Status of Versal project

Yun-Tsung Lai

KEK IPNS

ytlai@post.kek.jp

50th B2GM TRG session

24th Feb., 2025

Outline

- In this slide, I will briefly go through the progress of our study/development.
  - No technical details.

- Versal: Mainly about ML in AI engine
  - Hardware integration, and HLT

- HLS, ML, AIE general project: FINN

- Students' works

- Summary

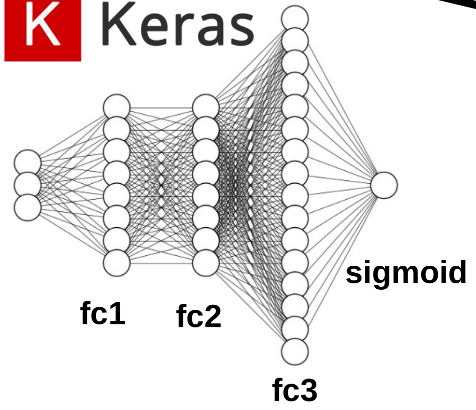
ML in AIE

- We finally have some basic understanding about this implementation.
- Material has been prepared.

- C++ programmable with single-precision floating point.
- No quantization loss, so the same performance as Keras model.
- Latency: 3.4 μs

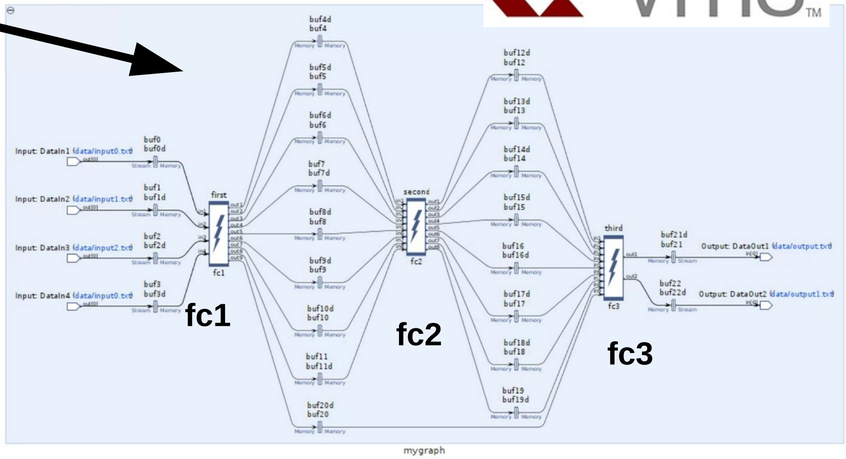
TensorFlow

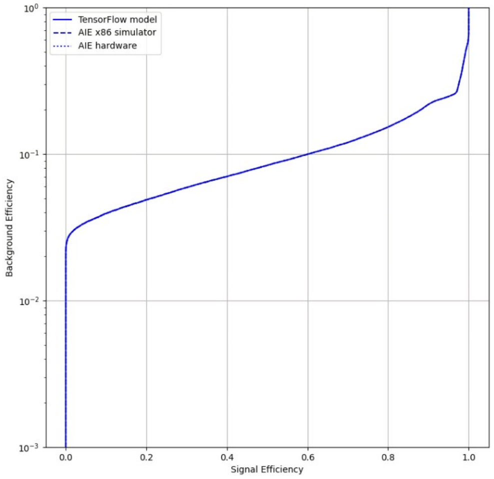
Keras



Exact math form
in C++ in float

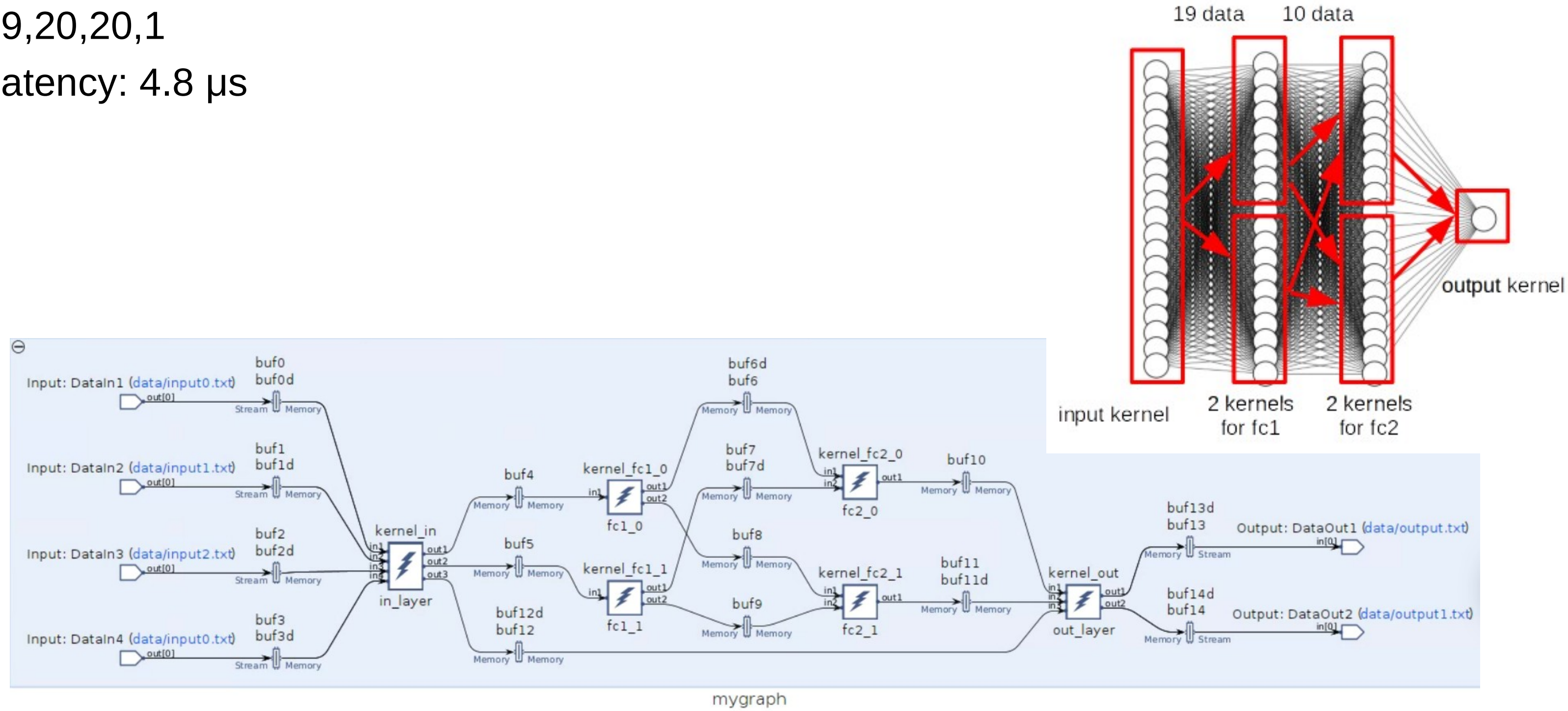
XILINX VITIS™

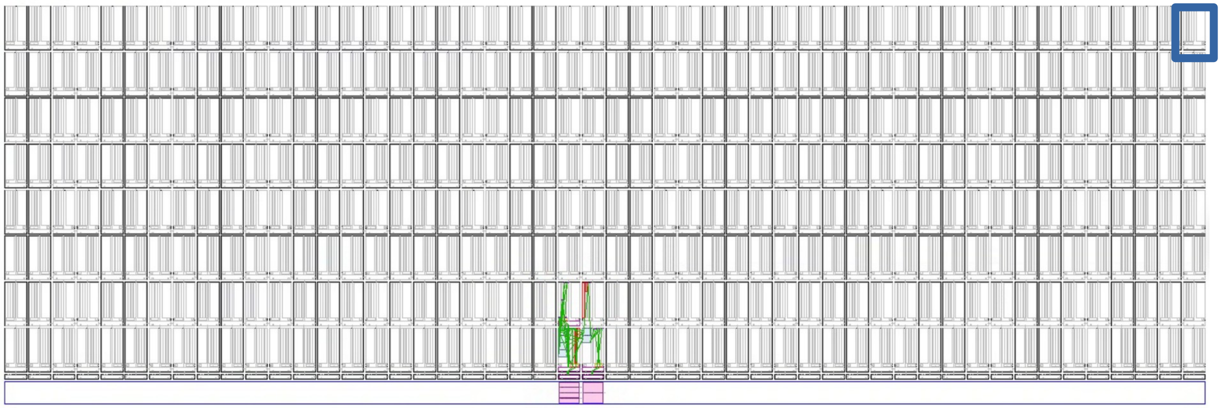




ML in AIE: GRL tau NN

- Use the pre-tained network by Nomaru-san, then implement the mathmatic formula in AIE.
- 19,20,20,1
- Latency: 4.8 μs



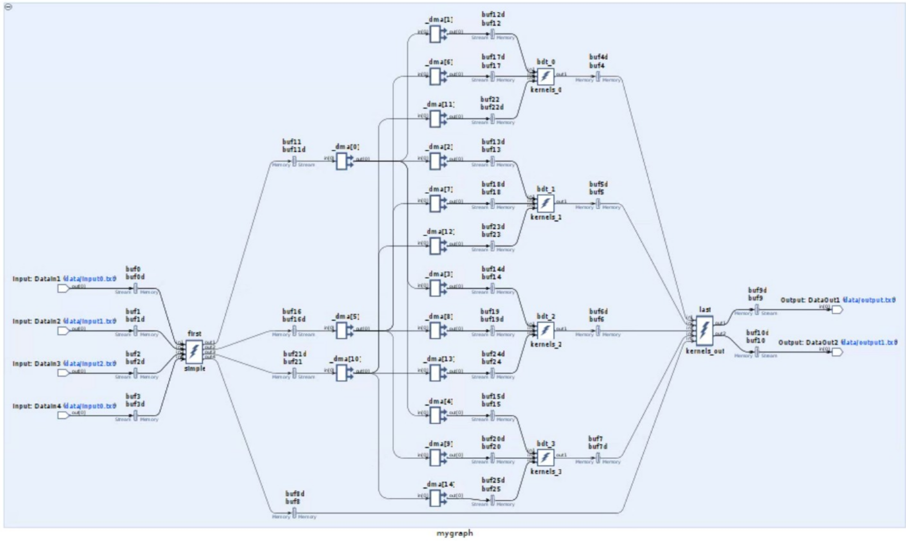


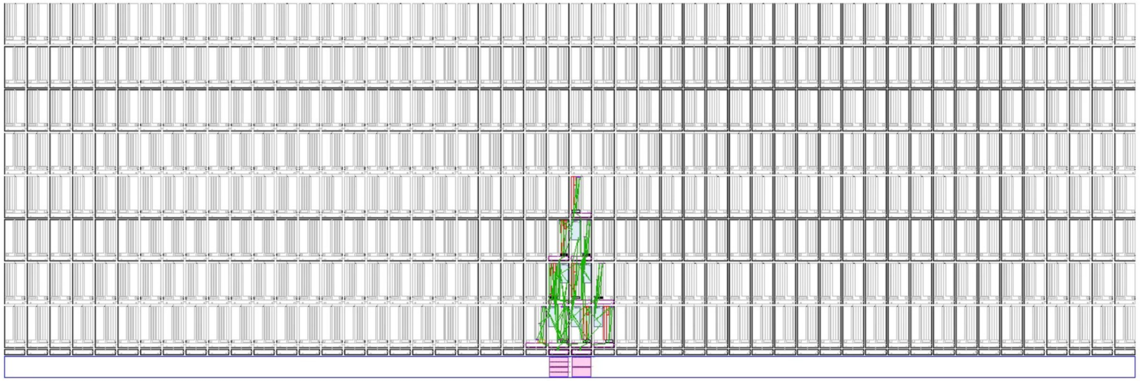
1 block = 1 "tile",

1 kernel in 1 tile at most

ML in AIE: BDT

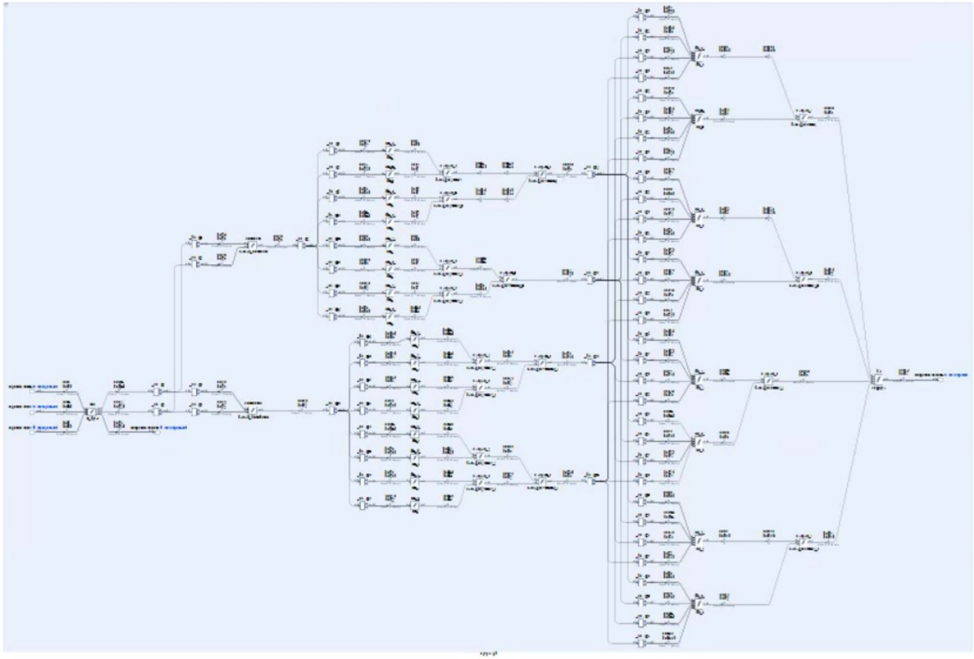
- BDT: Basically a large nested structure of if-else
- Using scikit-learn for model building. N_estimator = 10, depth = 3.
- Parallel kernels for separated estimators, then sum over all the outputs.
- Latency: 2.8 μs

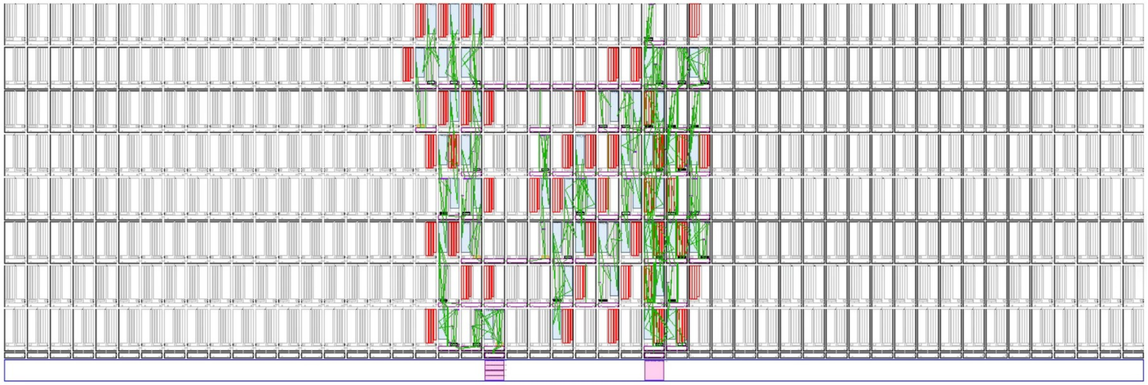




ML in AIE: KLMTRG NN

- Use the pre-tained network by Anthony, then implement the mathmatic formula in AIE.
- 8,64,16,3
  - Hidden layers use tanh.
    Output layer uses softmax.
- Complicated design!
- Latency: 10 μs





ML in AIE: KLMTRG NN (cont'd)

- The reason of the complexity: **exp() in math.h**

- In AIE design using Vitis, the regular C++ library (e.g. math.h) can be directly called.
  - However, for those can be simply written in CPU, they could be expensive and slow in AIE.

- tanh and softmax contains many exp().
  - 1 AIE tile (1 kernel) sometimes can contain only few exp().

- Workaround:
  - Use aprroximation for special function
  - Use LUT depending on the domain of input

ML in AIE: More to be studied

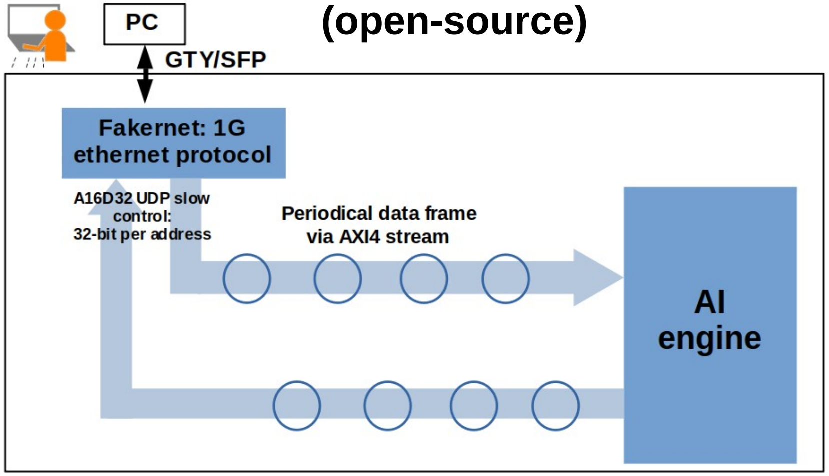
- Although we tried to learn about the basic utilization, there are still lots to be studied for AIE.
- I/O interface: Stream type v.s. buffer type
- Scaler v.s. vector
- MLIR-AIE and the IRON frameworks

- Students can explore them in depth once they move to relavant development with AIE.
  - WIll consult with Marc and Zhaozhi for their experience.

Hardware integration: AIE + PCIe or Ethernet

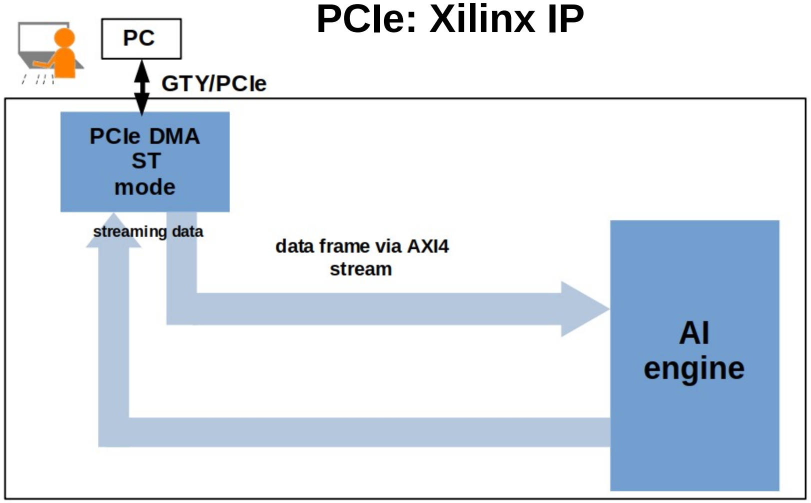
- We need to integrate FPGA + AIE with PC with efficient options for communication.
  - We tested with Ethernet link and PCIe for demonbstration.

Ethernet: Fakernet (open-source)

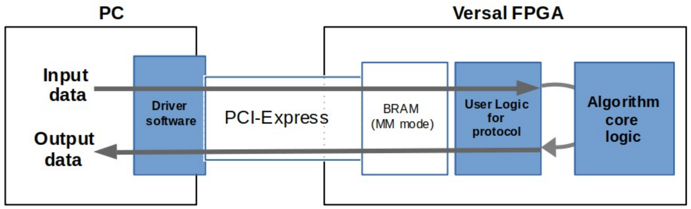

PC
GTY/SFP
Fakernet: 1G ethernet protocol
A16D32 UDP slow control:
32-bit per address
Periodical data frame via AXI4 stream
AI engine

- Support 1G and 2.5G
- GTY transceiver with optical SPF at FPGA, NIC at PC
- 1.5 hrs for 200,000 events
- Option for 10G is also discussed, but need to find good ptotocol

PCIe: Xilinx IP



- Self-defined protocol for data exchange.
- 50 min for 200,000 events.

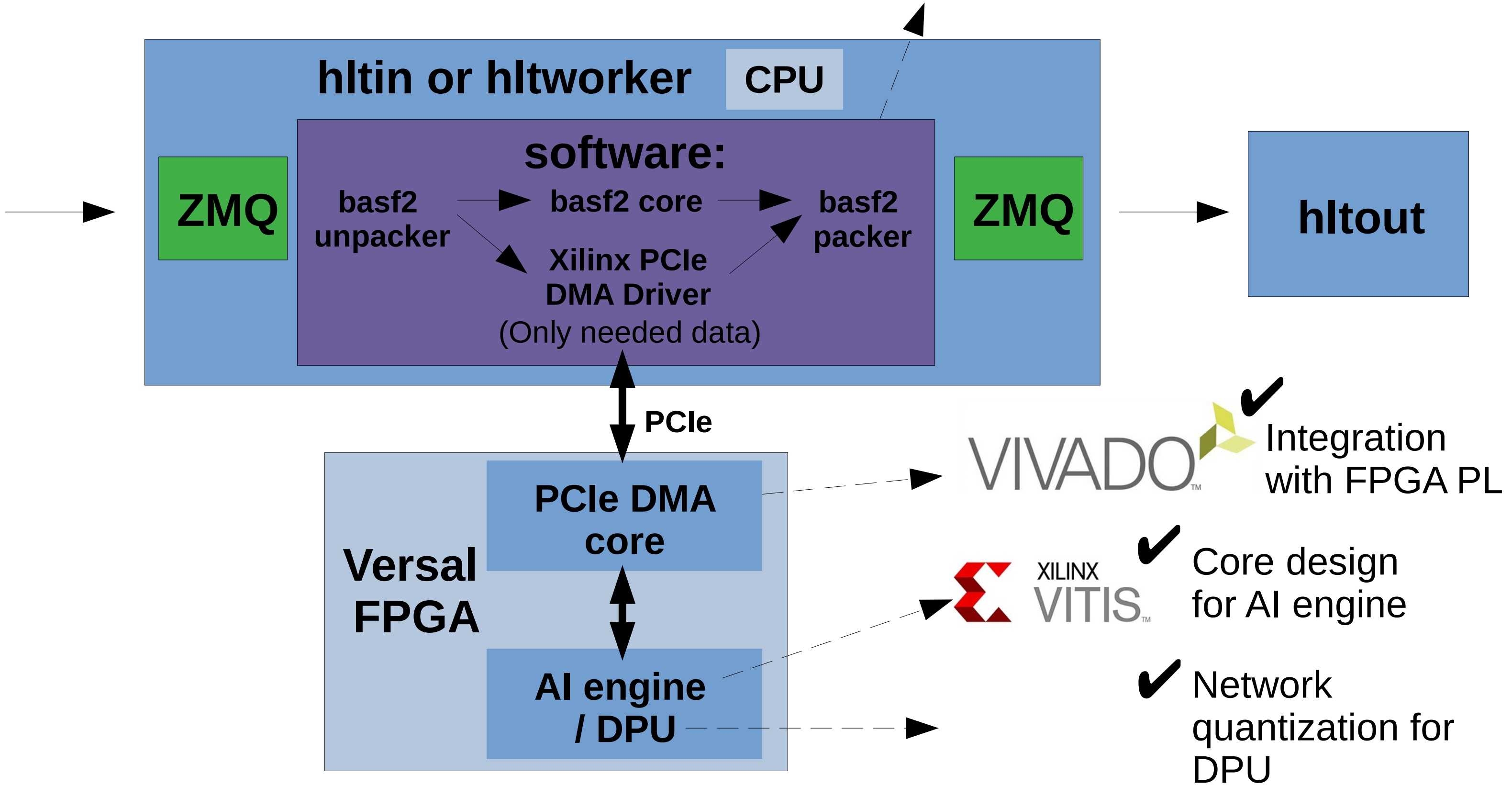


→ **Potential for HLT application.**

Plan for Belle II HLT

- For Belle II HLT, we are planning to design such kind of integrated framework in order to use "Versal in Belle II HLT".

✔ Software: integration in HLT data flow
basf2 unpacker, then transfer the needed event data to Xilinx PCIe driver to FPGA, get back the returned output, and pack the output to be sent to storage.



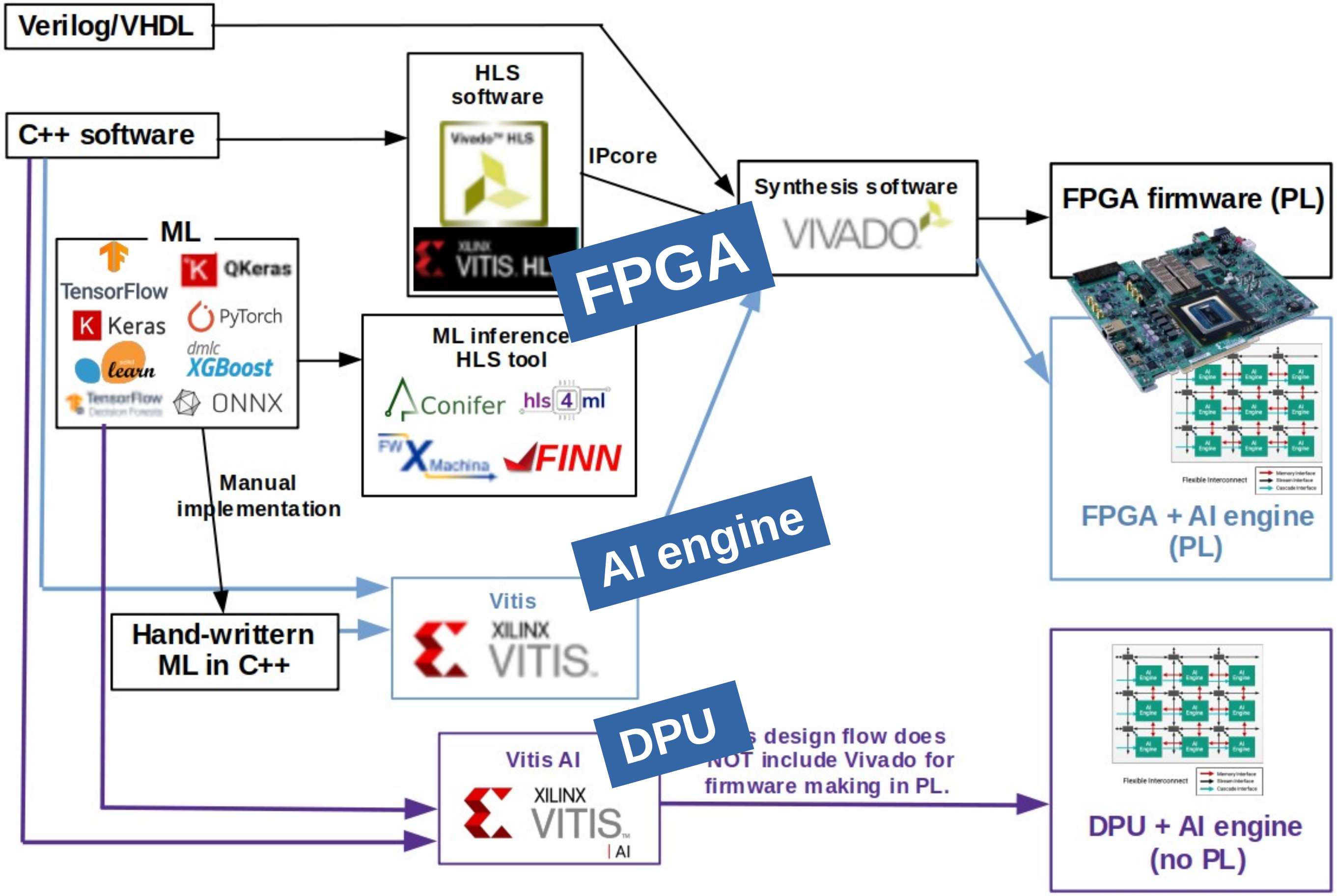
✔ Integration with FPGA PL

✔ Core design for AI engine

✔ Network quantization for DPU

HLS, ML, AI engine: roadmap of FPGA methodology

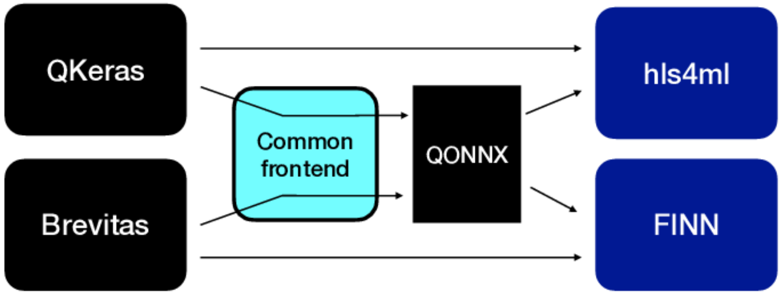
- We are almost done!
- Hand-on lectures are under preparation in 2025 fiscal year.
  - ~20 people.
  - For people in interst, collaboration is very welcome for local environment setup beforhand.



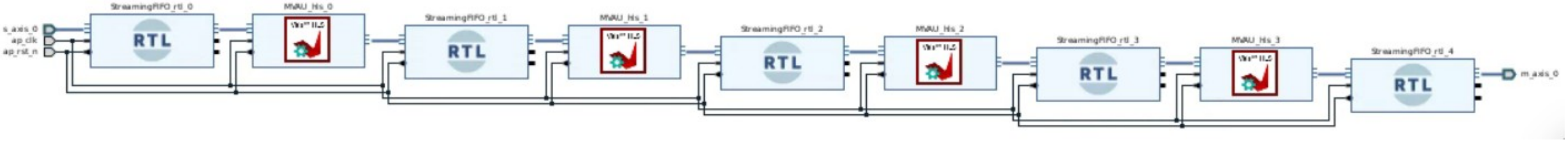
New study: FINN

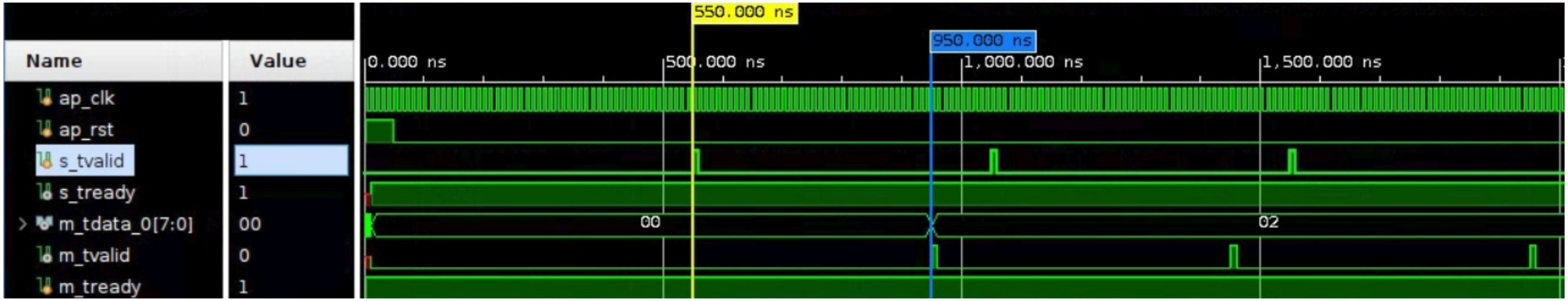
- Under development by AMD Xilinx.
- The core concept is matrix multiplication.
- Quantization based on Pytorch + Brevitas.
- Model representation by ONNX/QONNX.

- Material is ready.
- Will also encourage students to use it for our ongoing development in TRG.



source: 10.48550/arXiv.2206.11791





Name	Value
ap_clk	1
ap_rst	0
s_tvalid	1
s_tready	1
m_tdata_0[7:0]	00
m_tvalid	0
m_tready	1

Student activities

- Anthony Little (Sydney):
  - Already finished his work in KLMTRG NN
  - Imlpementaion and validation in UT4: Need new personpower or I will do it.
- Yang Yi (Fedan):
  - Now working on commissioning of DNN tracking
  - Plan: hls4ml/FINN for UT4/UT5 improvement, and also AI engine.
- Yongheon Ahn (KU):
  - Now working on "GRL tau BDT" with Conifer
  - Plan: BDT implementation in AI engine
- Junhyeok Song (KU):
  - Now working on a linear fitter design with HLS tool
  - Plan: fitter implementation in AI engine
- Ming-Chun Lin (NTU):
  - Now working on basf2 TSIM for 2D fitter
  - Plan: Improve 2D fitter, then use HLS, AI engine, or ML method for 2D Fitter design.

Summary

- Versal: Mainly about ML in AI engine
  - More basic studies on utilization are needed to have more understanding.
  - Hardware integration and plan for HLT

- HLS, ML, AIE general project:
  - FINN package studied
  - Hand-on lectures under planning

- Students' works: Try different options for logic construction and implementation
  - They will give report in TRG weekly meeting once ready.