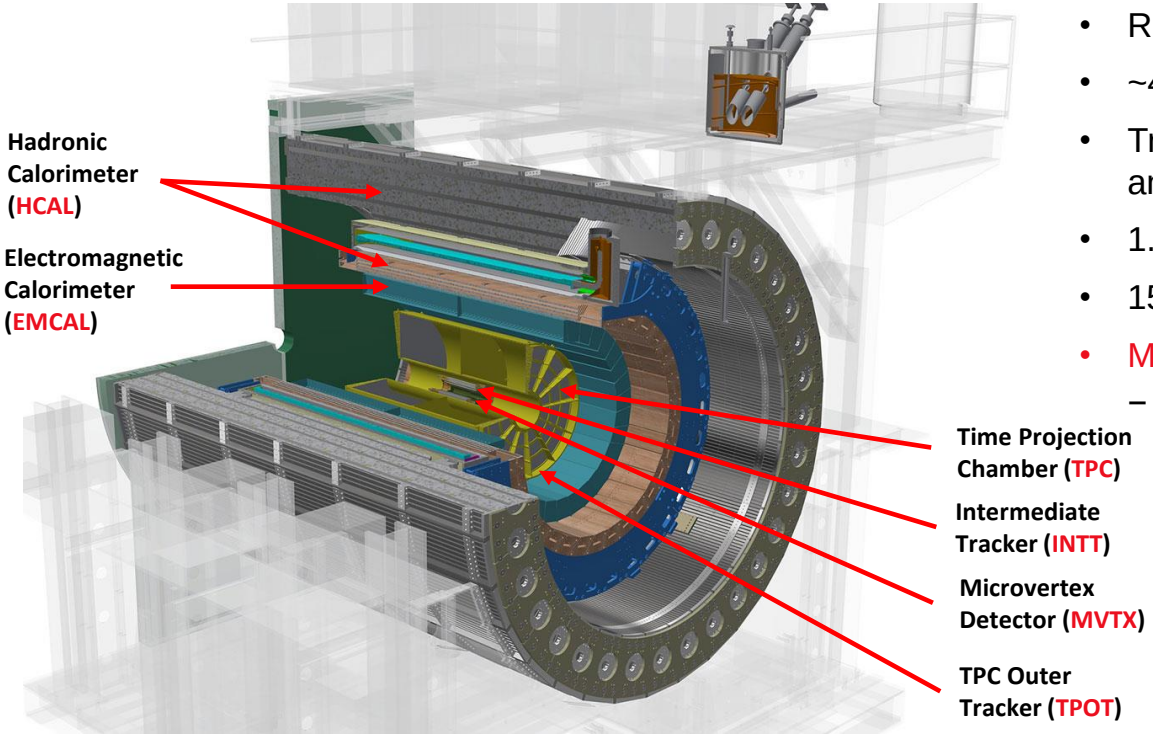
Intelligent experiments through real-time AI: Fast Data Processing and Autonomous Detector Control for sPHENIX and future EIC detectors

A proposal submitted to the DOE Office of Science
April 30, 2021

- The mission
  - Efficiently extract critical and strategic information from large complex data sets
  - Address the challenges of autonomous control and experimentation
  - Artificial Intelligence for data reduction of large experimental data

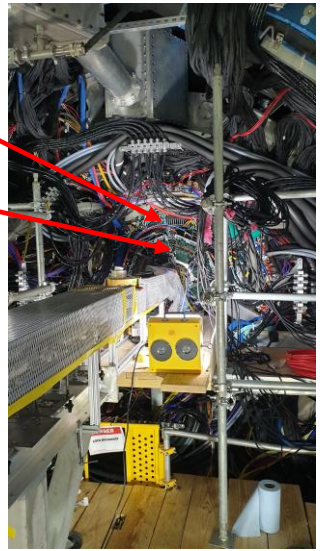
Extension proposal submitted 11th January 2023

sPHENIX experiment



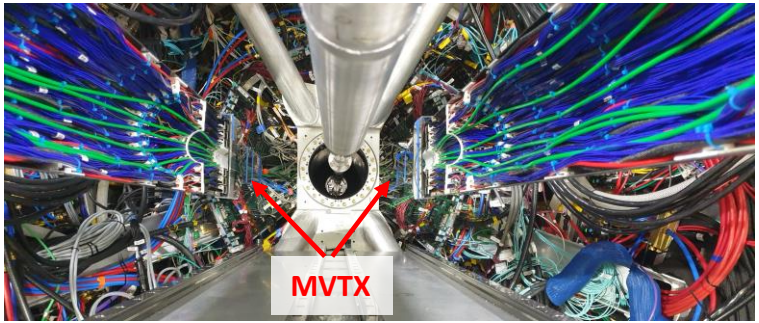
- Running period 2023-2025
- ~4m long, ~3m high, 1000 tons
- Tracking detectors (MVTX, INTT, TPC, TPOT) and calorimeters (EMCAL, HCAL)
- 1.4 T Magnetic Field, $|\eta| \leqslant 1.1$
- 15 kHz Trigger Rate (due to calorimeter limit)
- MVTX and INTT
  - Silicon detectors capable of streamed readout

MVTX and INTT



MVTX - Active area ~1685 cm^2

- Based on ALICE ALPIDE chips, with ATLAS FELIX backend
  - Monolithic Active Pixel Sensors
  - Very fine pitch (27 μm x 29 μm)
  - Good Time resolution ~5 μs
- 3 layers
  - Number of staves in each layer: 12, 16, 20
  - 9 chips per layer ~ 230M channels



INTT

- Silicon Strip Detector
  - Hamamatsu silicon modules
  - Pitch 27 μm x 16 (or 20) mm
  - Excellent Time resolution ~100 ns
- 2 layers
  - 56 ladders total
  - 360k channels

The sPHENIX Physics

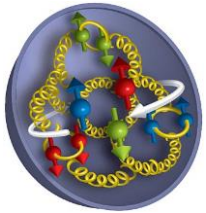
For more information see Antonio talk Wednesday at 3:50 PM

1. Hot QCD

Energy flow in QGP, transport coefficients, phase transitions, critical point

2. Cold QCD

Proton spin, cold nuclear matter effects



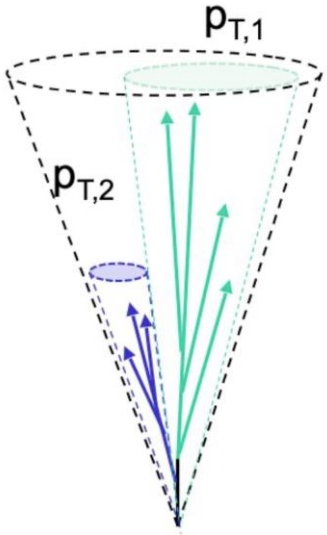
Heavy-flavour (HF) hadrons and jets

Energy flow in QGP, transport coefficients, HF hadronization, fragmentation, and flow, constrain QCD based models, dead-cone effect

Jet structure and substructure

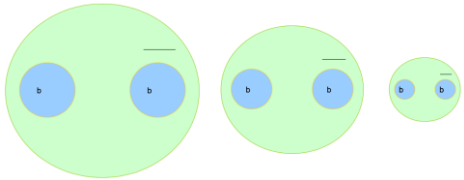
Vary momentum/angular scale of the probe

Look at energy/momentum distribution inside jet



Quarkonium spectroscopy

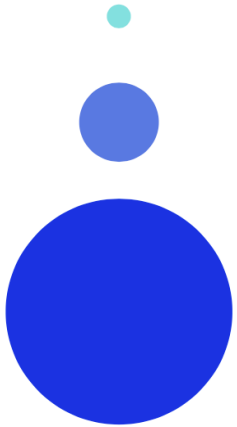
Vary size of the probe

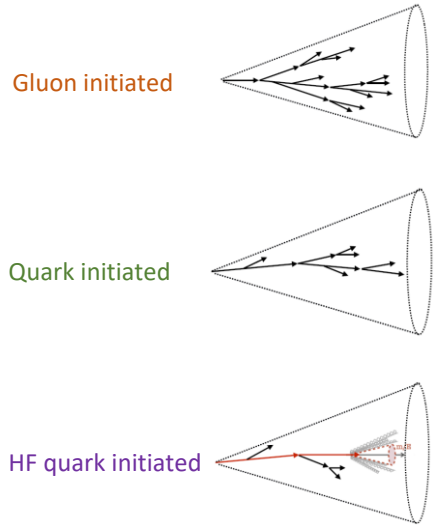


Y(3s) 0.78fm Y(2s) 0.56fm Y(1s) 0.28fm

Parton energy loss

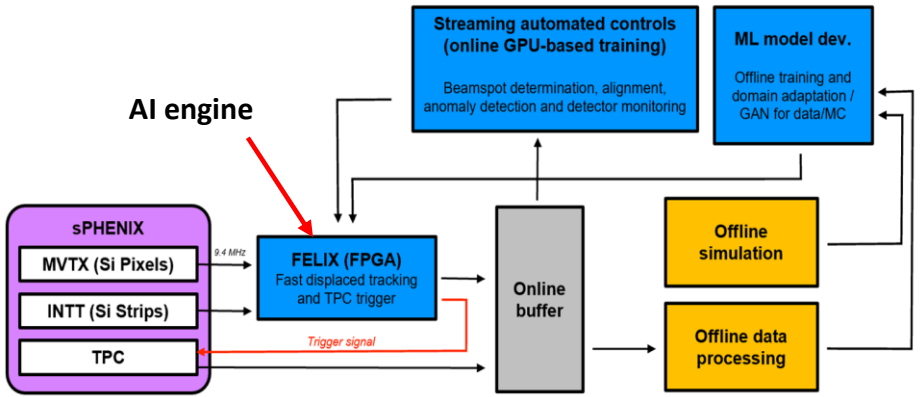
Vary momentum/mass of the probe



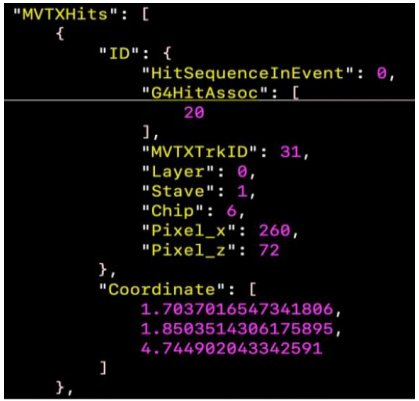


The DAQ Data Flow

- MVTX 144 links @ 3.2 Gbps and INTT raw data stream will feed two AI engines (one for each hemisphere)
  - 8b10b protocol with links driven @ 10Gbps (tested up to 14 Gbps)
  - 24 links for MVTX and 24 links for INTT per AI engine
- The decision signal of heavy flavor event from the AI-Engine will be sent out via the LEMO connectors to the sPHENIX DAQ global trigger system to initiate the TPC readout in the triggered mode



Simulated events

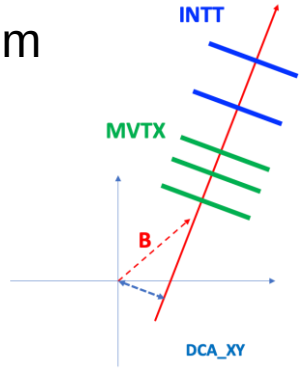


- Events were simulated in order to train the neural network
- JSON format and store the information
  - event by event
  - event header (event ID, collision vertex)
  - raw data (all hits produced by charged particles through the MVTX and INTT)
  - truth track information.
- Two data sets were simulated using the PYTHIA 8 event generator and GEANT 4 configured for minimum-bias p+p at $\sqrt{s}$ = 200GeV
  1. Data set $D^0 \rightarrow K^- \pi^+$ candidates from cc events within the sPHENIX acceptance
  2. Background simulation with no HF events.
- Next simulate also bb events

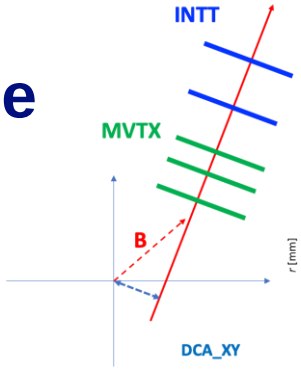
The ML algorithm – The overview

For more information see Giorgian talk Today at 1 PM

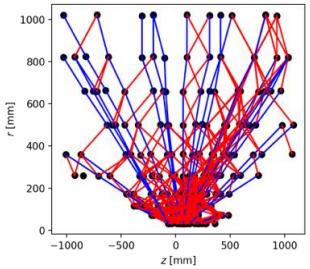
- Based on Graph Neural Network (GNN)
  - Detector and physics knowledge improves prediction
  - Based on PyTorch and PyTorch Geometric
- Topological selection of HF signals on FPGA
  - Tracking and clustering must be done on FPGA
- Beam-spot and anomaly detection on GPU based feed-back system
- Initial training on simulated data from MVTX and INTT
  - On GPU
  - NVIDIA Titan RTX, A500, and A6000

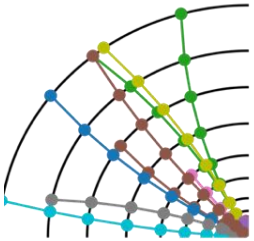


The ML algorithm – The pipeline

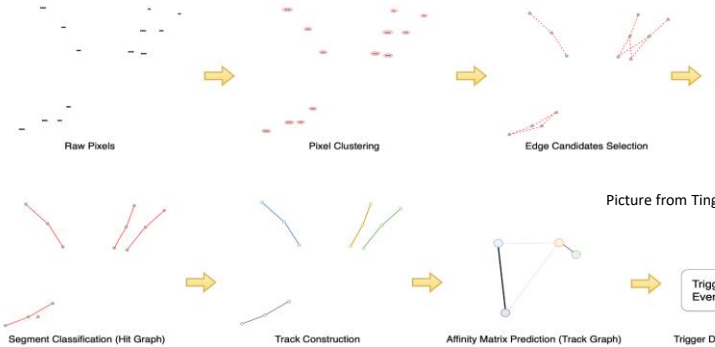


For more information see Giorgian talk Today at 1 PM





- Three stages of event processing
  1. Hits clustering
  2. Track reconstruction + outlier hits removal
     - by connecting the hits across different detector layers into hit pairs.
     - apply geometric constraints and down select the hit
  3. Trigger detection
- Graph Neural Network to solve
  - the track reconstruction problem
  - the inter-track adjacency matrix prediction
  - the graph level trigger detection



Picture from Tingting Xuan

- We propose a novel method to treat the events as track graphs instead of hit graphs. This method is driven by the physics (transverse momentum)
  - Estimate momentum based on silicon hits -> 15% improvement on trigger decision

The ML algorithm – The challenge

For more information see Giorgian talk Today at 1 PM

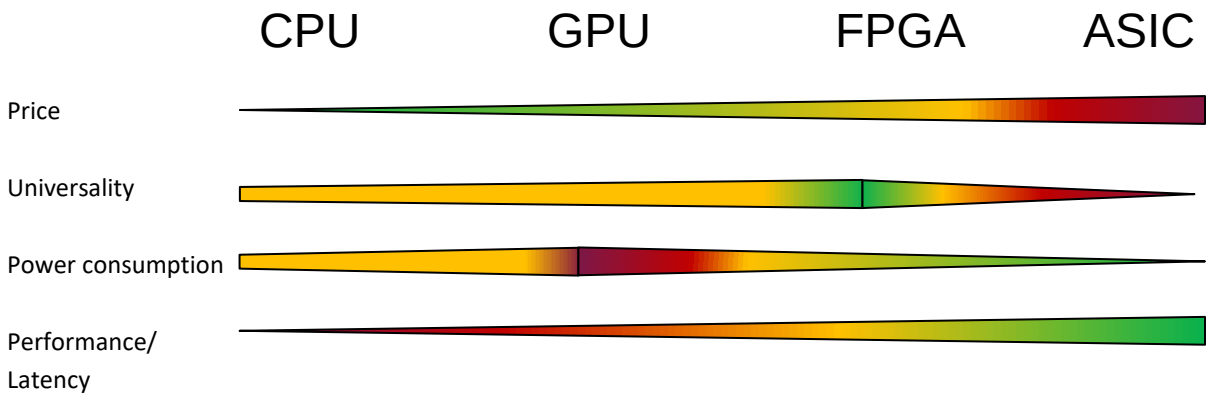
- Challenges
  - To provide an end-to-end solution that uses raw detector readout hit information to make trigger decisions for data collection.
  - To design a neural network compatible with the given detector readout and capable of learning a broad spectrum of physics properties
- using low-level hits to build the high-level trigger decision.
- Growing sub-field of geometric deep learning

$D^0 \rightarrow K^- \pi^+$ sample

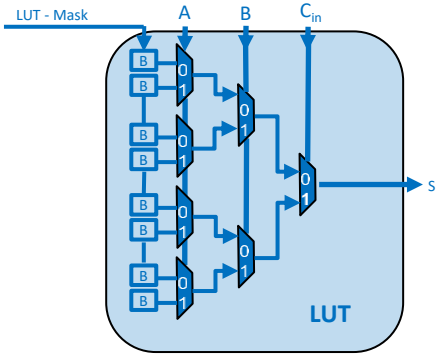
1% signal/background ratio			0.1% signal/background ratio		
Background Rejection	Efficiency	Purity	Background Rejection	Efficiency	Purity
90%	72.5%	7.25%	90%	78%	0.78%
95%	48.9%	9.78%	95%	50%	1.0%
99%	15.0%	15.0%	99%	17%	1.7%
99.33%	10.5%	15.74%	99.33%	11.0%	1.65%

What is an FPGA?

- Field Programmable Gate Array (FPGA)
  - Different electronics components that can be connected as user wants
- CPU (sequential, parallel, fixed instruction, high memory)
- GPU (sequential, highly parallel, fixed instruction)
- FPGA (concurrent, highly parallel, no instruction, reconfigurable)
- ASIC (concurrent, highly parallel, custom, not reconfigurable)

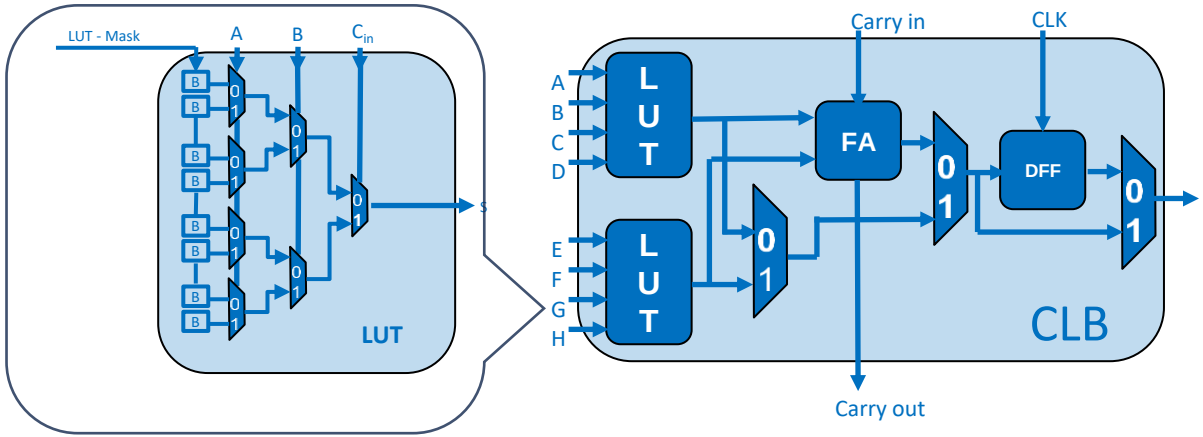


What is an FPGA made of?



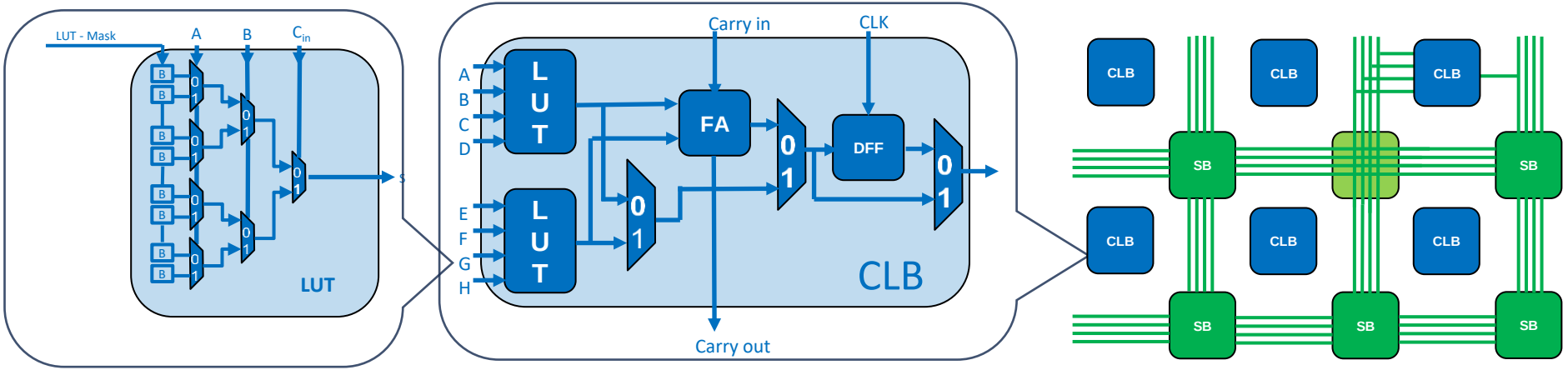
- Look-Up table (LUT)
  - Smallest FPGA unit
  - Electronic “switch”
  - Each LUT has its own mask
    - All results are precalculated
    - Input is an address in a memory
  - Used for Boolean operations, arithmetic, and memory

What is an FPGA made of?



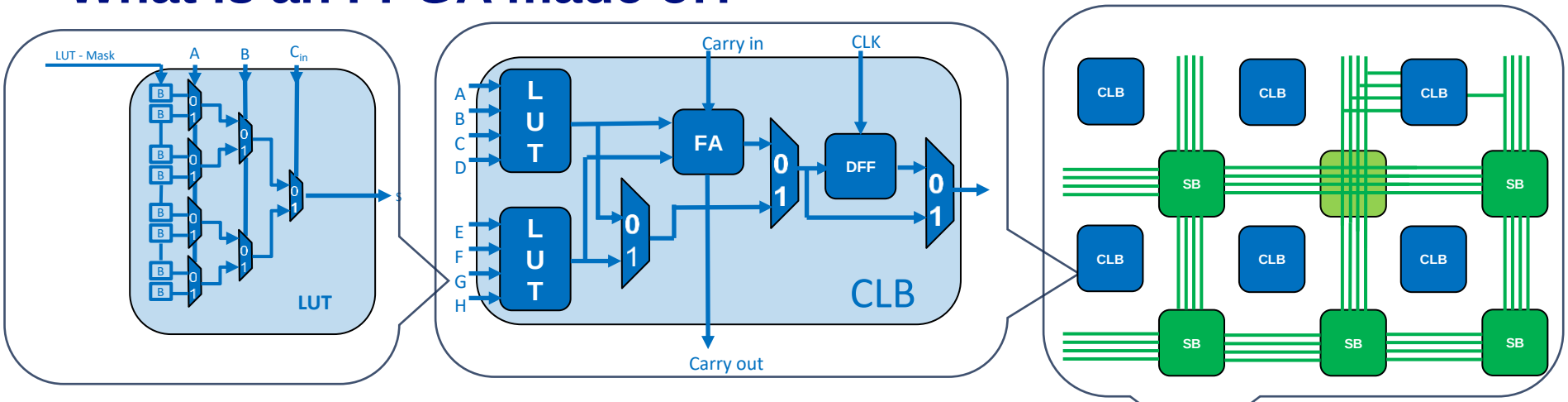
- Configurable Logic Block (Xilinx) or Logic Array Block (Altera)
  - LUTs are grouped into logic blocks usually with D type flip-flops (DFF) and Full adders (FA)
    - Vendor specific, trade secret
  - Flip-Flops to register data with a FPGA clock cycle
  - Full adders for fast connection between blocks

What is an FPGA made of?



- Switch Boxes (SB) connect components on FPGA
  - Switch box can connect any incoming wire to any outgoing wire
  - Full customization
  - Obviously, there is a limited number of wires, that is why large designs can be sometimes challenging
  - Bitstream (binary representation of FPGA code) set up all the LUT masks, switches and components

What is an FPGA made of?

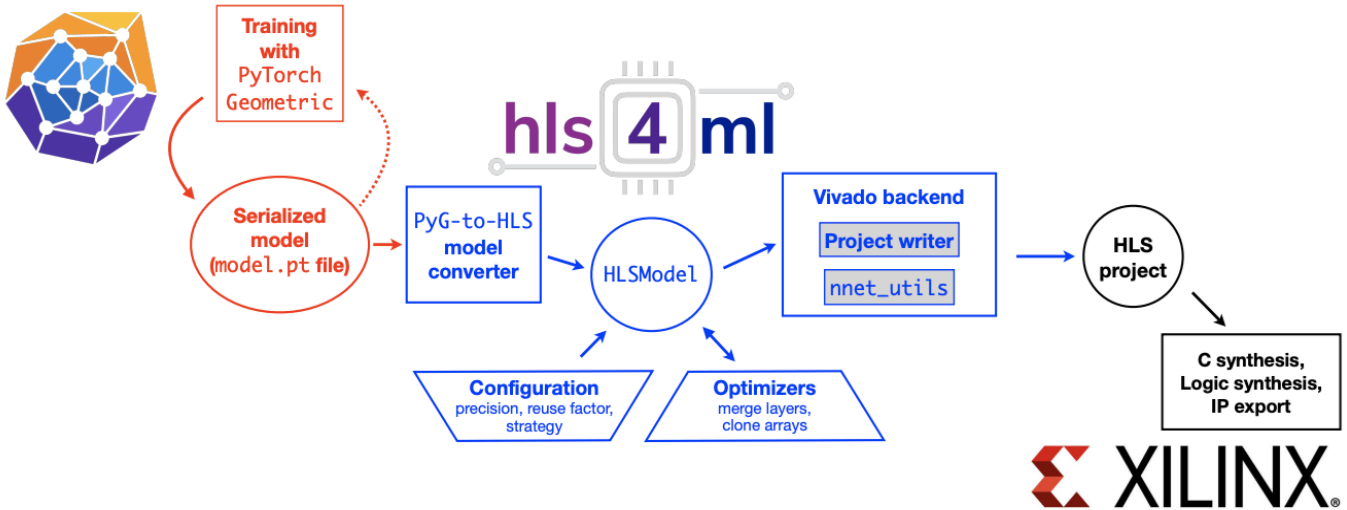


- Digital Signal Processor (DSP)
  - Specialised units for multiplication and arithmetic
  - The driving component of NN matrix multiplication
- Block Random Access Memory (BRAM)
  - Specialised unit of small fast memory (RAM, ROM, FIFO)

hls4ml Translation to Firmware

arXiv:2112.02048

- High Level Synthesis for Machine Learning (hls4ml)
  - Python package for machine learning inference in FPGAs
- Hls4ml translates NN algorithm into high level synthesis and generates IP (Intellectual Property) core
  - translate it to the FPGA synthesizable high-level synthesis code.



hls4ml – Planned Upgrade

- High Level Synthesis for Machine Learning (hls4ml)
  - Python package for machine learning inference in FPGAs
- Hls4ml translates NN algorithm into high level synthesis and generates IP (Intellectual Property) core
  - translate it to the FPGA synthesizable high-level synthesis code.
- Third main upgrade underway, focusing on 3 examples
  - Example 1: Tri-muon reconstruction with the LHC (muon endcaps)
  - Example 2: Heavy flavor tracking at sPHENIX
  - Example 3: Silicon strip tracking at LHC

Reuse factor

multiplier

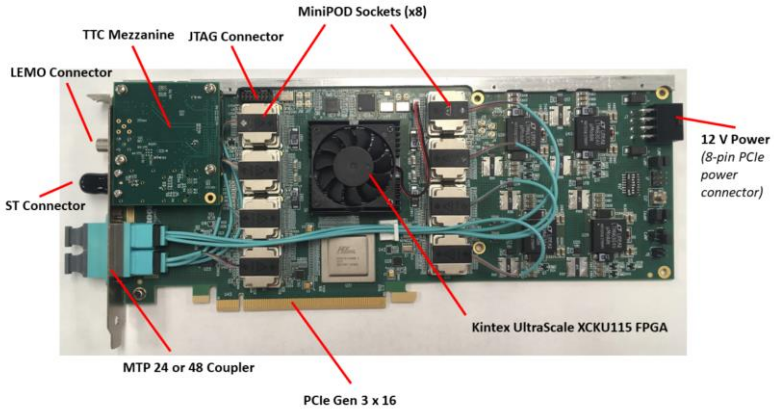
Design	$(n_{\text{nodes}}, n_{\text{edges}})$	RF	Precision	Latency [cycles]	II [cycles]	DSP [%]	LUT [%]	FF [%]	BRAM [%]
Throughput-opt.	(28, 56)	1	ap_fixed<14,7>	59	1	99.9	66.0	11.7	0.7
Throughput-opt.	(28, 56)	8	ap_fixed<14,7>	75	8	21.9	23.8	4.7	0.7
Resource-opt.	(28, 56)	1	ap_fixed<14,7>	79	28	56.6	17.6	3.9	13.1
Resource-opt.	(448, 896)	1	ap_fixed<14,7>	470	174	56.6	25.0	7.4	16.5
Resource-opt.	(448, 896)	8	ap_fixed<14,7>	1590	520	5.6	25.0	7.4	16.3

total width integer

@200 MHz, 1590 Cycles → 7.5μs

Hardware implementation

- Initial tests with Felix-711 (FLX-711) board by FNAL group
- Ongoing implementation on FLX-712 board which is also used for the MVTX and INTT backend
- Motivation to use FELIX board:
  - To reuse the PCIe implementation (16-lane Gen-3) and software tools provided by the FELIX developers
  - on-board FPGA is a Kintex Ultrascale XCKU115FLVF1924-2E
- Recent setup at BNL-1008 (sPHENIX counting house) so it is ready to receive MVTX and INTT data



ATL-DAQ-PROC-2020-005

Building the AI engine – initial 711 tests

- The AI Engine needs to provide a trigger signal in ~ 10µs and fit within the FPGA resources
- We use this standard well-understood benchmark model “Jet Tagger” to test the workflow arXiv:1804.06913
  - QKeras and converted to hls4ml to create an IP
  - 16 inputs (expert variables) and 3 dense hidden layers with 64, then 32, then 32 neurons;
  - Algorithm must have low latency and resource use
- PCIe and Jet Tagger tested on cropped version of FLX-711

Table 3: FPGA Resources Utilization post routing rm-4.11 FLX-711 Firmware.

Resource	Utilization	Available	Utilization %
LUT	343623	663360	51.80%
LUTRAM	39.24	293760	1.34%
FF	515551	1326720	38.86%
BRAM	1289	2160	59.68%
IO	126	728	17.31%
GT	40	64	62.50%
BUFG	82	1248	6.57%
MMCM	4	24	16.67%
PCIe	2	6	33.33%

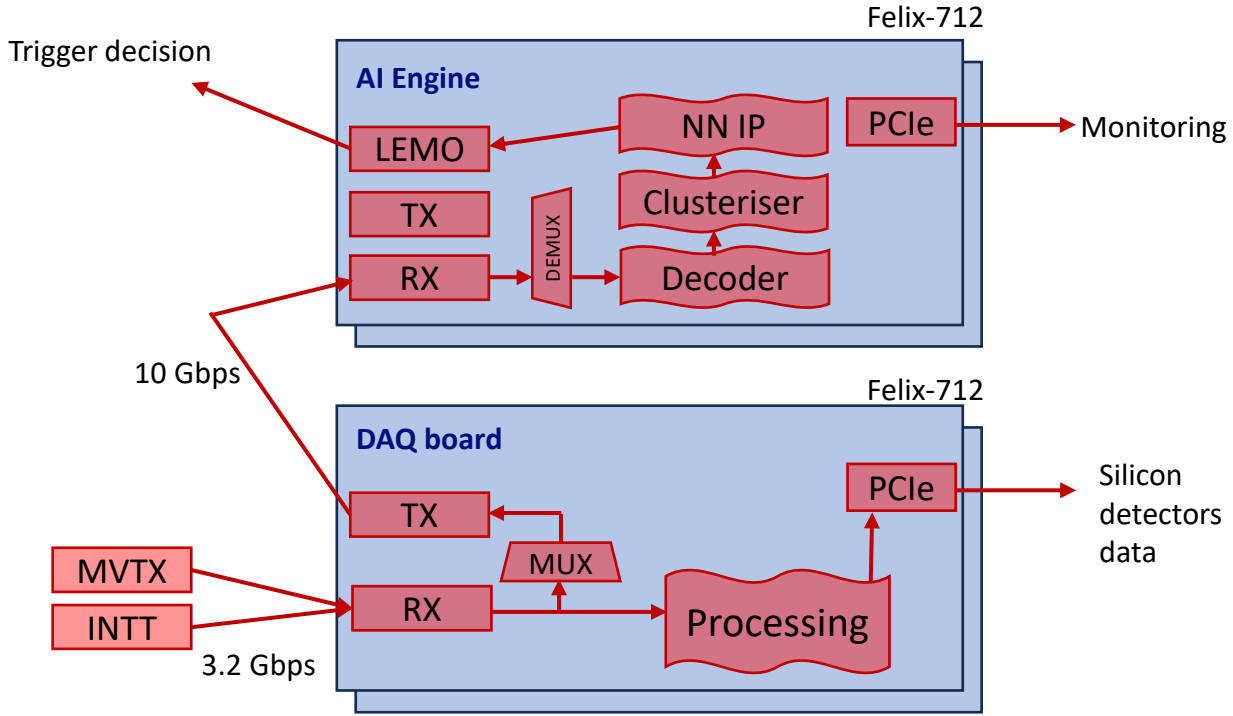
Building the AI engine – initial 711 tests

- The FELIX card is a data router. It needs an application to be instructed on the data movement.
- The application relies on an OS and a driver interfacing the FELIX card.
  - The testing relies on the interplay of Hardware, Firmware, and Software
- The challenge
  - Interfacing the NN to the board
  - Meet the timing constrains of the FELIX card

Table 3: FPGA Resources Utilization post routing rm-4.11 FLX-711 Firmware.

Resource	Utilization	Available	Utilization %
LUT	343623	663360	51.80%
LUTRAM	39.24	293760	1.34%
FF	515551	1326720	38.86%
BRAM	1289	2160	59.68%
IO	126	728	17.31%
GT	40	64	62.50%
BUFG	82	1248	6.57%
MMCM	4	24	16.67%
PCIe	2	6	33.33%

The firmware design - data flow



- Data needs to be decoded, clustered and feed the neural network
- Raw MVTX and INTT data packets:
  - 1 MVTX packet @5 us strobe
    - ~20 pp collisions (MB events) @4MHz pp collisions
  - 50 INTT packets (RF) @ 100 ns strobe
- Very challenging project to fit in the FPGA resources!
  - On CMS the AI engine is on 6 FPGA that are 2-3x bigger than FELIX FPGA

Predicted timeline

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

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

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

- Deploy device at sPHENIX pp/pA run
- EIC preliminary TDR (CD2)

- Final design for EIC TDR (CD3)
- Take advantage of new technology if required

- Deploy device at EIC

We are here!

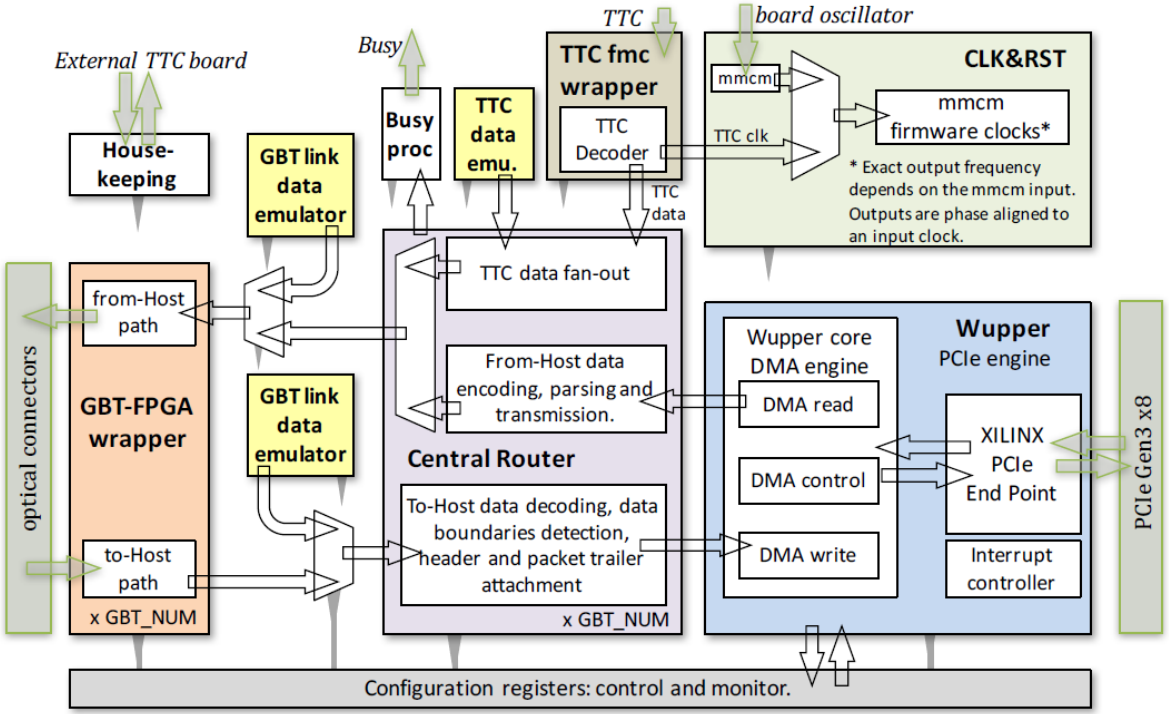
Summary

- Simulations are done for bb events
- NN algorithm tested to provide good precision while analysing two hemispheres independently
  - Testing the NN algorithm with higher pileup due to MVTX longer strobe window
- hsl4ml undergoing upgrade to provide better performance
- Initial tests with FLX-711 board done by FNAL group
- Ongoing implementation on FLX-712 board in sPHENIX counting house
- A new FLX-812 board arrived to BNL which will be the base for EIC development

Thank you for your attention

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

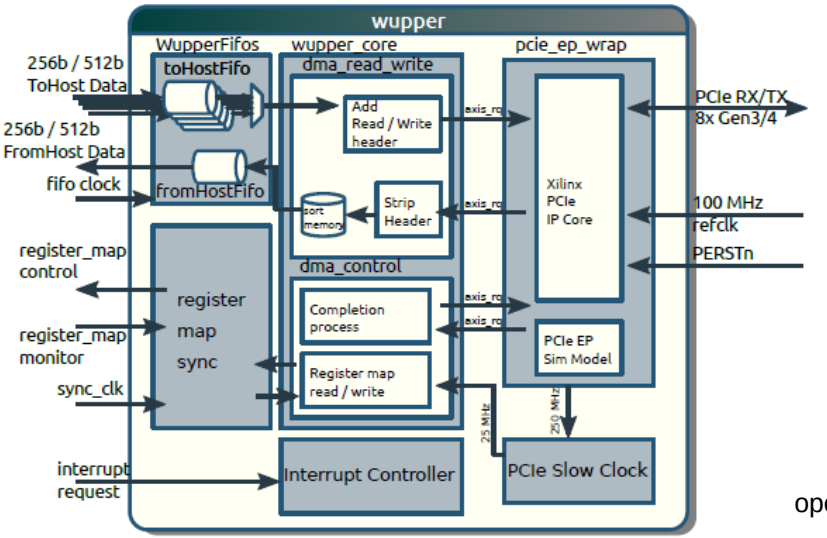
FLX-712 block diagram



TWEPP 2018

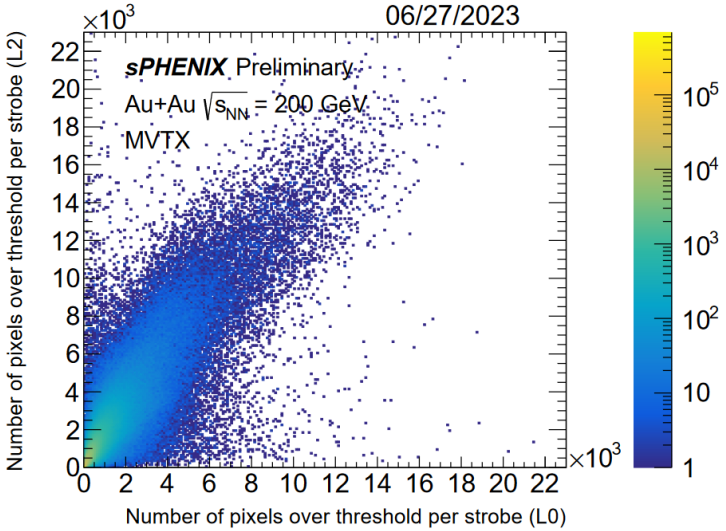
FELIX PCIe interface – The Wupper

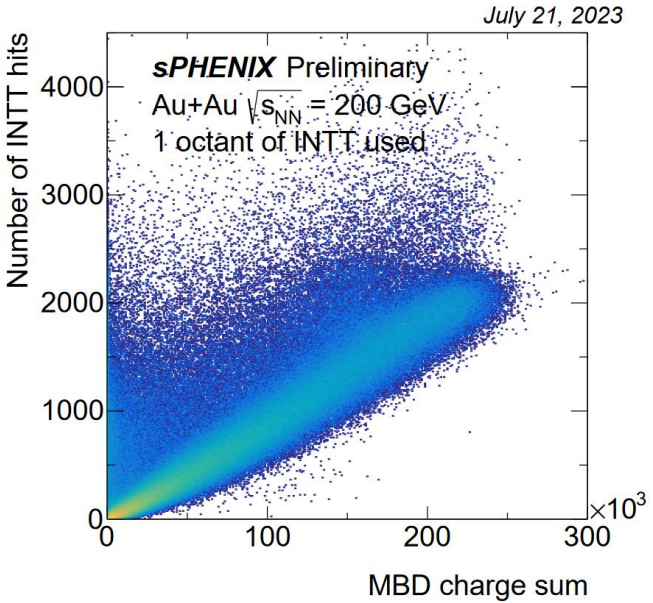
- Wupper is designed for the ATLAS / FELIX project to provide a simple Direct Memory Access (DMA) interface to a standard FIFO. This FIFO has the same width as the Xilinx AXI4-Stream interface (256 or 512 bits) and runs at 250 MHz.



opencores.org/projects/virtex7_pcie_dma

MVTX and INTT commissioning performance





- Timing in detectors on good track