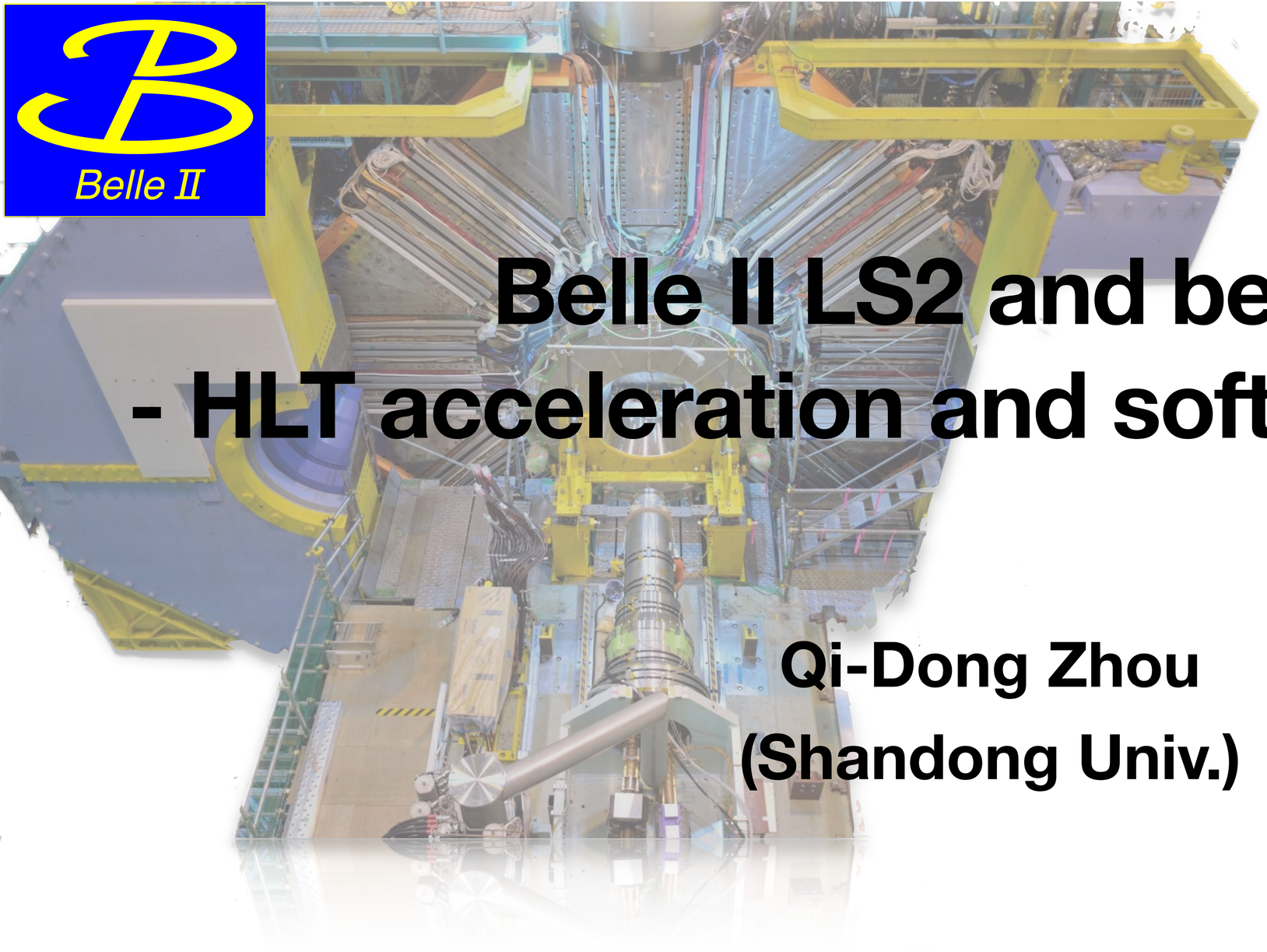


Belle II LS2 and beyond
- HLT acceleration and software trigger -

Qi-Dong Zhou
(Shandong Univ.)

24 Oct. , 2025
Belle II Trigger and DAQ Workshop

Motivations of trigger-DAQ upgrade

TRG/DAQ Joint Meeting Oct. 2024

Physics

- Tau trigger efficiency is 70~90%
- Low multiplicity trigger efficiency (single photon trigger pre-scaled)
- Low-momentum track (slow pion) trigger efficiency
  - SVD trigger !
- “Anomaly” trigger
  - Design a special trigger line for some specific physics channel
- Trigger efficiency of displaced vertex

Hardware limitation:

- DAQ system is designed to handle 30 kHz (Limited by SVD APV25 buffer)
  - L1 latency 4.3 us (SVD APV25 buffer)
    - CDC DNN trigger latency ~500 ns, latency already limited more large model
- L1 trigger rate will reach to ~20 kHz at 0.9×10^{-35} cm^{-2} s^{-1} (13 HLT units, w/o Hyperthreading)
  - Planed full HLT: 15 units (8000 CPU cores)
  - Low multiply: single photon trigger already pre-scaled during exp. 24

Short term solution (w/ SVD)

TRG/DAQ Joint Meeting Oct. 2024

L1 trigger upgrade (limited by latency and 30 kHz trigger rate):

- CDC noise filtering (DNN ready for physics)
- GNN for CDC and ECL trigger (In progressing)
- UT5 board (preparing)

HLT upgrade (limited by computing power):

- Increase CPU cores (2 more units)
- Improve methodology/performance of the software algorithm for reconstruction
  - Speed up the reconstruction algorithm (first round)
  - GNN for CDC tracking and ECL cluster reconstruction (CDC ready, ECL progressing)
- Introduce new Level 3 trigger
  - Software solution (talks in this meeting)

Motivations of trigger-DAQ upgrade

TRG/DAQ Joint Meeting Oct. 2024

Physics performance estimation after short term upgrade (personal options!!)

- Tau trigger efficiency is 70~90% -> ~90%
- Low multiplicity trigger efficiency (single photon trigger pre-scaled) -> expect slightly improvement
- Low-momentum track (slow pion) trigger efficiency -> expect slightly improvement
- “Anomaly” trigger -> not yet
  - Design a special trigger line for some specific physics channel
- Trigger efficiency of displaced vertex -> expect a mount of improvement

Middle term solution (VXD upgrade)

TRG/DAQ Joint Meeting Oct. 2024

My person option!!!! necessary if VXD is upgraded
L1 trigger upgrade (latency ~10 us?):

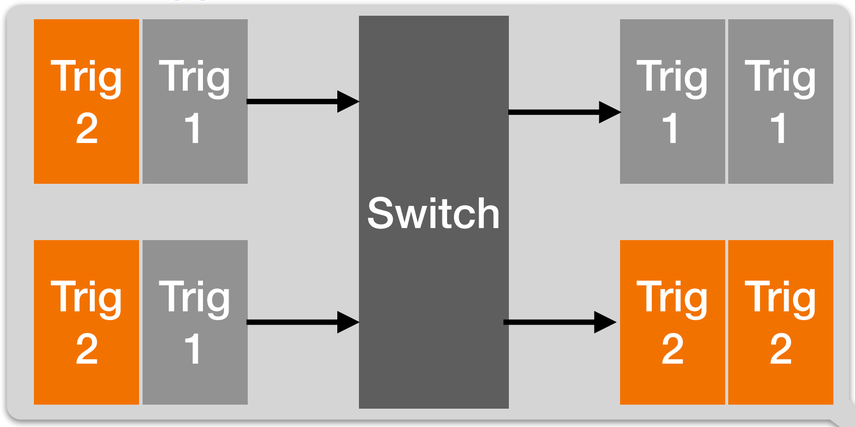
- UT5 board
- VXD trigger
- L1 trigger rate: ~100 kHz?

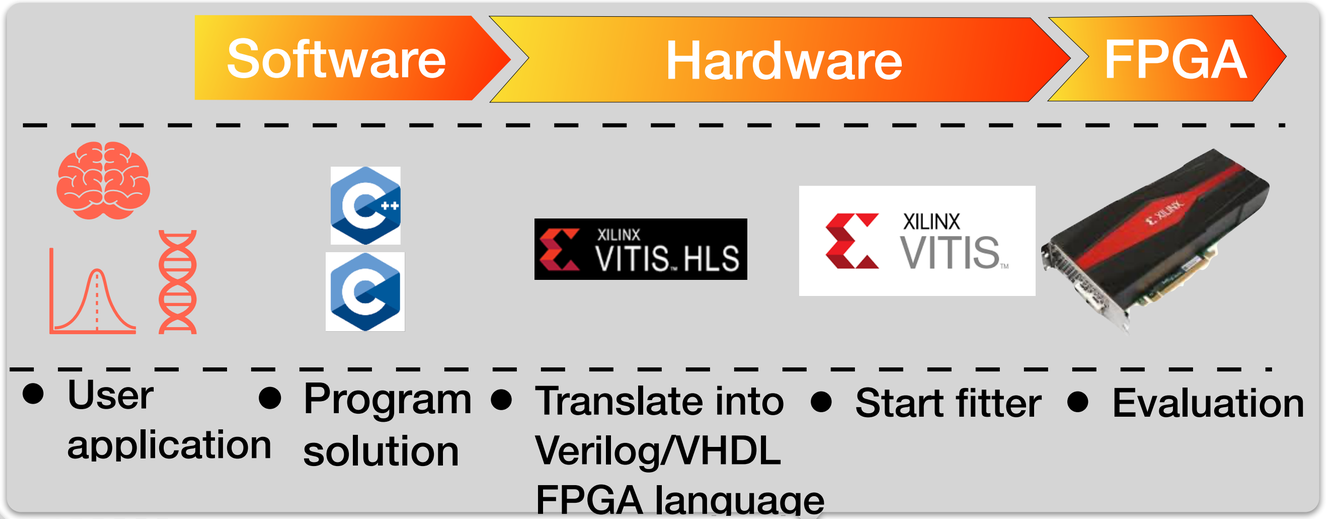
HLT upgrade (acceleration):

- Level 3 trigger:
  - Hardware + software solution
  - Dedicated subsystem: e.g. tracks quick reconstruction
  - FPGA + CPU (FPGA is preferred from latency point view)

Level 3: hardware + software solution

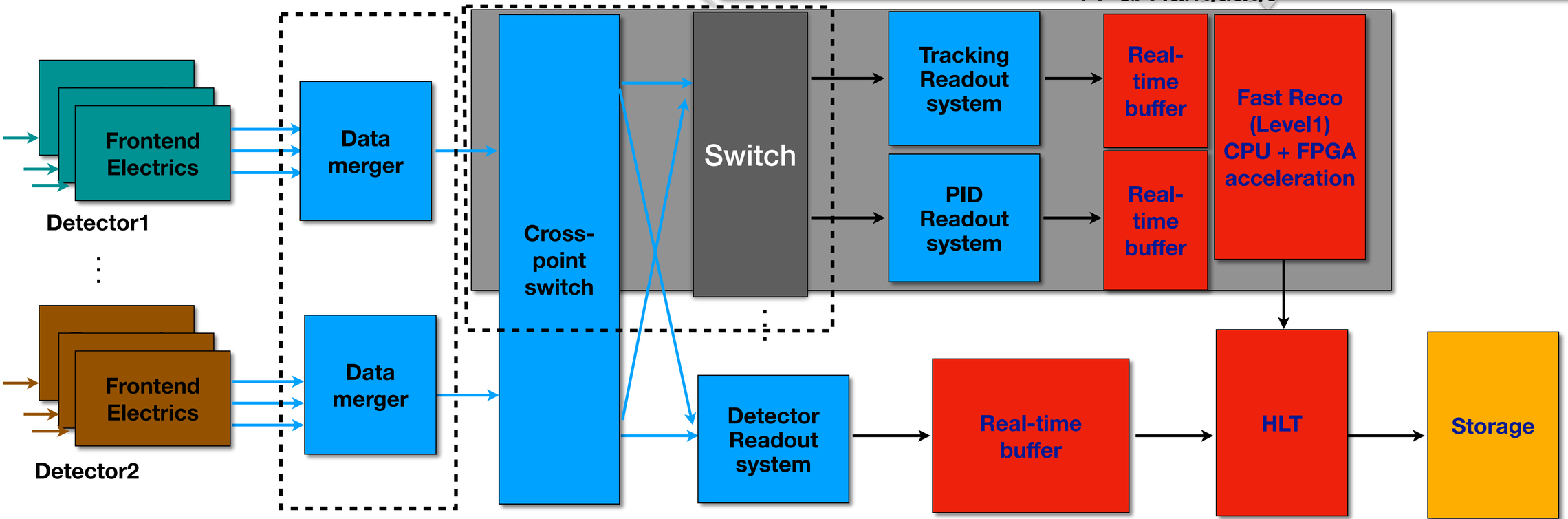
[Proposed on trigger DAQ workshop 2022]





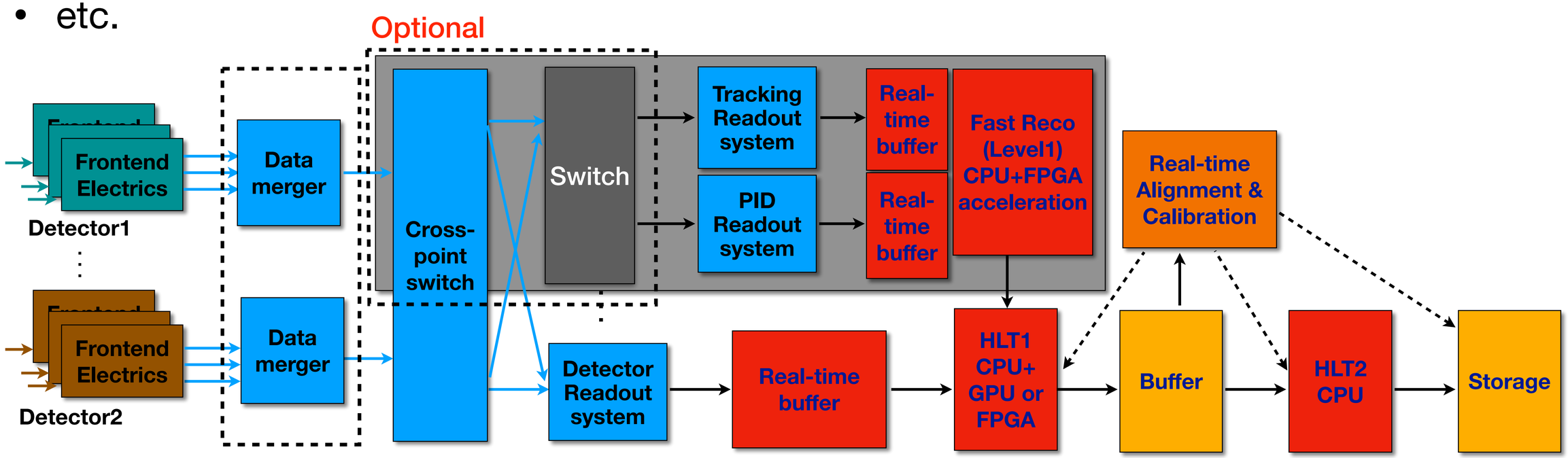
Sorting

Optional



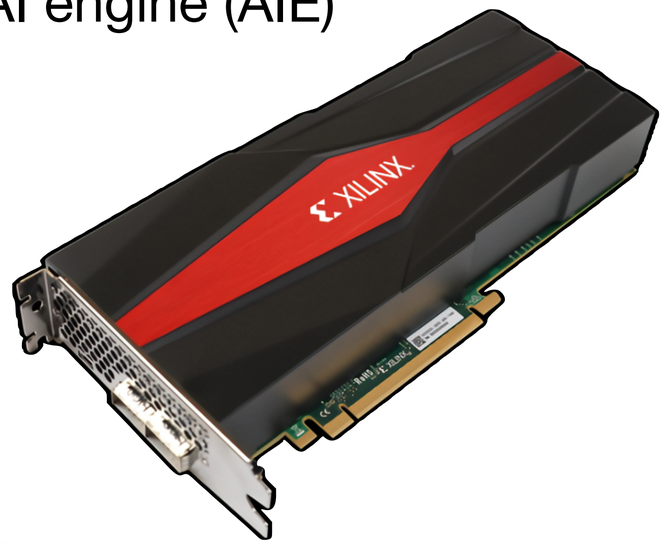
Long term: HLT with CPU + GPU + FPGA (DPU)

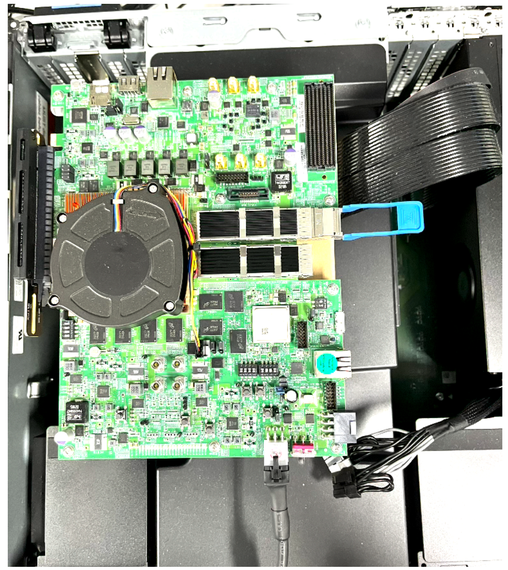
- Motivations:
  - Real-time alignment/calibration + skim scheme
    - HLT1 acceleration
  - Higher luminosity and background
  - Carbon neutral
- BASF2-Heterogenous computing
  - GPU/FPGA (DPU) based reconstruction algorithm
  - A large mount of disk as BUFFER system
  - High throughput network (e.g. infiniBand)
  - etc.

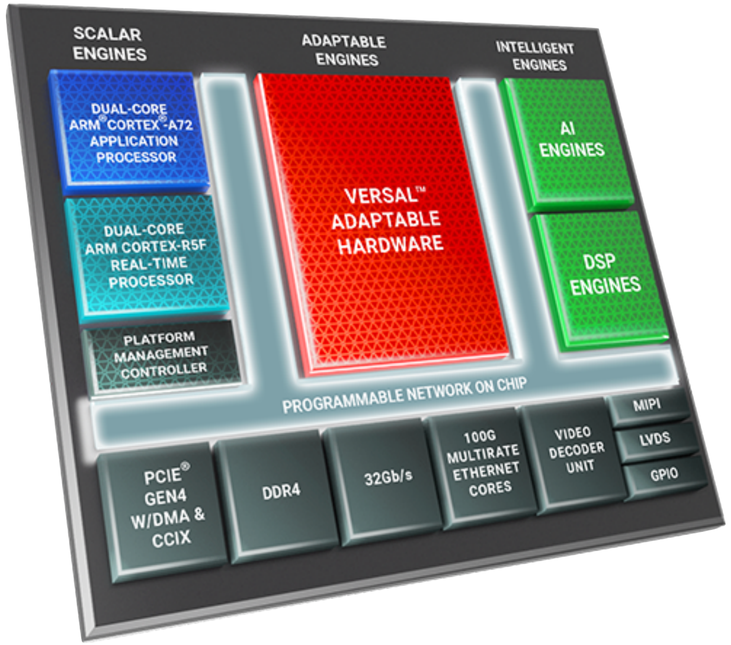


Heterogenous computing platform

- R&D of a new general FPGA device using the AMD Versal ACAP
  - Heterogenous acceleration (VCK190, VCK5000 evaluation kit)
    - AI engine (AIE)







UG1079

Figure 2: **AI Engine Array**

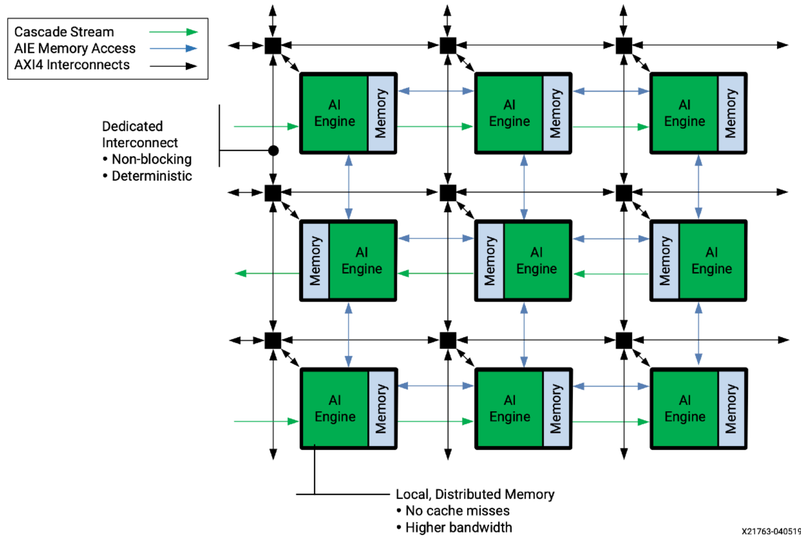
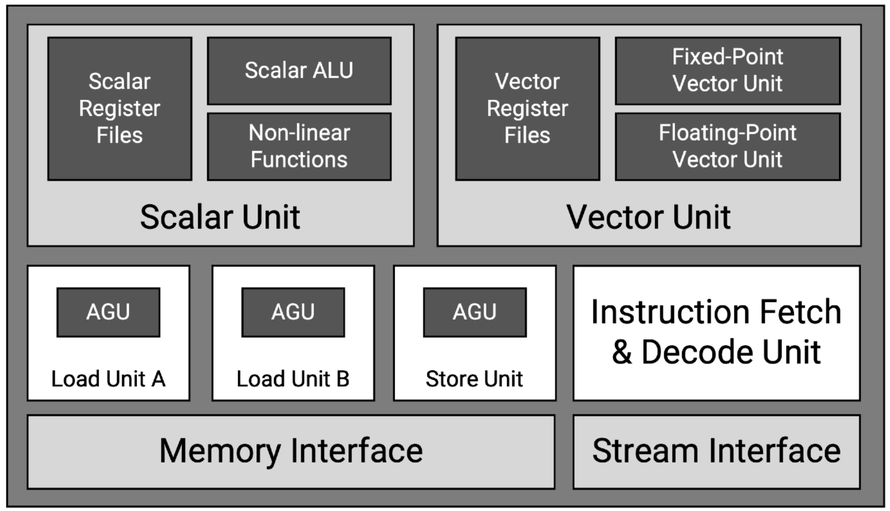


Figure 4: **AI Engine**



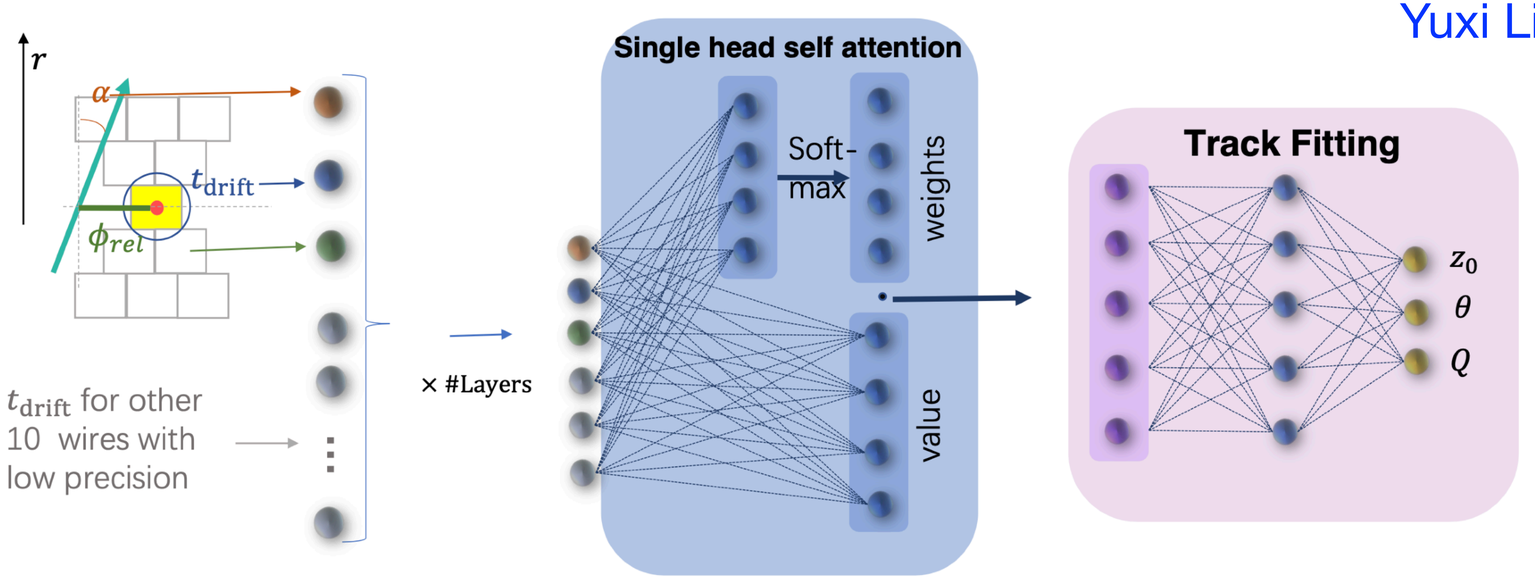
HLT: CPU vs. GPU vs. FPGA

Trigger DAQ workshop 2022

System	Processing power / HLT unit	Price (¥) / HLT unit	Ratio	Comment
CPU (Intel Xeon E5 2660)	480 cores	18,000,000	~6.5	Software update: ⭐⭐⭐ Power usage: ⭐⭐ https://arxiv.org/pdf/1906.11879.pdf
GPU (GeForce RTX 3090)	12 GPU GPU : CPU ~ 40 : 1	GPU: ~180,000 x 12 = 2,160,000 Server: 600,000 x 3 = 1,800,000 Total : 3,960,000	~1.5	Software update: ⭐⭐ Power usage: ⭐
FPGA (VCK5000, Versal ACAP VC1902)	5 FPGA card Versal : CPU ~ 100 : 1	FPGA card : ~300,000 x 5 = 1,500,000 Server: 600,000 x 2 = 1,200,000 Total : 2,700,000	1	Software update: ⭐ Power usage: ⭐⭐⭐

Deep Neural Network for Z trigger

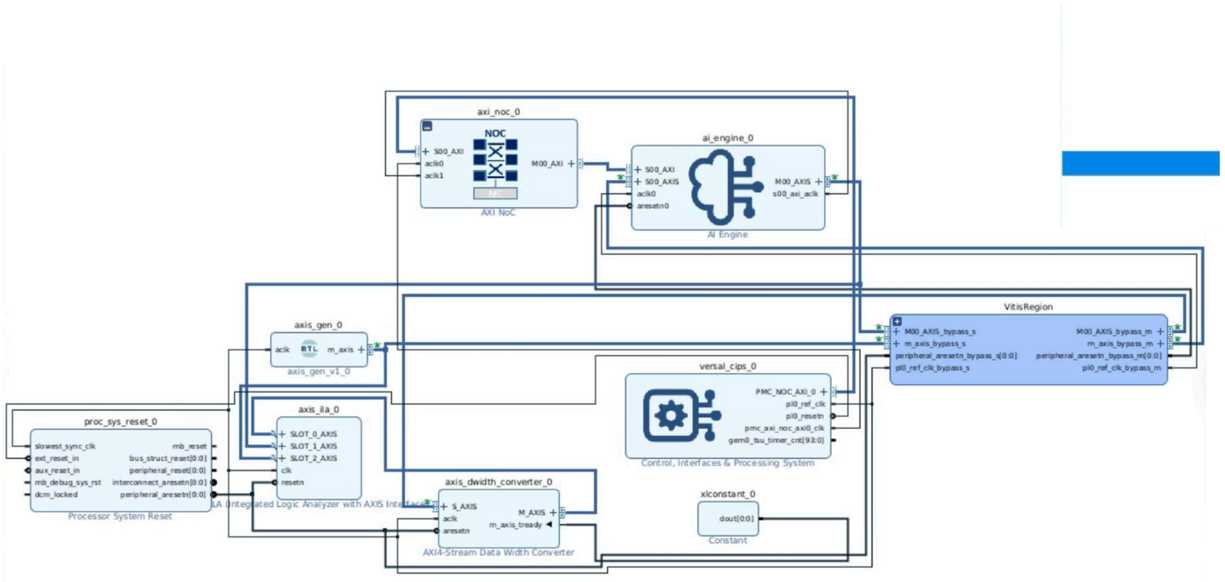
Yuxi Liu (KEK) @ IEEE RT 2024

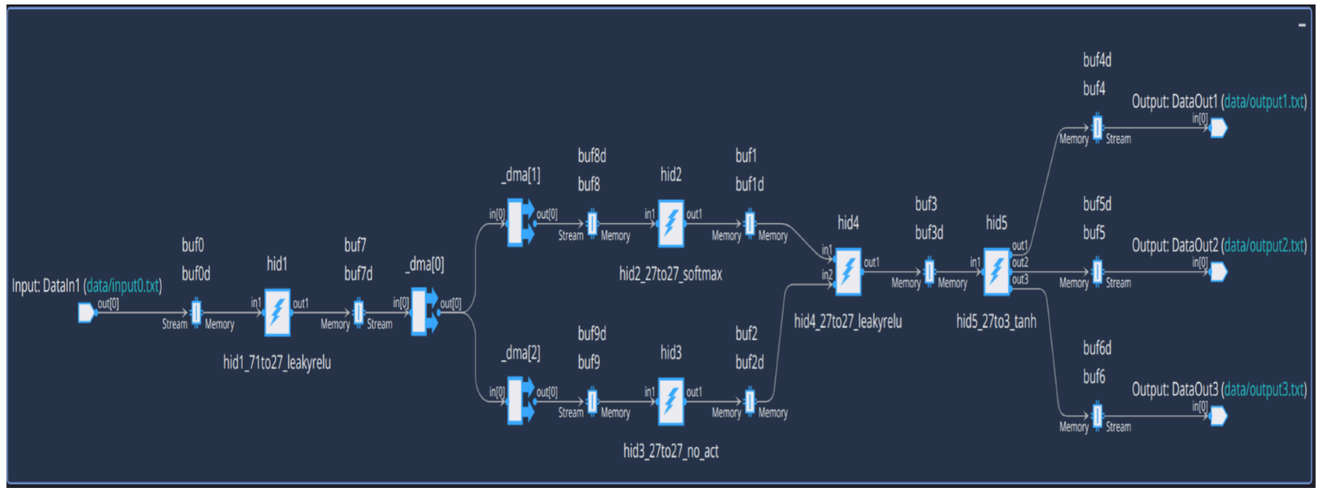


- Inputs: **Drift time t_{drift}**, **wires relative location ϕ_{rel}**, **Crossing angle α** for priority wires + **Drift time for all other wires**
- Introduce the **self-attention architecture** to "focus" on certain inputs
- Output track vertex z_0, track θ and **signal/ background classifier output (Q)**

- Latency : 76 clock = **551.2 ns** ;require: < 600ns
- FPGA resource (UT4: Virtex UltraScale XCVU160) usage:
  - DSP: ~75%, LUT: ~45%, others <30%
- At signal efficiency ~95%
  - Background rejection rate ~85%

DNN acceleration on Versal ACAP





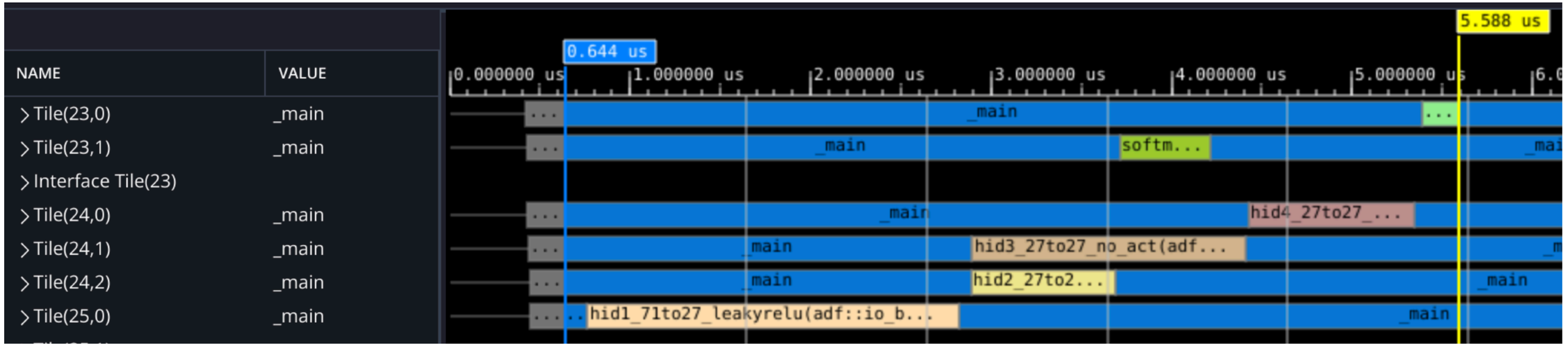
- DNN implementation:
  - Model on a “graph”
  - Dense layer on a “kernel”
- AI engine: C++ based coding on Vitis
  - AI engine libraries
  - AI engine specific functions
  - Scaler, Vector engines, pipelining, etc.

	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
Input nodes	71	27	27	27	27
Output nodes	27	27	27	27	3
Active Func.	LeakyReLU	Softmax	—	LeakyReRU	Tanh

AI Engine Resource Utilization

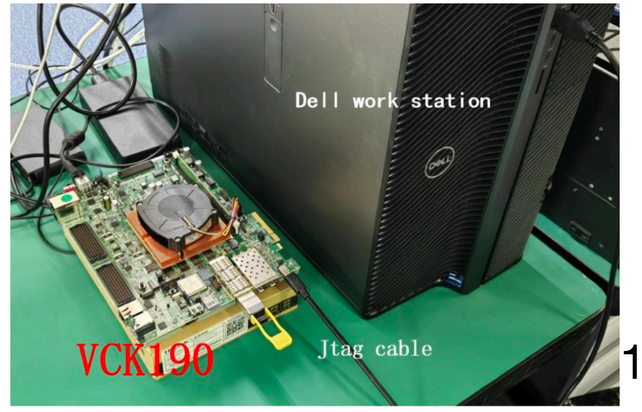
Tiles used for AI Engine Kernels:	5 of 400 (1.25 %)
Tiles used for Buffers:	7 of 400 (1.75 %)
Tiles used for Stream Interconnect:	8 of 450 (1.78 %)
DMA FIFO Buffers:	0
Interface Channels used for ADF Input/Output:	4 (PLIO: 4)
Interface Channels used for Trace data:	0

Latency optimization on Versal ACAP



	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Total
Input nodes	71	27	27	27	27	—
Output nodes	27	27	27	27	3	—
Active Func.	LeakyReLU	Softmax	—	LeakyReRU	Tanh	—
Ver.0 latency	~12us	~66us	~1.5us	~5.5us	~9.9us	~86us
Ver.1 latency	~2.1us	~1.3us	~1.5us	0.9us	~0.2us	~5us
Ver.2 latency	~0.48 us	~0.93us	~0.33us	~0.40us	~0.10us	~2.1us

Total **305 clk cycles** one instance. Clock period **10ns**.
Latency running on Versal ACAP is **3.05us**

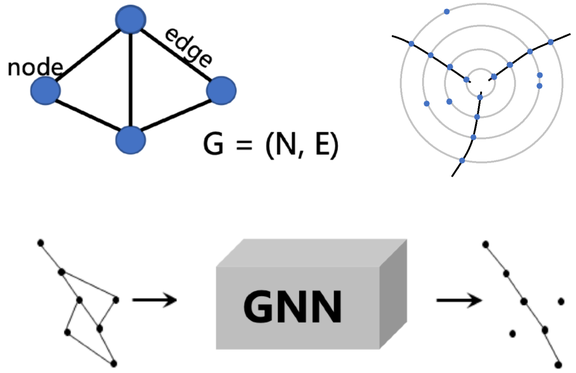


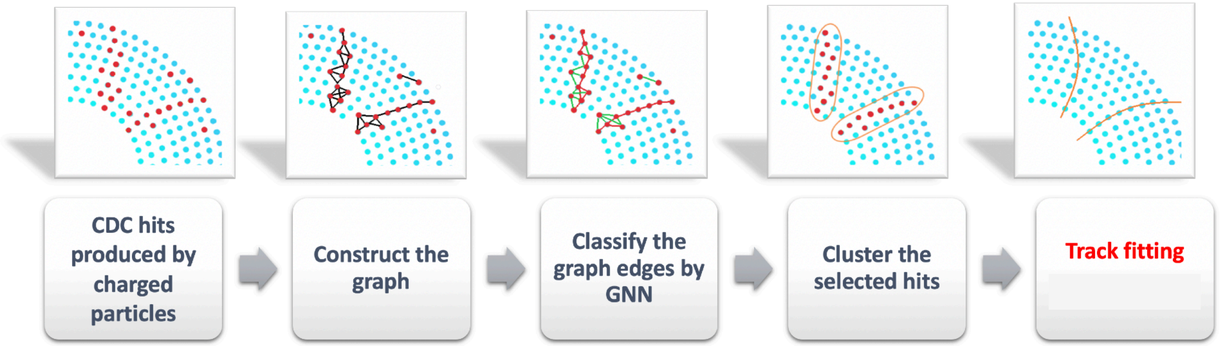
Machine Learning for Software Track Trigger (SFOTWARE)

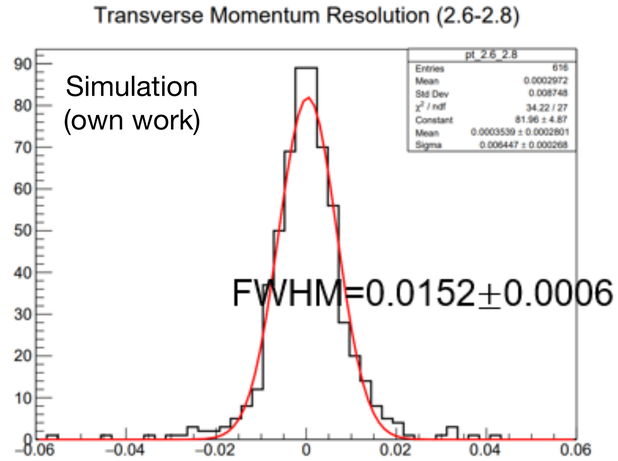
GNN for CDC track background filtering

- Developed a GNN algorithm (based on BESIII's algorithm) for Belle II CDC hits clean up
  - Inputs: TDC, position coordinates r, ϕ

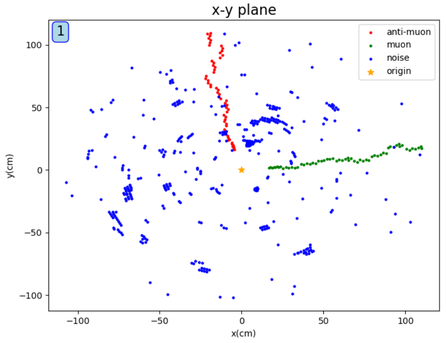
Noise filtering → Clustering → Fitting



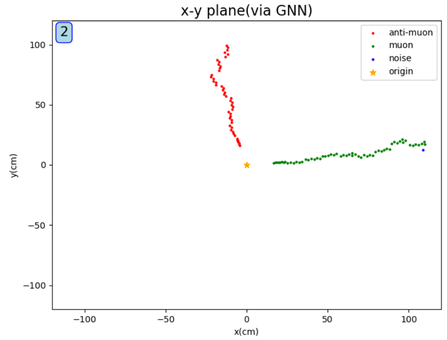




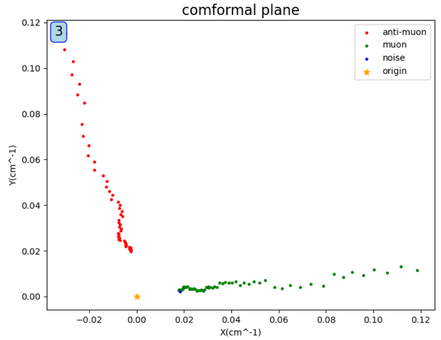
Simulation (own work)



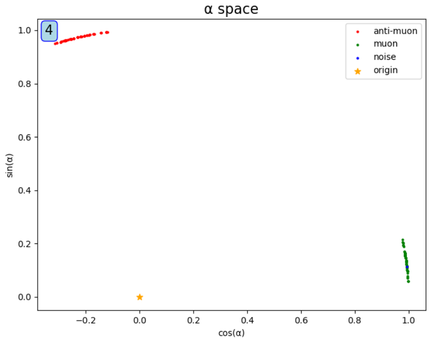
μ+ μ- (particle gun)



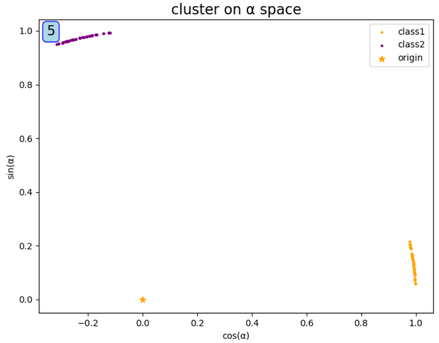
GNN noise filtering



Transform space



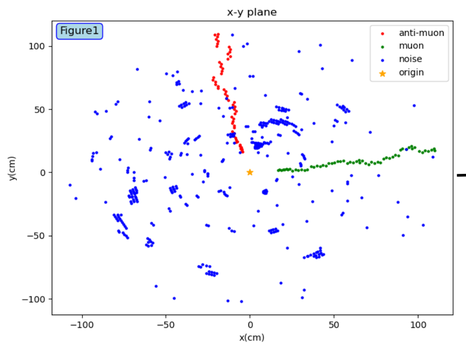
Transform α space



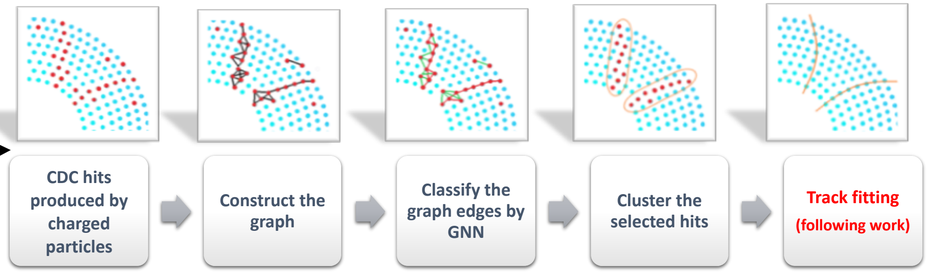
DBSCAN clustering

Acceleration on Versal ACAP platform

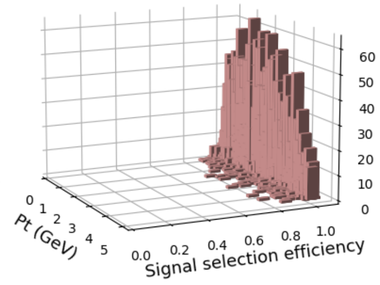
μ+ μ- (particle gun)



Belle II simulation (own work)

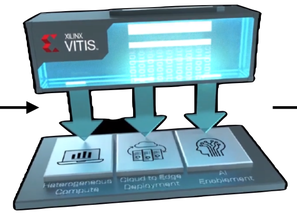
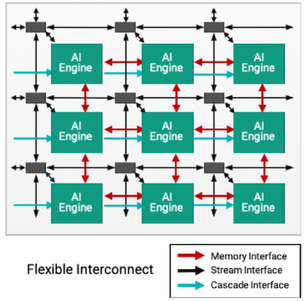


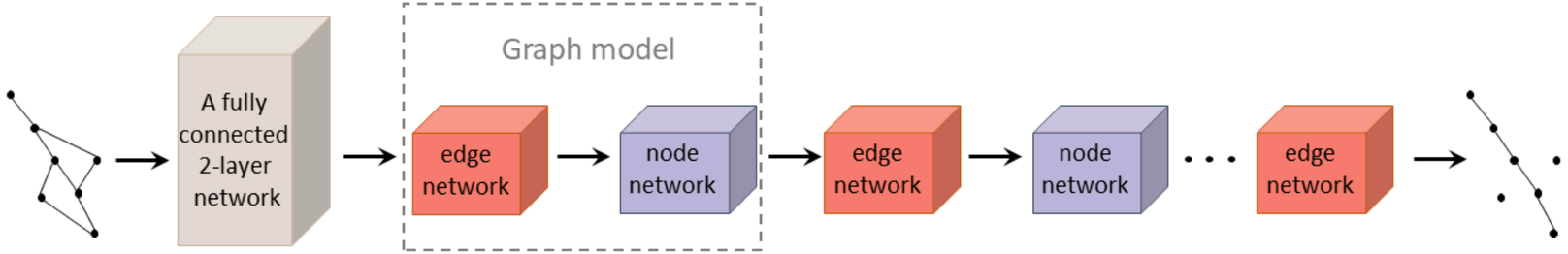
CDC hits produced by charged particles → Construct the graph → Classify the graph edges by GNN → Cluster the selected hits → Track fitting (following work)

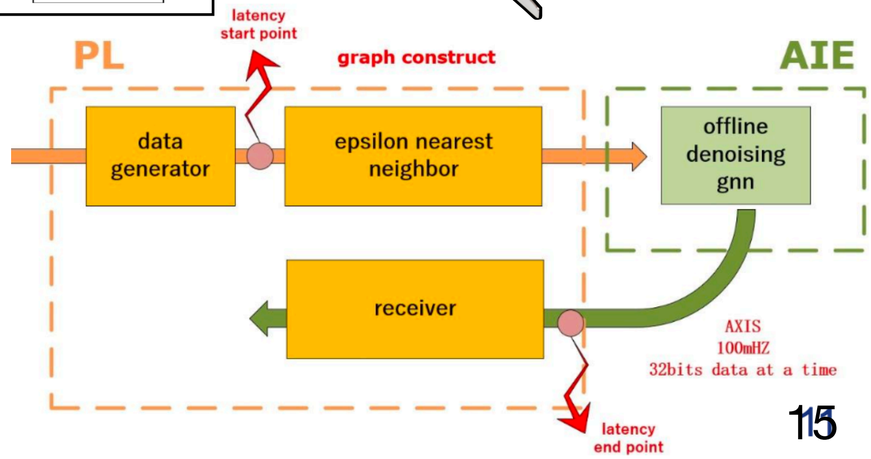


Hit selection efficiency: 98.4%
Hit selection purity : 97.9%

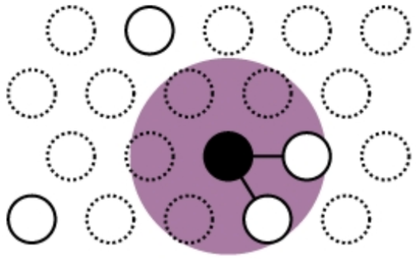
AI engine

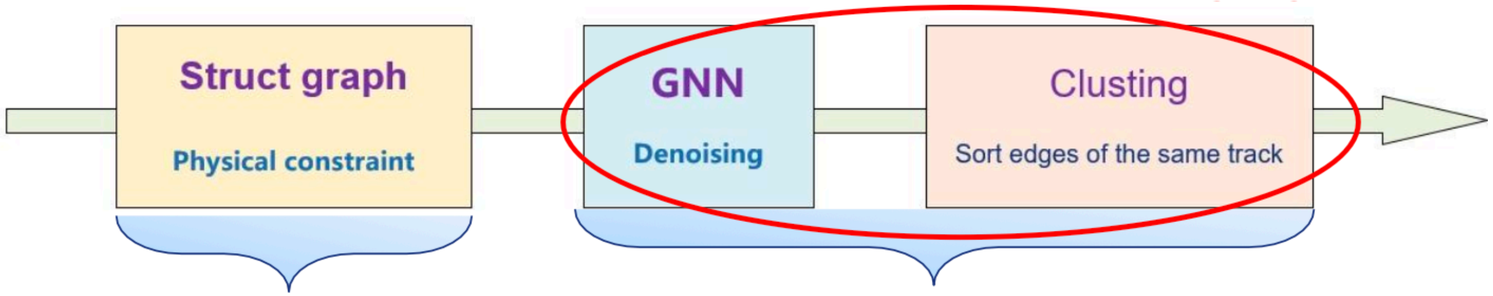
C++ software →  → VIVADO™ → 





GNN implementation on Versal ACAP

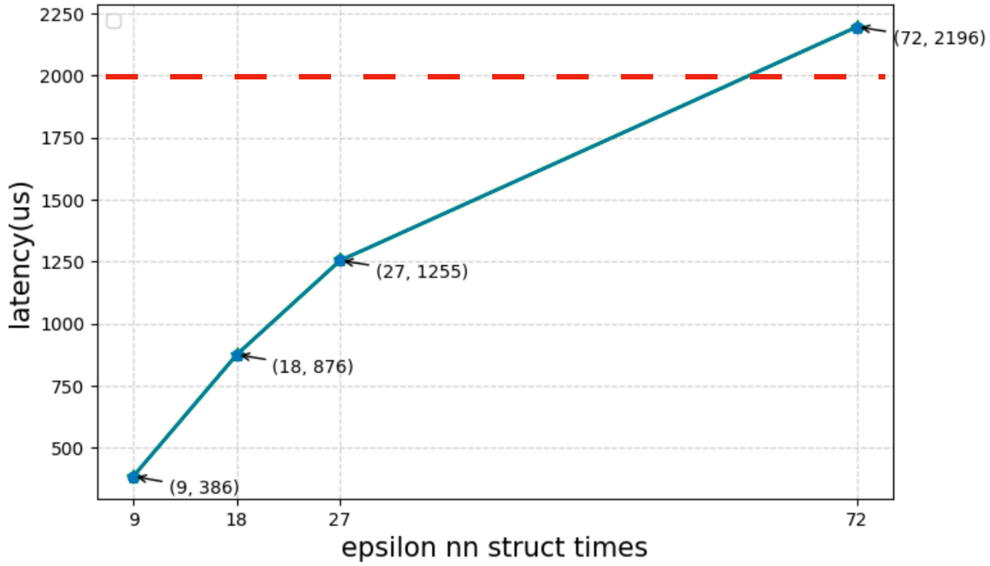


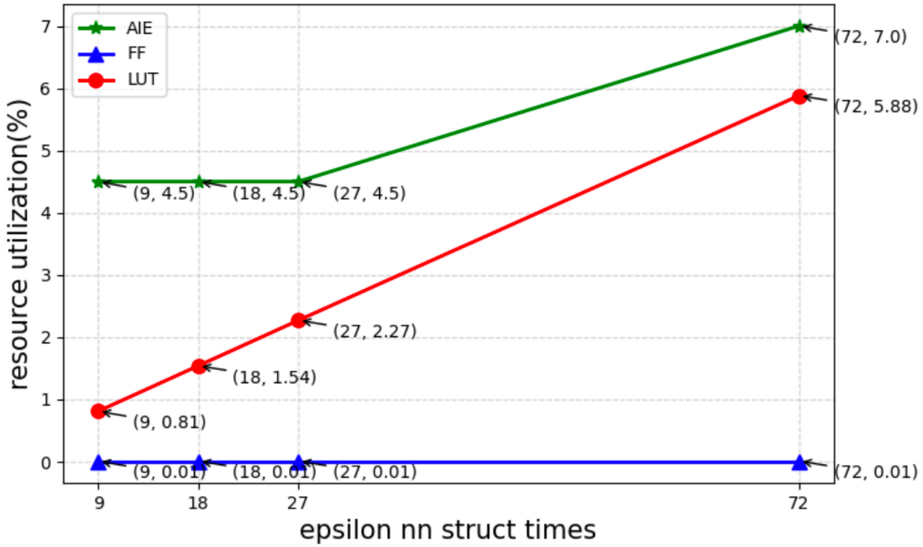


VIVADO

Vitis™

2 ms





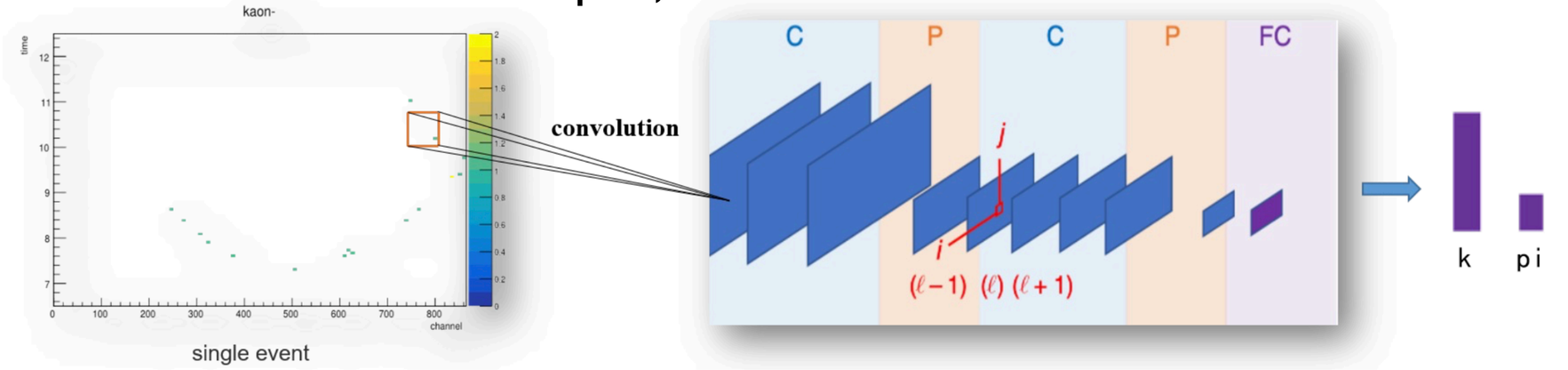
Latency of simple implementation with only 1 iteration of GNN algorithm on Versal is in order of **~ms**, shows GNN can be used for online data processing

Machine Learning for HLT or Offline Reconstruction (SFOTWARE)

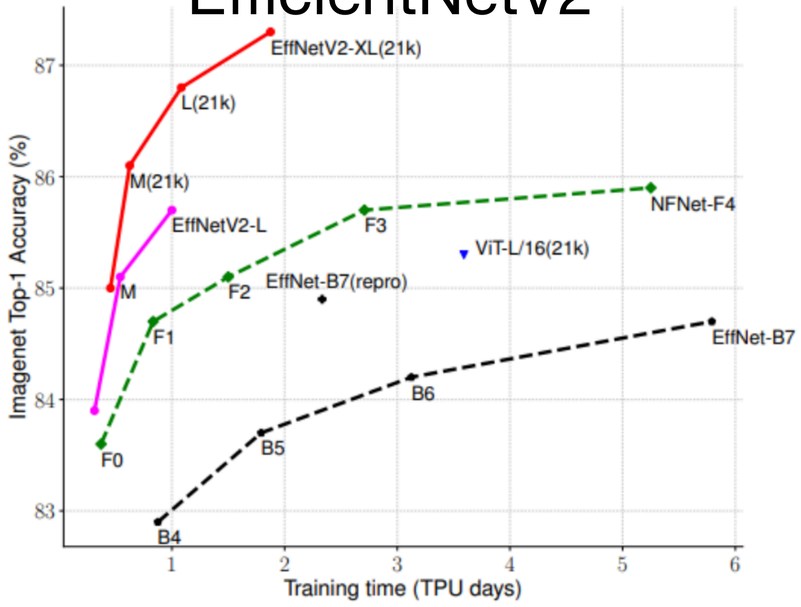
CNN algorithm for STCF PID

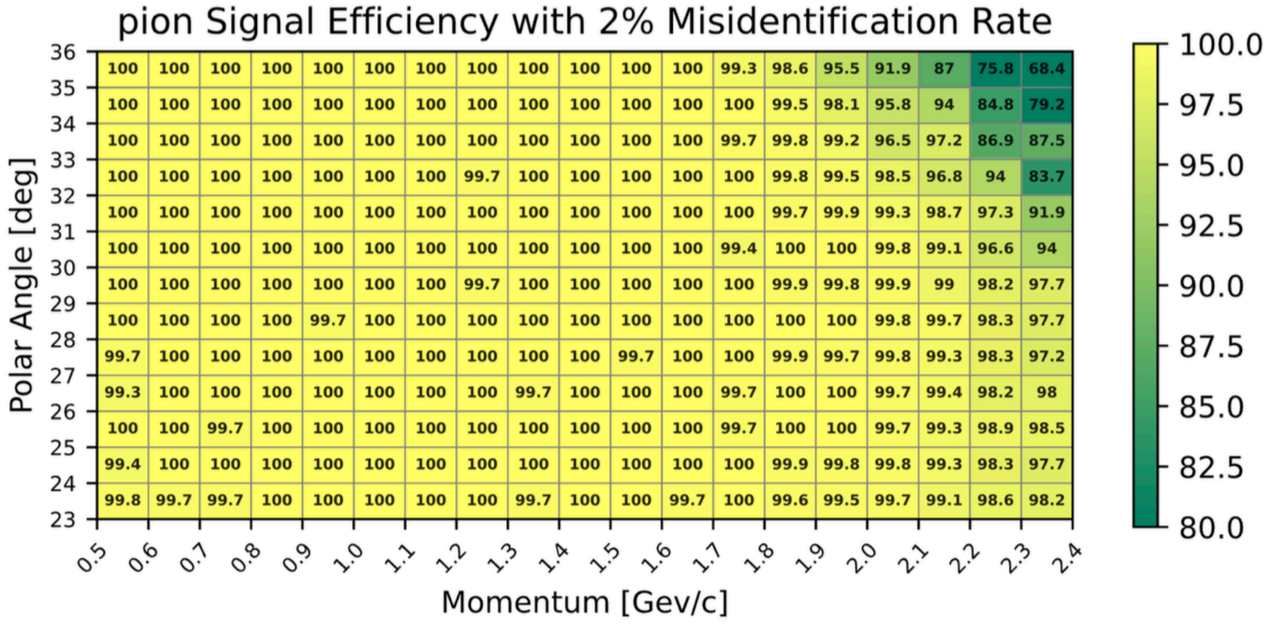
- DTOF as a PID subdetector of STCF
- CNN algorithm developed for Kaon/pion identification
- Kaon/Pion MC simple, 800w

Z. Yao et al.@SDU

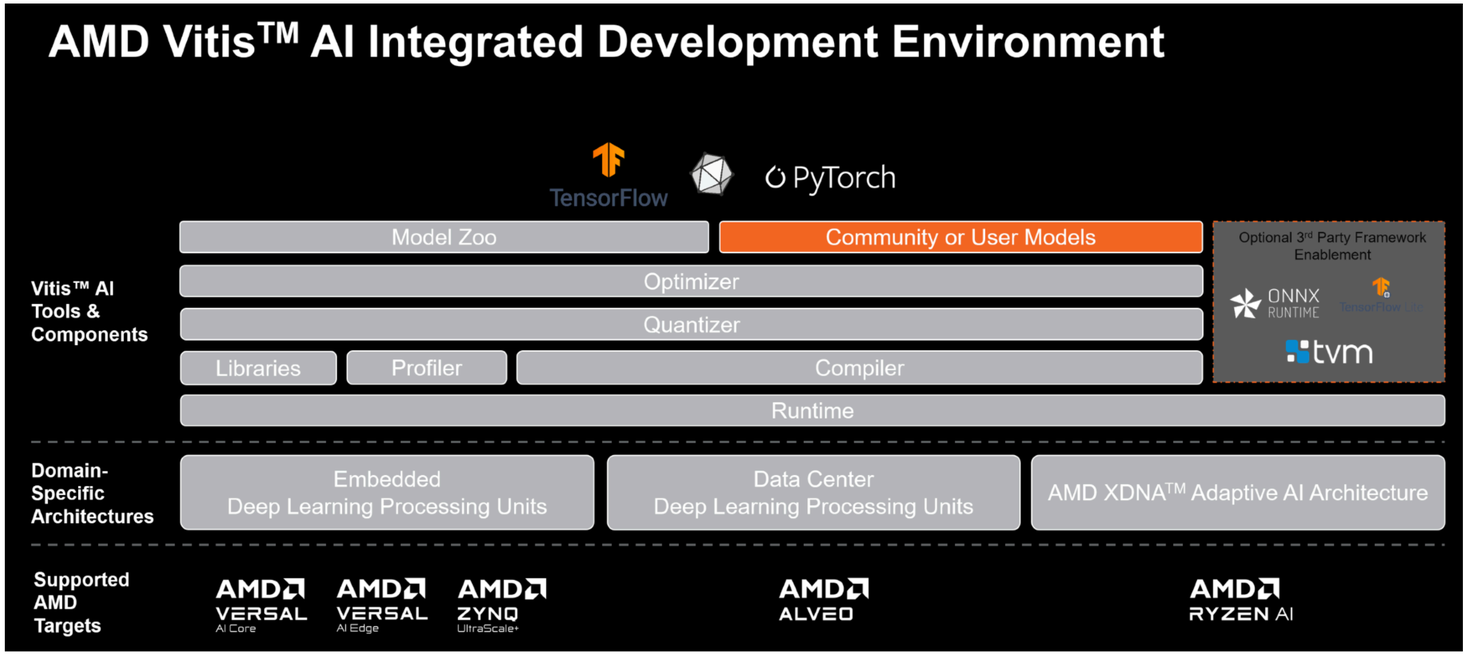


EfficientNetV2





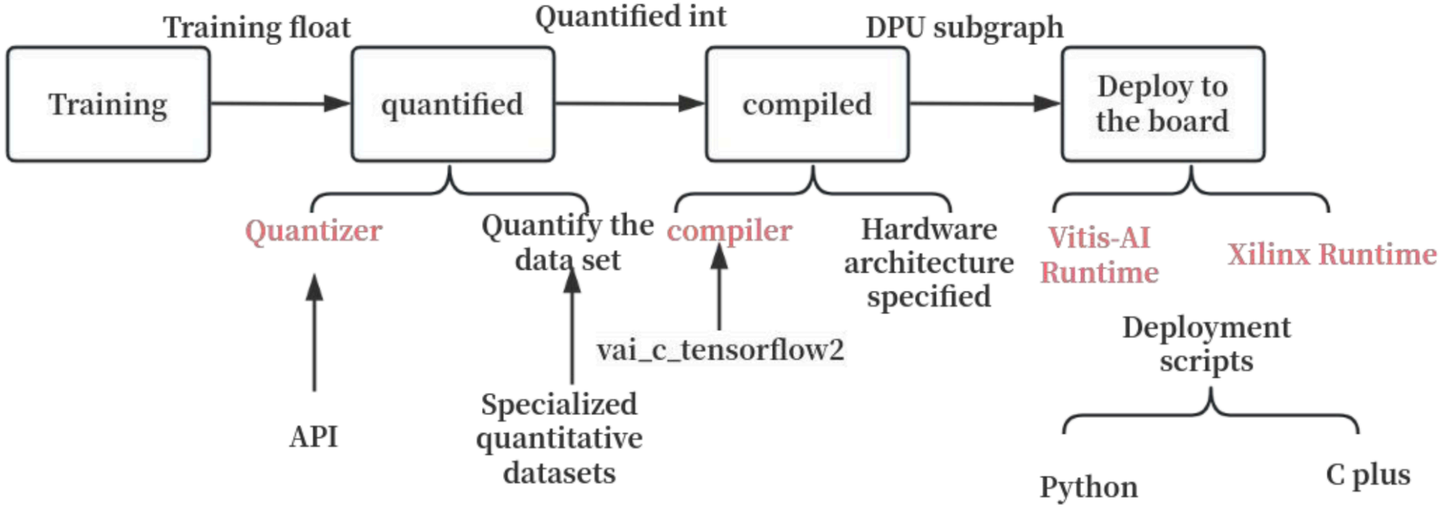
CNN algorithm implementation on DPU

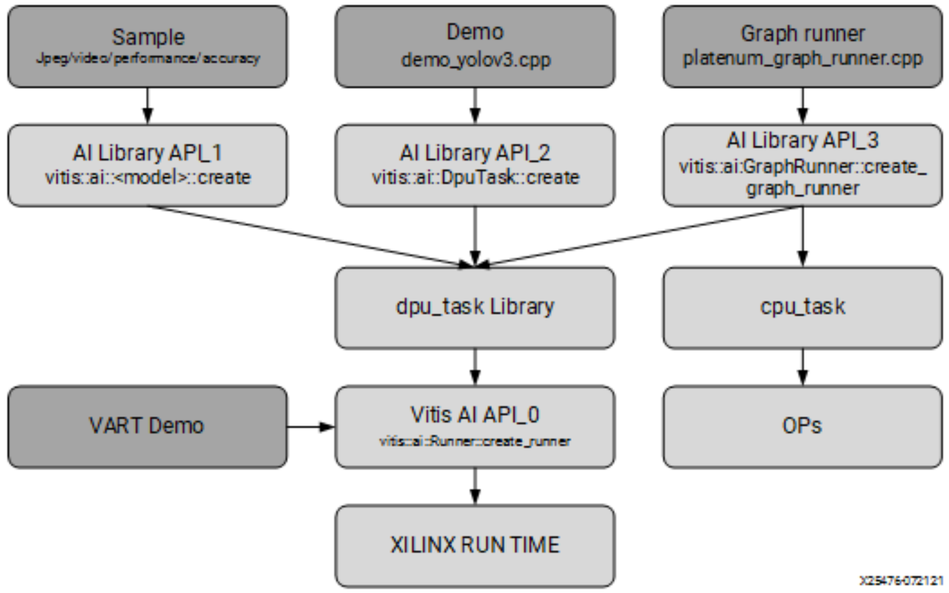


arXiv: 2506.11441

	DPUCZ DX8G[2]	XVDPU [1]		DPUV4E Ours		
Device	ZCU102	VCK190		VCK5000		
Config	B4096 CU3	C32B3	C32B6	6PE+ DWC	6PE	8PE
Freq(MHz)	281	333	300	350	350	350
LUT(K)	160	205	403	631	223	674
FF(K)	315	266	507	648	287	696
BRAM	771	0	678	780	684	912
URAM	0	332	343	402	390	424
DSP	1686	407	809	424	34	524
AIE (MAC)[1]	0 (0)	96 (96)	192 (192)	384 (192)	384 (192)	384 (256)
TOPS	3.7	32.7	61.4	72.5	68.4	91.2

[1] MAC refers to the number of AIEs utilized for performing Conv MAC operations.

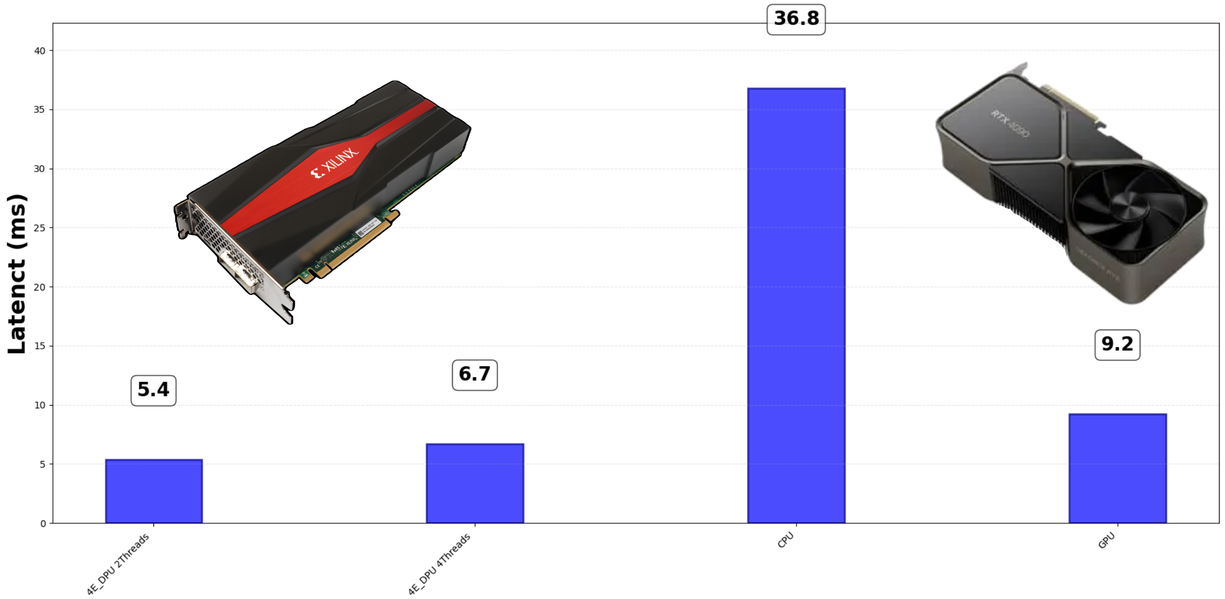




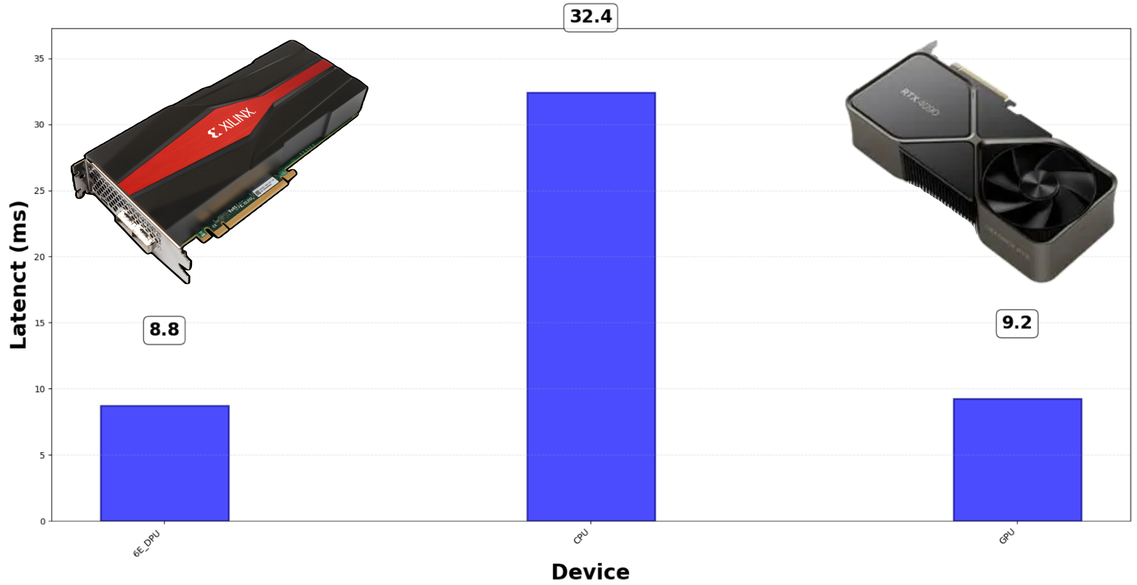
CNN algorithm inference

Inference result based on 10000 sample

- Batch size 4 or 6 due to DPU PE limitation



(VCK5000 4PE) (CPU) (RTX4090)



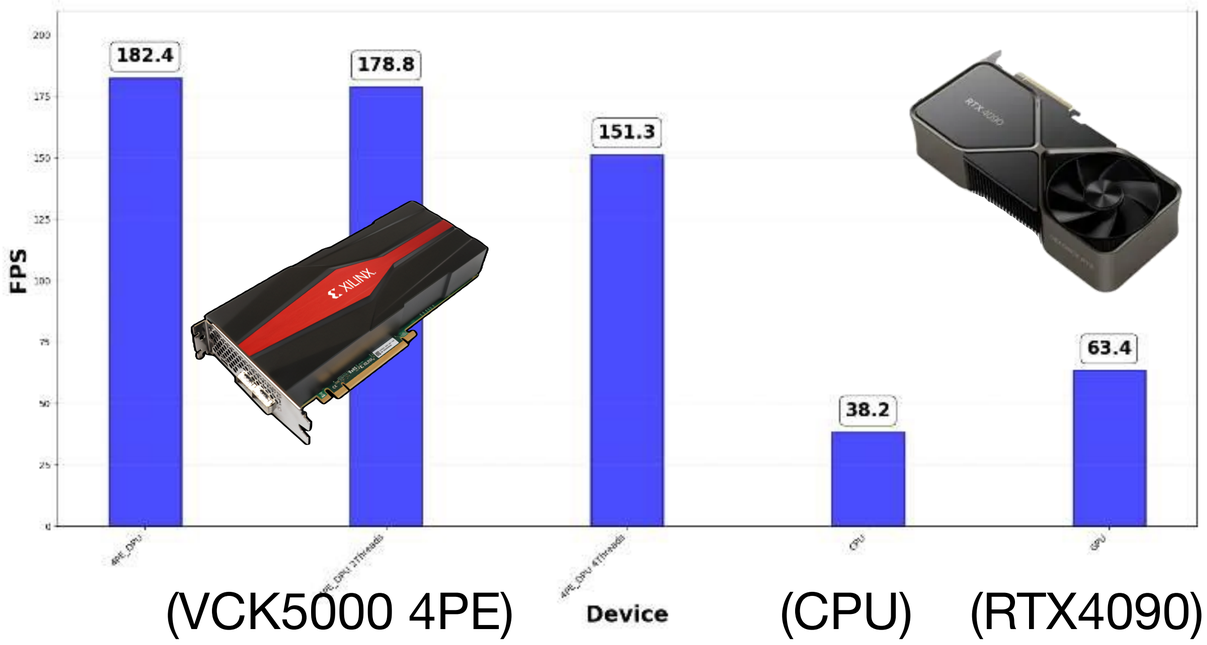
(VCK5000 6PE) (CPU) (RTX4090)

- DPU based on AMD Versal ACAP shows ~7 times(CPU)/~2(GPU) faster inference time
  - GPU RTX4090 should have better performance with increasing batch number
  - RTX4090 and cuda córę is fully optimized for EfficientNet model
  - VCK5000 is designed for data center (good at power cost efficient)

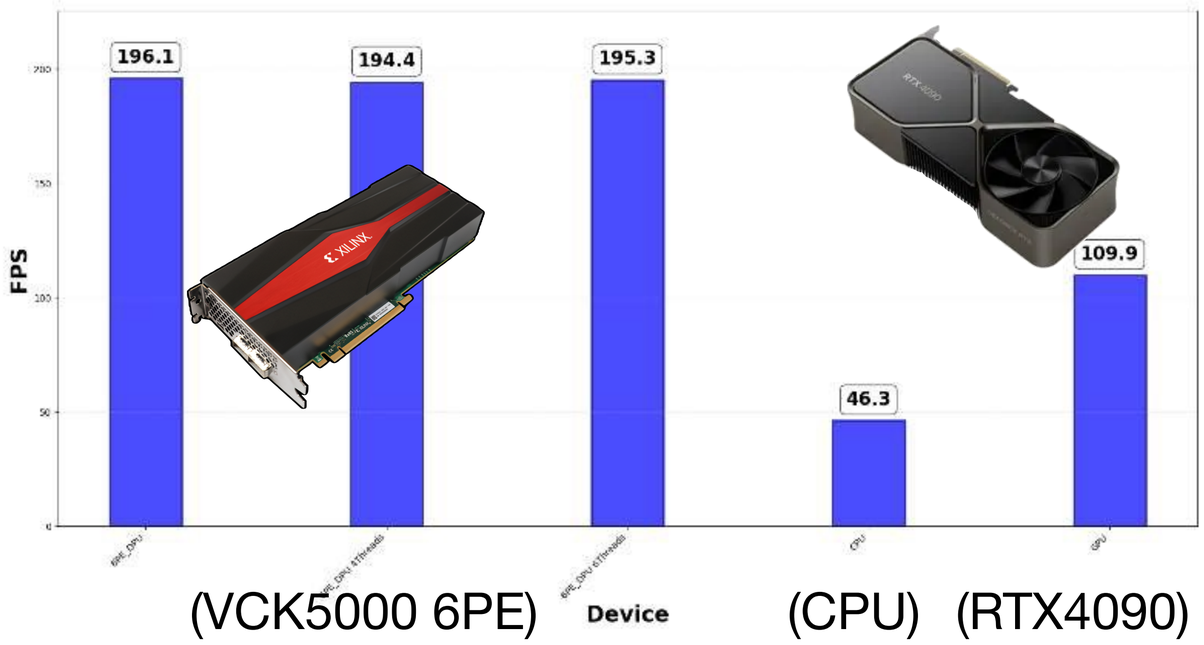
CNN algorithm inference

Inference result based on 10000 sample

- Batch size 4 or 6 due to DPU PE limitation



(VCK5000 4PE) (CPU) (RTX4090)



(VCK5000 6PE) (CPU) (RTX4090)

- DPU based on AMD Versal ACAP shows ~7 times(CPU)/~3(GPU) higher throughput
  - GPU RTX4090 should have better performance with increasing batch number
  - RTX4090 and cuda córę is fully optimized for EfficientNet model
  - VCK5000 is designed for data center (good at power cost efficient)

Summary and conclusion

- Next upgrade plan is necessary that trigger, DAQ and software (basf2, tracking, etc.) experts to get a consensus
- Before LS2 (SVD APV25 buffer is bottleneck)
  - Increasing CPU cores
  - Speed up reconstruction algorithms
  - Developing new algorithms for both L1 and HLT
    - Software Level 3 trigger
- After LS2
  - VXD upgrade is decided, we SHOULD consider the hardware HLT acceleration
  - CPU+GPU should be consider as the ordinary computing resource
  - FPGA+Versal ACAP should be used at specific part at earlier stage of reconstruction
    - e.g. Hit noise filtering, track seed finding, etc.
    - FPGA, Versal ACAP is difficult to use that need to load database (geometry, etc.)
  - The software platform is necessary when using heterogeneous computing for HLT