清華大學
Tsinghua University

PulseDL-II: A system-on-chip neural network accelerator for timing and energy extraction of nuclear detector signals

Pengcheng Ai (Speaker), Zhi Deng, Yi Wang, Hui Gong, Xinchi Ran, Zijian Lang

Department of Engineering Physics, Tsinghua University

8/1/2022

Three Elements of Design Perspectives

- Three independent elements:

A. **Nuclear Electronics**: Readout system and signal features

B. **Neural Network**: Architectural research, network training…

C. **Digital Design**: NN accelerator and system-on-chip scheme

- Overlay of elements:

AB. **Application Training**: NN algorithm research for nuclear signals, selection and optimization of network architecture

BC. **Hardware Mapping**: Accelerator hardware implementation of NN

AC. **System Prototype**: Hardware design in the context of readout system

ABC. **Joint Validation**: Synthesis of the above three

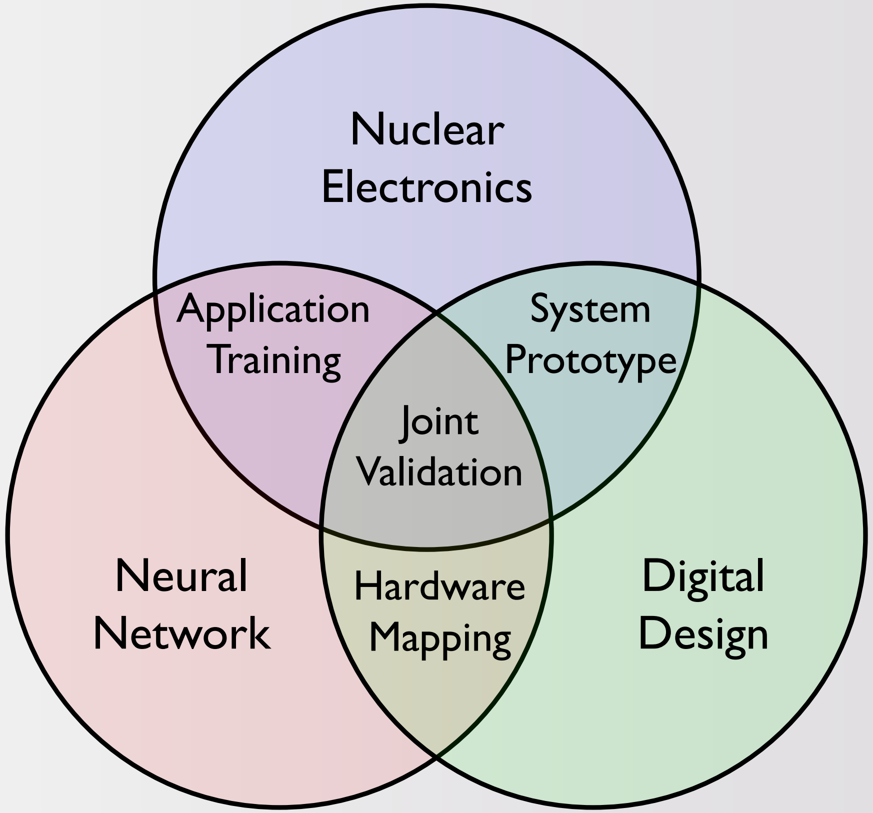


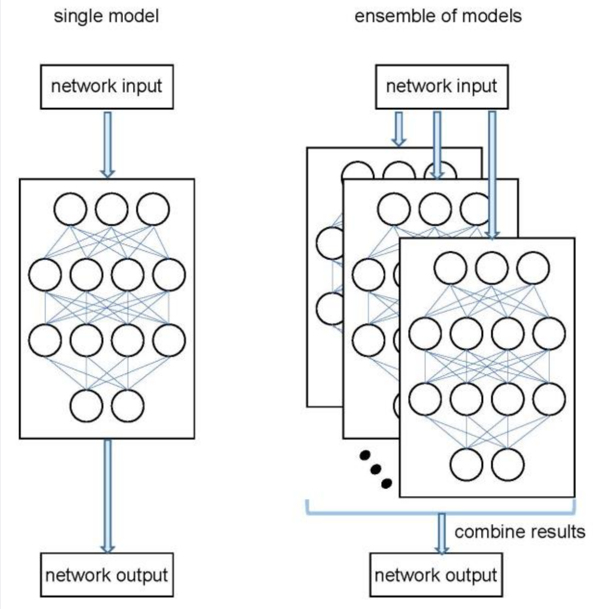
TABLE OF CONTENT

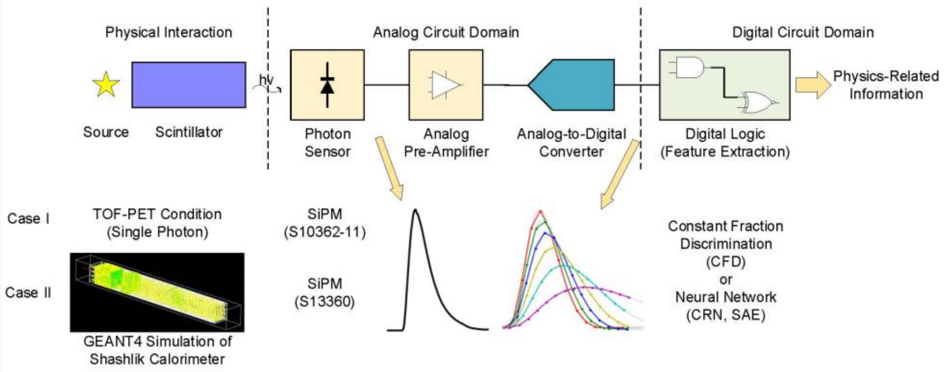
- Signal feature extraction with NN
- NN accelerator-based readout system
- System-on-Chip Accelerator Design
- System Validation
- Summary

What and to What Extent Neural Nets Can Do (AB)

Estimation of heterogeneous uncertainty of nuclear detector signals with ensemble of NNs

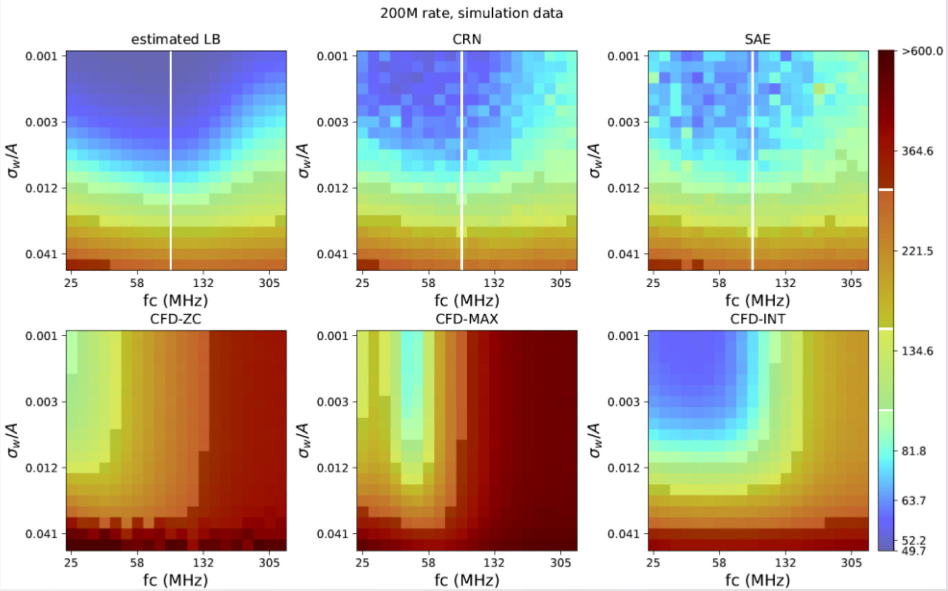
P. Ai *et al* 2022 *JINST* **17** P02032



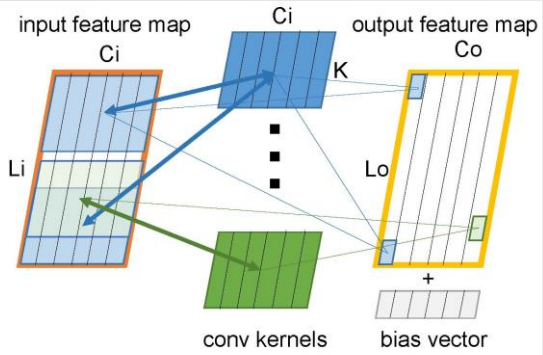

Physical Interaction
Analog Circuit Domain
Digital Circuit Domain
Source
Scintillator
hν
Photon Sensor
Analog Pre-Amplifier
Analog-to-Digital Converter
Digital Logic (Feature Extraction)
Physics-Related Information
Case I
Case II
TOF-PET Condition (Single Photon)
GEANT4 Simulation of Shashlik Calorimeter
SiPM (S10362-11)
SiPM (S13360)
Constant Fraction Discrimination (CFD) or Neural Network (CRN, SAE)

Computation of the Cramer Rao lower bound of timing to find out limits for NN and traditional methods

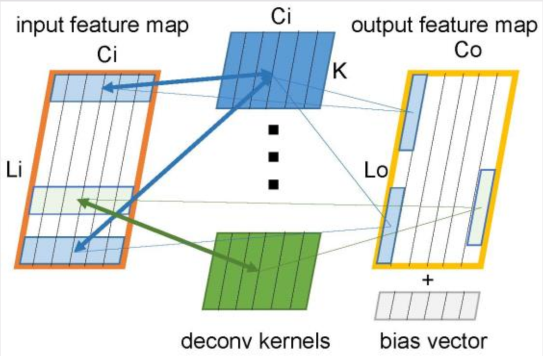
P. Ai *et al* 2021 *JINST* **16** P09019



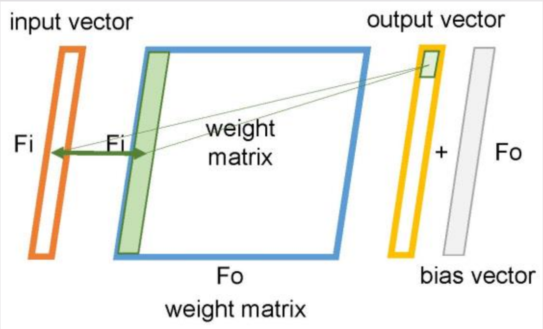
Building Blocks of Network Structure (and Why They Work) (AB)



one-dimensional convolution



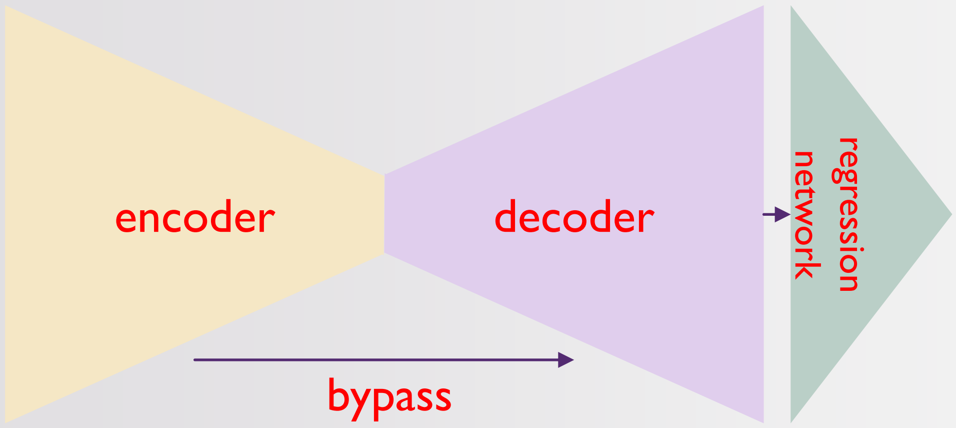
one-dimensional deconvolution



fully-connected matrix multiplication

- We choose Convolutional Neural Networks (CNN) because they succeeded in many ML tasks and facilitated parallel computing
- We select four representative building blocks:
  - 1d convolution layer
  - 1d deconvolution layer
  - fully-connected layer
  - nonlinear activation (ReLU)
- Nonlinearity is the key for **Inductive Learning** (and thus intelligent signal processing) J.C.Ye (2022) Geometry of Deep Learning, Springer
  - Without nonlinearity, the weights in the mapping function are the same for any input sample. Once learned, they never change. (transductive)
  - With nonlinearity, weights in the mapping function are selectively turned off/scaled by nonlinear function. (inductive)

Autoencoder-Based Network Architecture (AB)

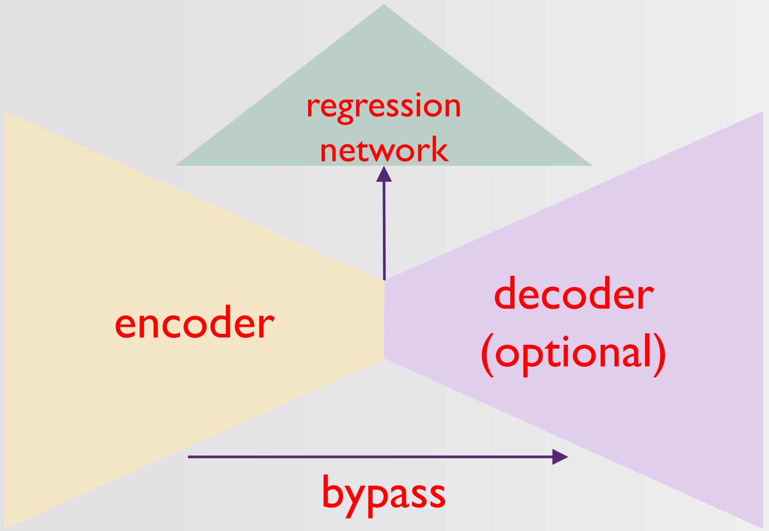


encoder

decoder

regression network

bypass



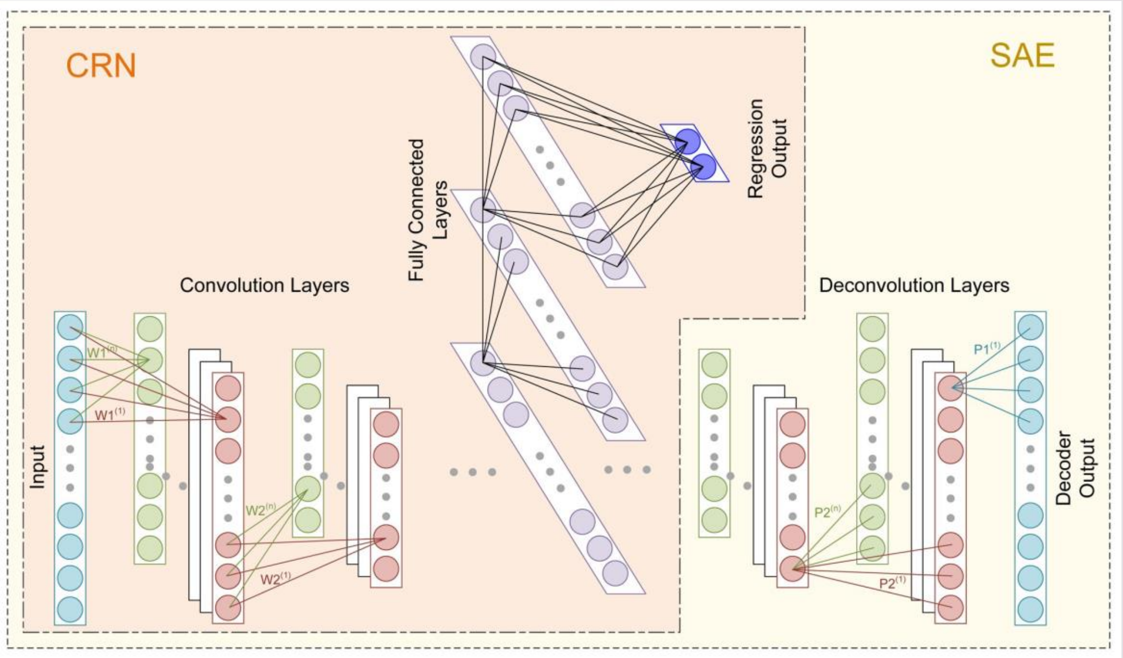
regression network

encoder

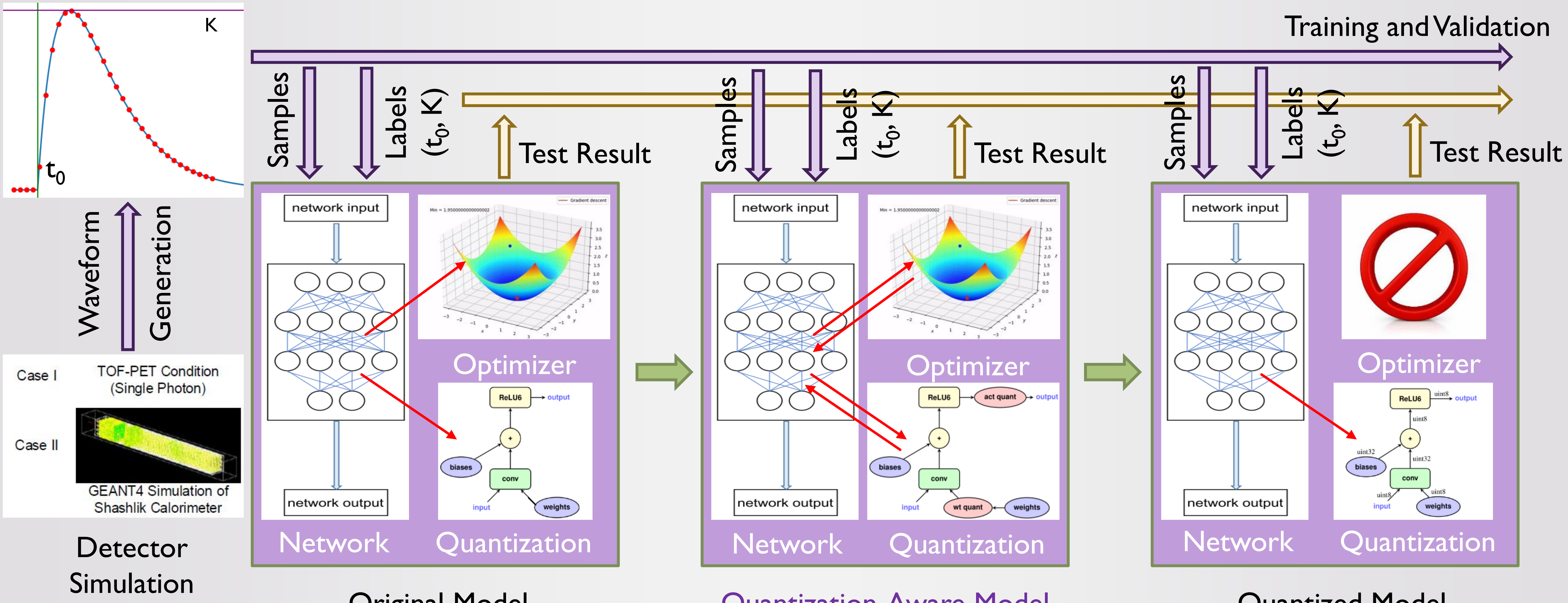
decoder (optional)

bypass

- Regression network can be located at the far-end of the decoder if an accurate noiseless waveform can be obtained.
- Regression network can also be located at the bottleneck if we only have original waveform (and the decoder is optional).



Quantization-Aware Training and Validation (AB)



Original Model
(infer with floating number)

Quantization-Aware Model
(update step and range)

Quantized Model
(infer with integer/fixed number)

Why Bringing Them to Front-End Electronics (AC)

- A case study of electromagnetic calorimeter (ECAL) of NICA-MPD
  - 64-channel, 12-bit, 62.5M-rate ADC
  - Waveform data readout, triggering & timing by optical fiber
  - power consumption: 250 mW/channel, water cooling system is needed for heat dissipation

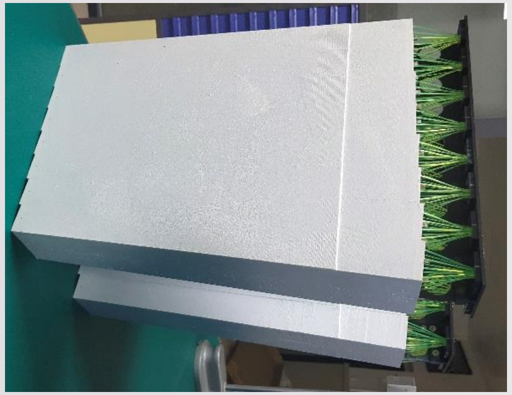
Waveform Sampling

64-channel ADC board

Fiber Transmission

Event Reconstruction

Cable Connection

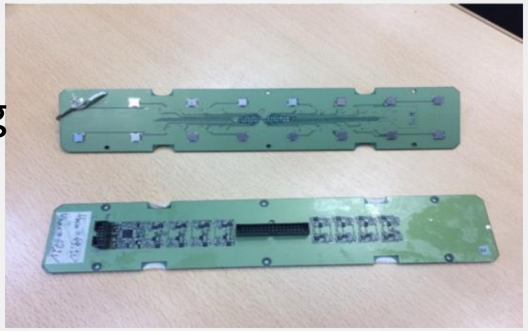


Shashlik Calorimeter

Coupling

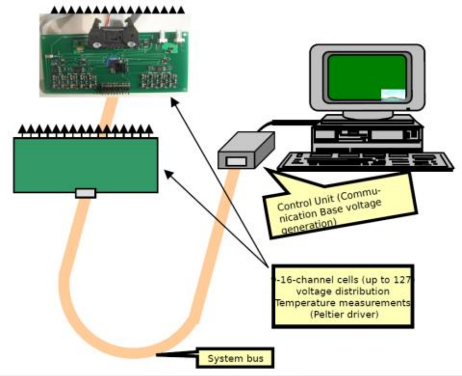
SiPM

Soldering



Front-End PCB

Bias Voltage

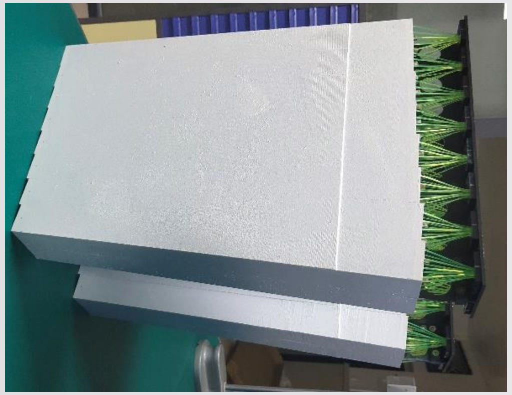

Control Unit (Communication Base voltage generation)
-16-channel cells (up to 127 voltage distribution Temperature measurements (Peltier driver)
System bus

High Voltage System

Why Bringing Them to Front-End Electronics (AC)

- A case study of electromagnetic calorimeter (ECAL) of NICA-MPD
  - 64-channel, 12-bit, 62.5M-rate ADC
  - Waveform data readout, triggering & timing by optical fiber
  - power consumption: 250 mW/channel, water cooling system is needed for heat dissipation
- Front-end upgradation with ASIC

pre-amplifier, 200M-rate ADC & NN accelerator

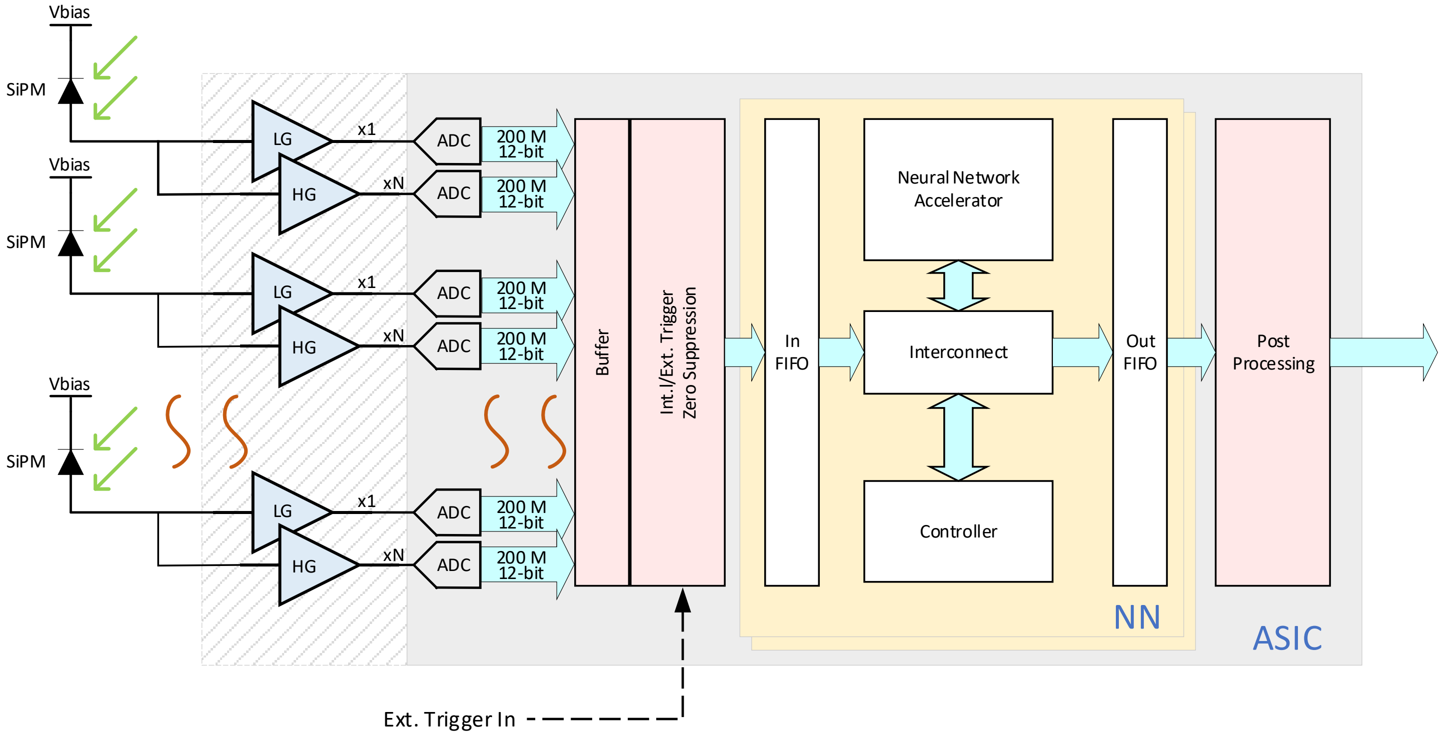


Shashlik Calorimeter

Coupling

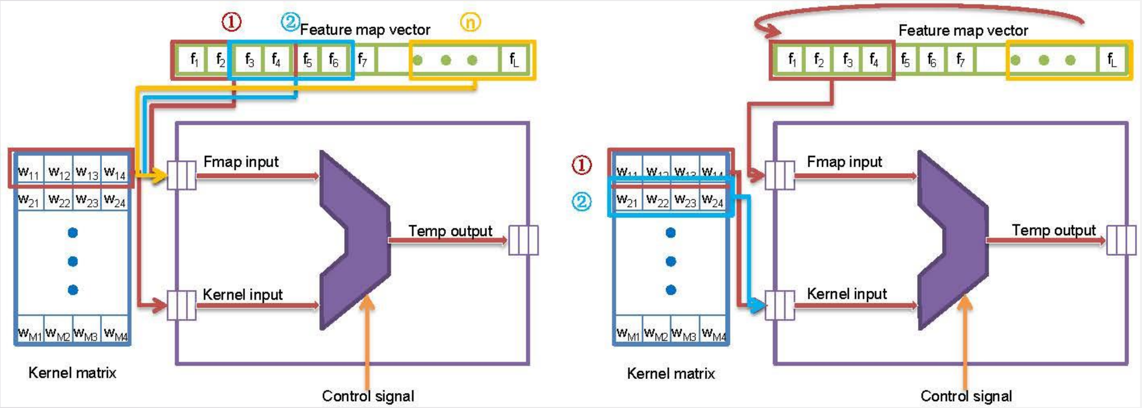
SiPM

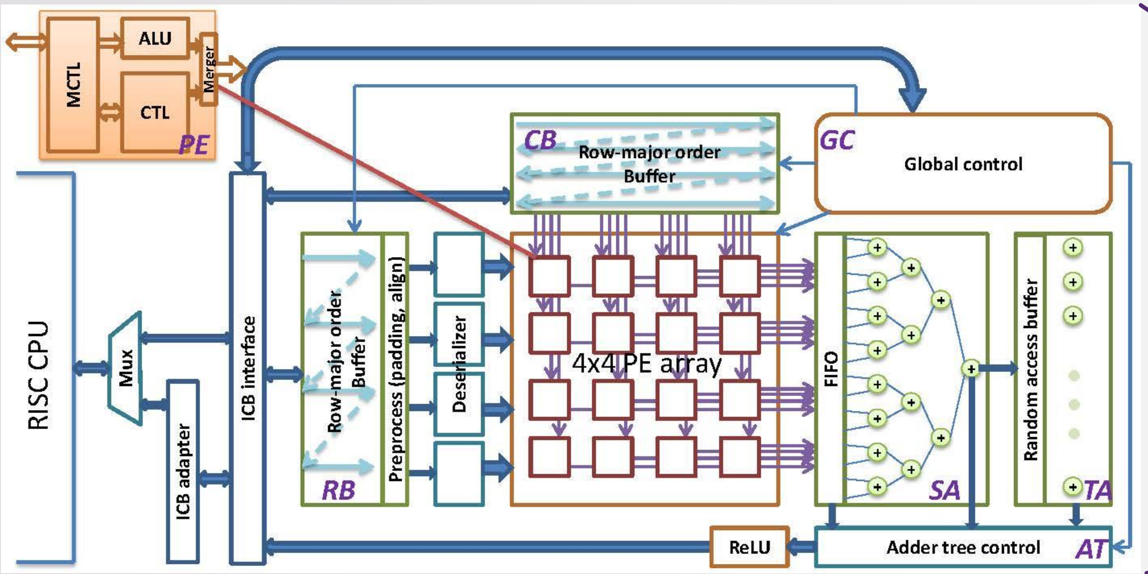
Sold

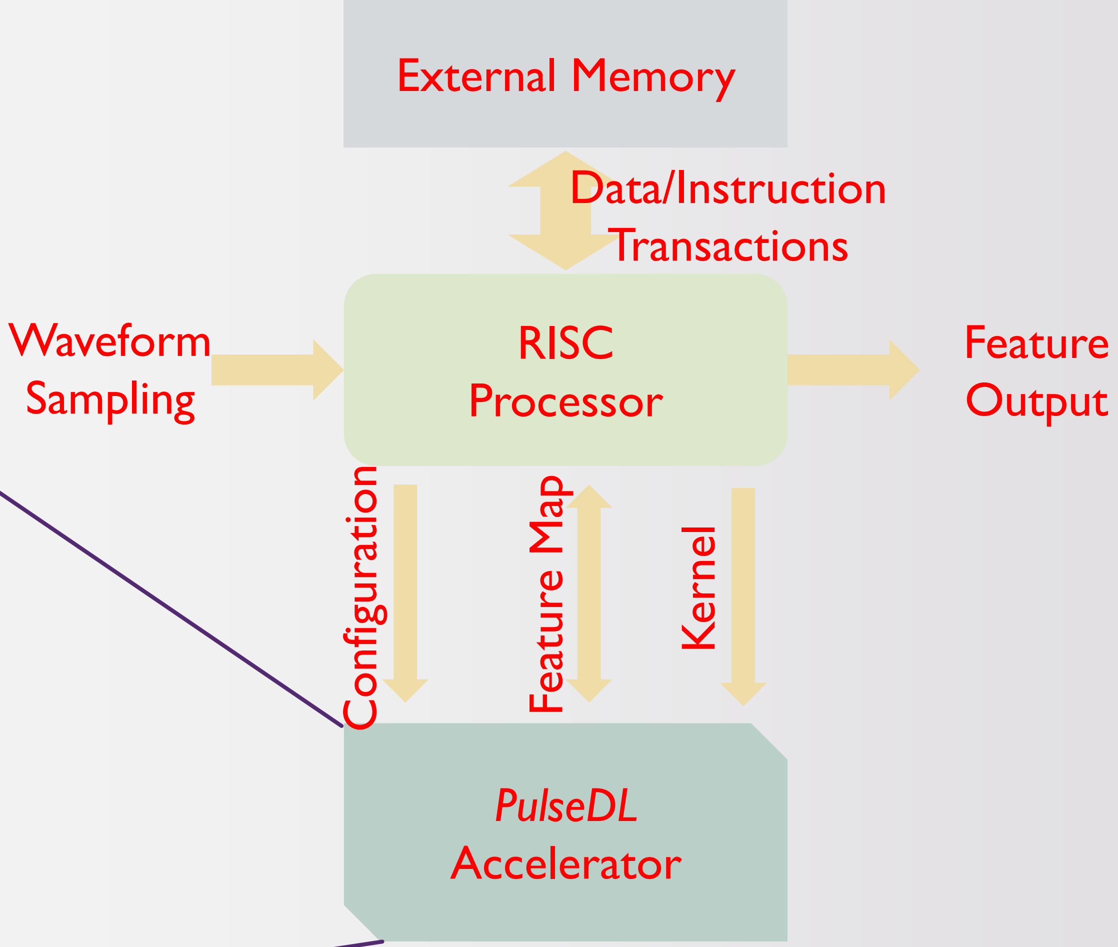


Reduce power and bandwidth and improve performance

A Brief Review of *PulseDL* (BC)


①
②
ⓝ
Feature map vector
f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_L
Fmap input
Temp output
Kernel input
Control signal
Kernel matrix
w_{11} w_{12} w_{13} w_{14}
w_{21} w_{22} w_{23} w_{24}
w_{M1} w_{M2} w_{M3} w_{M4}





External Memory

Data/Instruction Transactions

Waveform Sampling

RISC Processor

Feature Output

Configuration

Feature Map

Kernel

PulseDL Accelerator

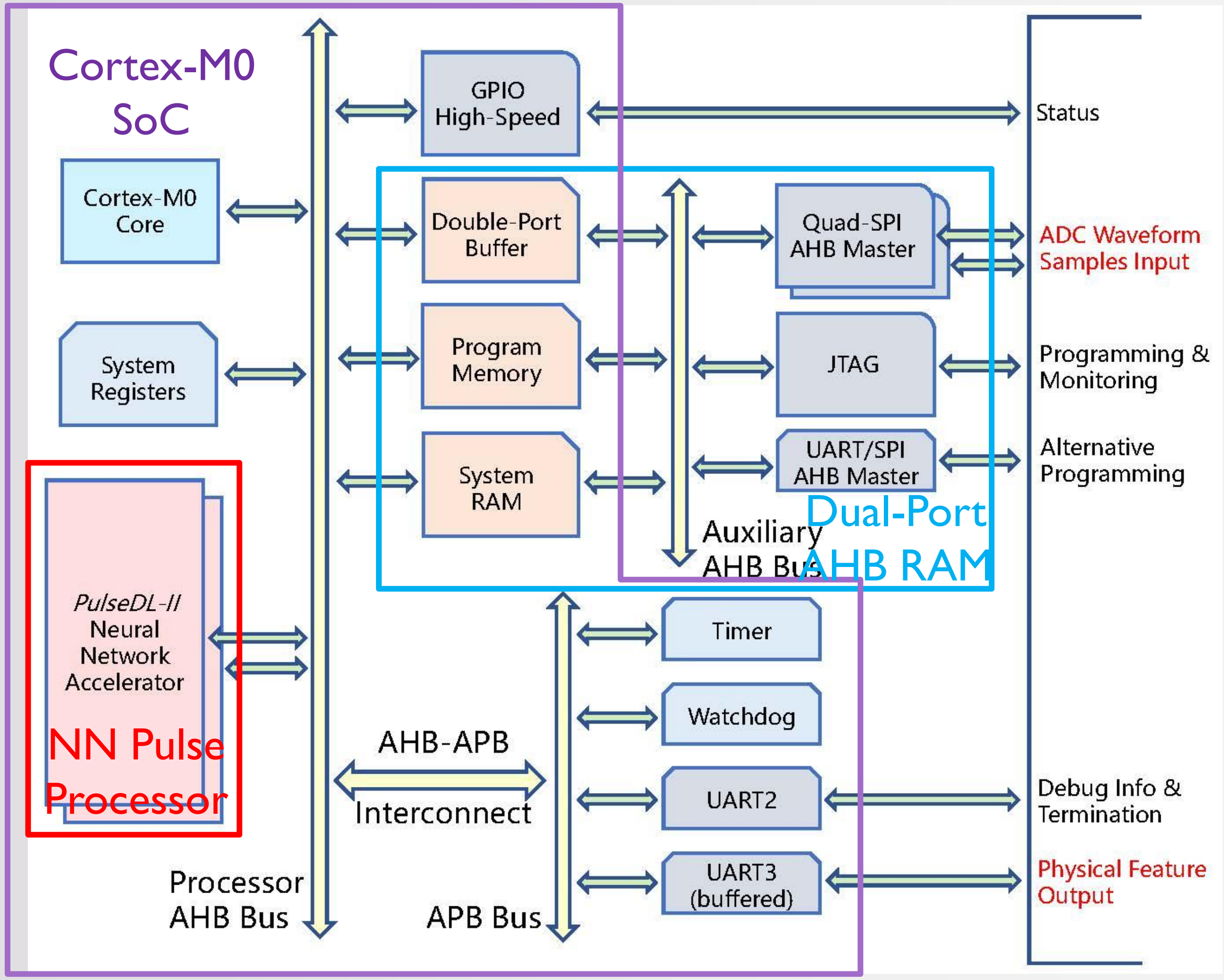
P. Ai *et al* 2020 *NIMA* **978** 164420

Limitations of *PulseDL* (BC)

- The first version of the chip, although a successful practice, has the following limitations:
  - A RISC CPU outside the chip (or NN accelerator) is needed to schedule transactions
  - Dynamic quantization scheme is adopted and may bring about additional time budget
  - The adder tree structure has much space for improvement (especially the temporal adder tree)
  - Only manual configuration was done, and deep learning framework had not been supported yet

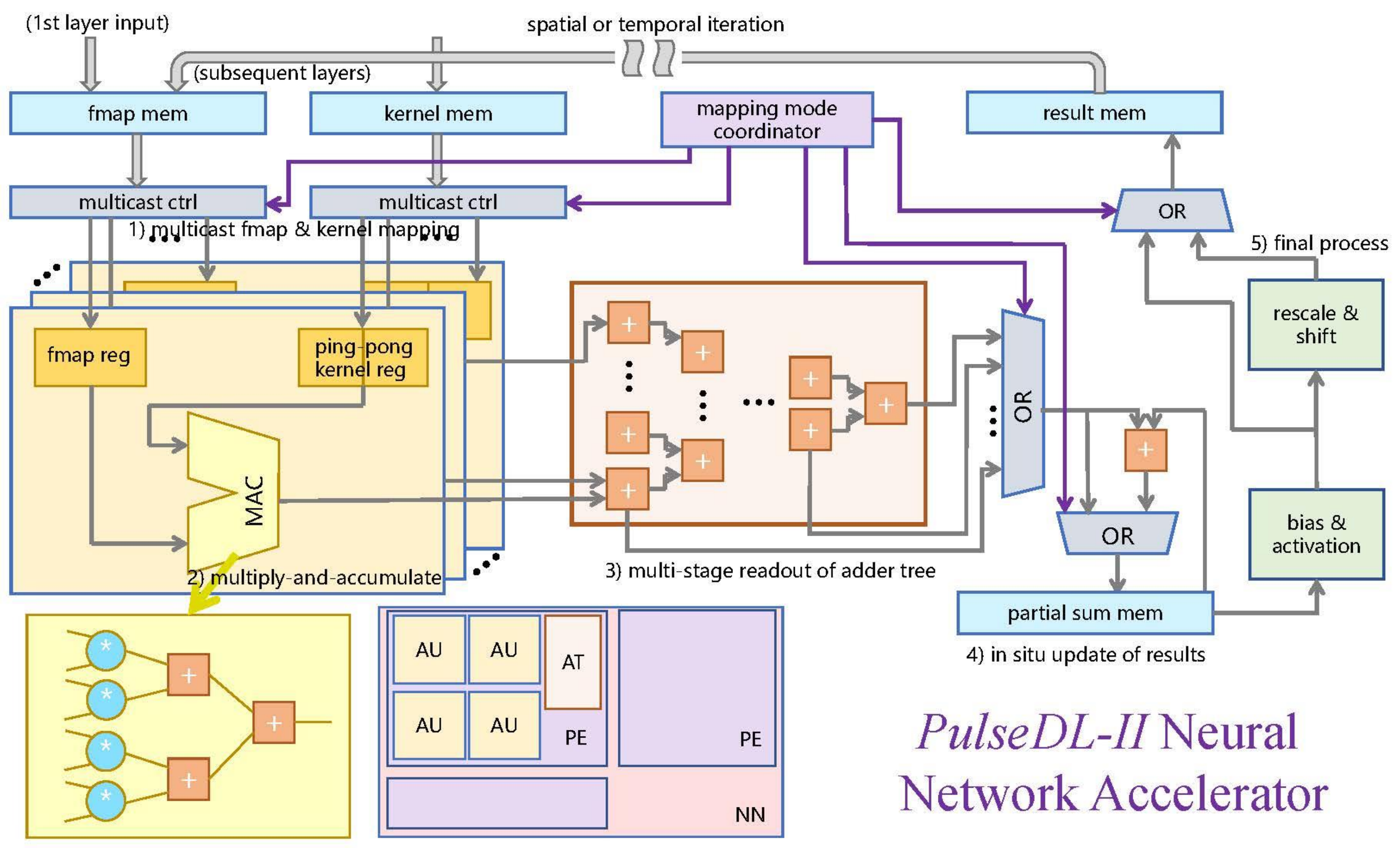
- The above limitations motivate us to develop *PulseDL-II*, the new version of the chip

PulseDL-II: Improvement in System Structure (BC)



- Integrate an RISC CPU into the digital design to form System-on-Chip (SoC)
  - RISC CPU: ARM Cortex-M0
  - System Bus: AHB/APB
- The *PulseDL-II* NN accelerator is mounted on the processor AHB bus as a peripheral
- Input/Output peripherals:
  - Quad/Normal SPI
  - UART (with or without internal buffer)
  - JTAG
  - GPIO

PulseDL-II: Improvement in Accelerator Architecture (BC)



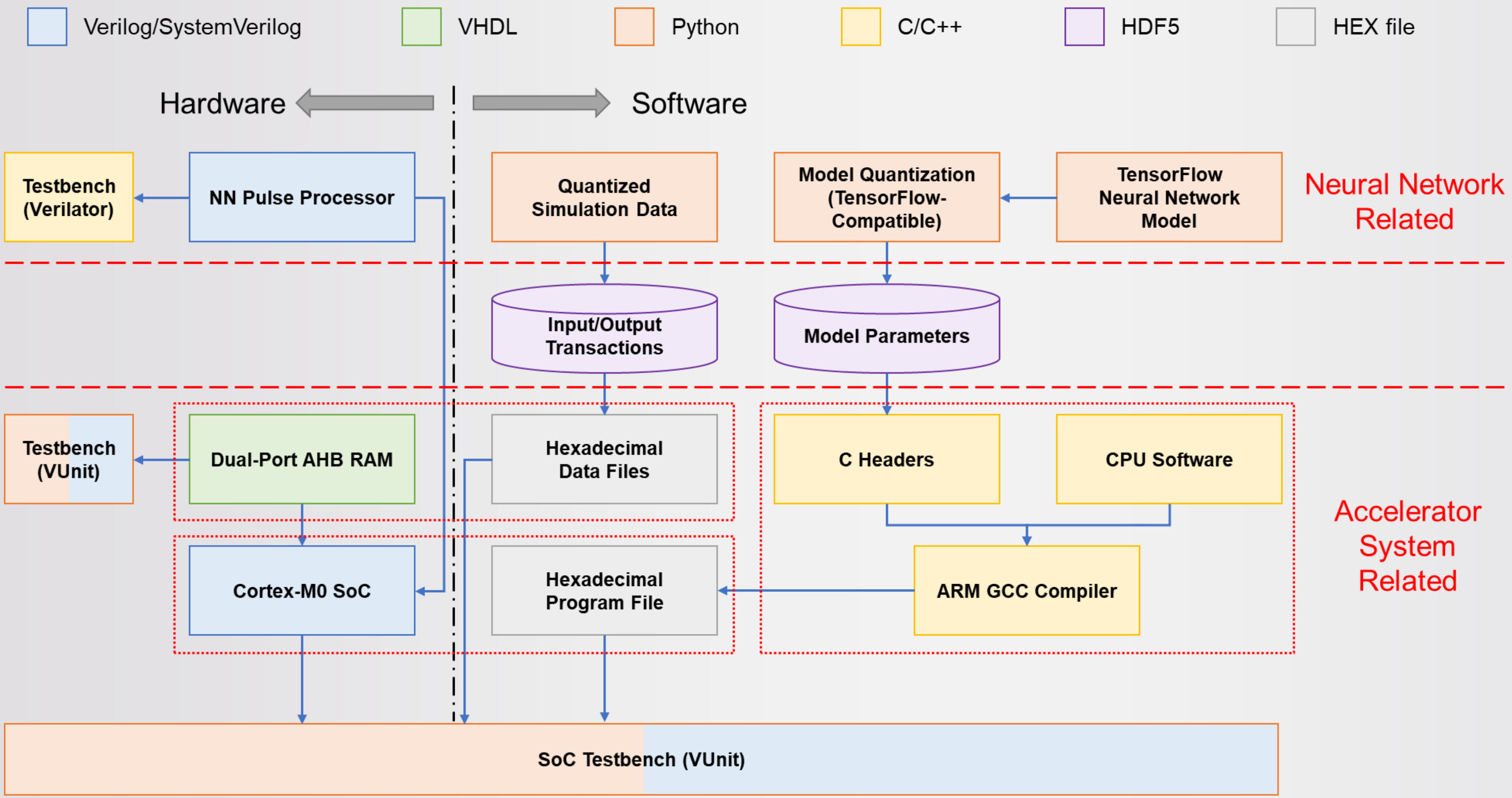
PulseDL-II Neural Network Accelerator

Compared to the last version:

- Adding a new topological level: Arithmetic Unit (AU)
- Broadcasting of input feature map and kernel
- Optimizing the adder tree with partial sum accumulator
- Adding function blocks for bias addition and activation
- For quantization compatible with TensorFlow or other deep learning frameworks, rescale and shift are supported

B.Jacob *et al* 2018 *CVPR* 2704

Hardware-Software Codesign (BC)



HDL to design hardware components

Verilator & VUnit for simulation

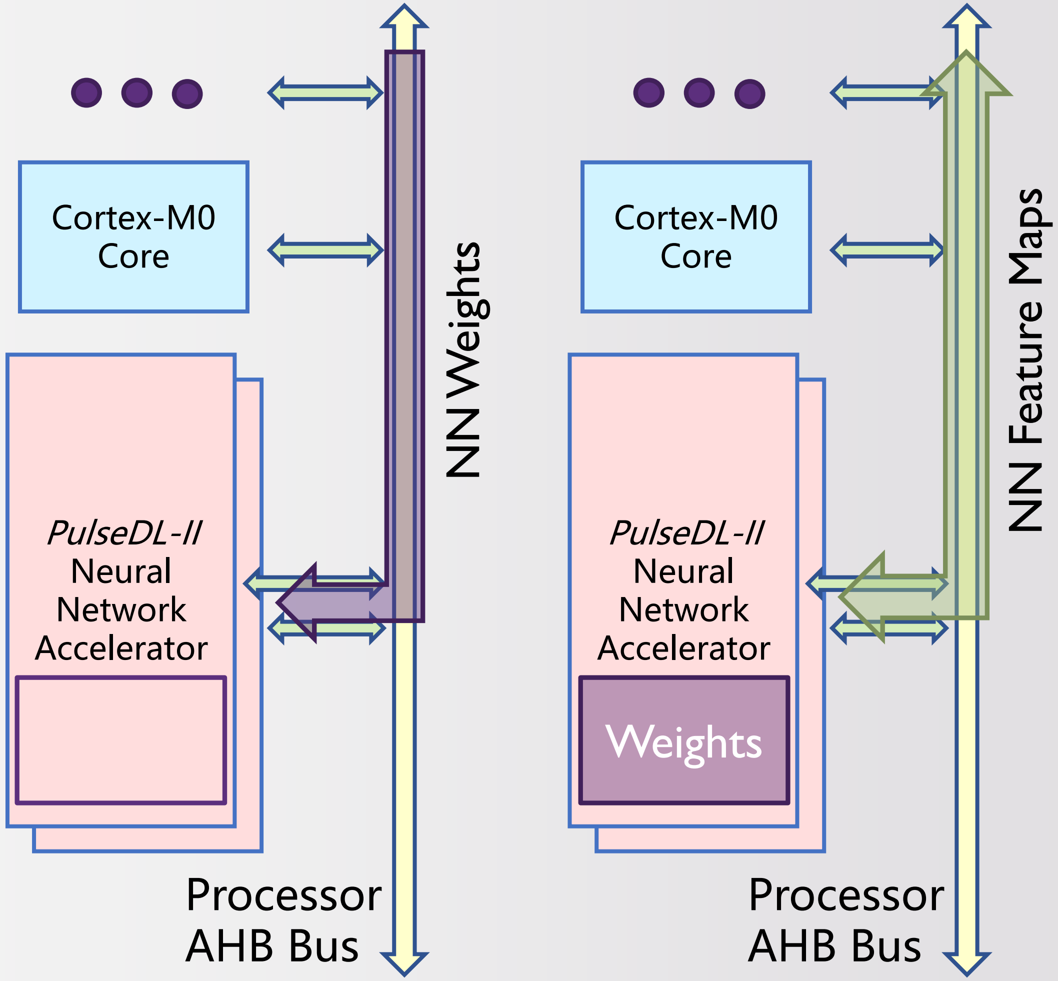
Python to develop NN in TensorFlow framework

HDF5 to save design database

C to design ARM embedded software

Embedded Software with Weight-Stationary Mapping (BC)

- The designed hardware allows different mapping rules
- For NNs with small/medium size, a weight-stationary mapping scheme can be adopted
  - Weights are stored into PEs before samples come in (Preparation Phase)
  - Only input data, output data and intermediate feature maps are transferred during inference (Inference Phase)
- The embedded software enables following features:
  - Layer-wise inference pipelining: weights for different layers are mapped to different groups of PEs, and they can operate simultaneously
  - Event-level parallelism: Each event is assigned a unique token, which will be passed in company with feature maps along the pipeline



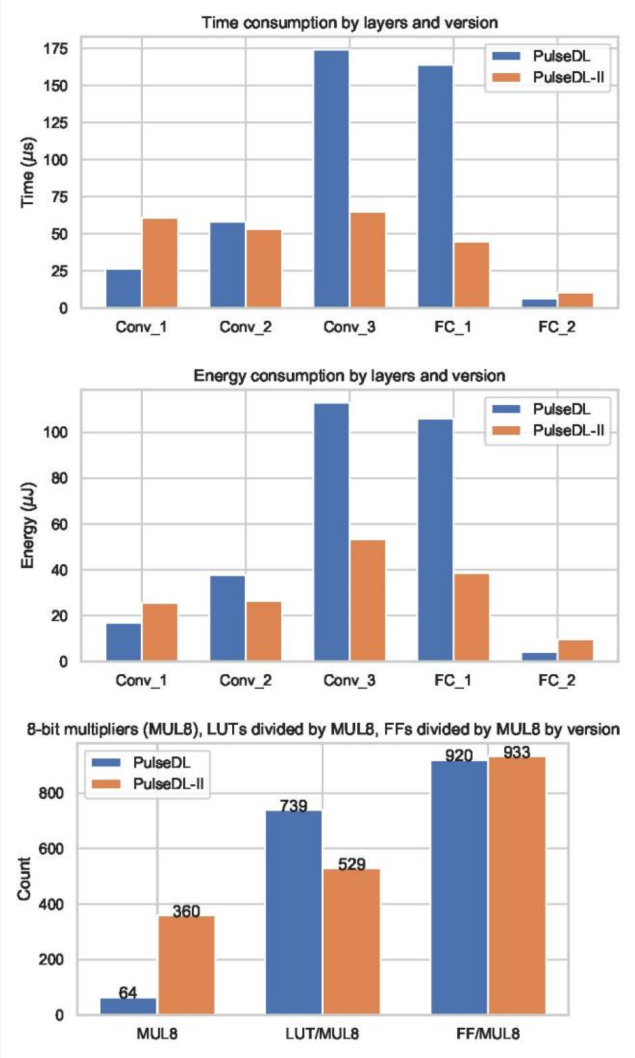
Preparation Phase

Inference Phase

Evaluation of Performance, Power and Area (BC)

- Evaluation settings
  - Xilinx ZCU104 Evaluation Board
  - 100 MHz working frequency
  - post-synthesis

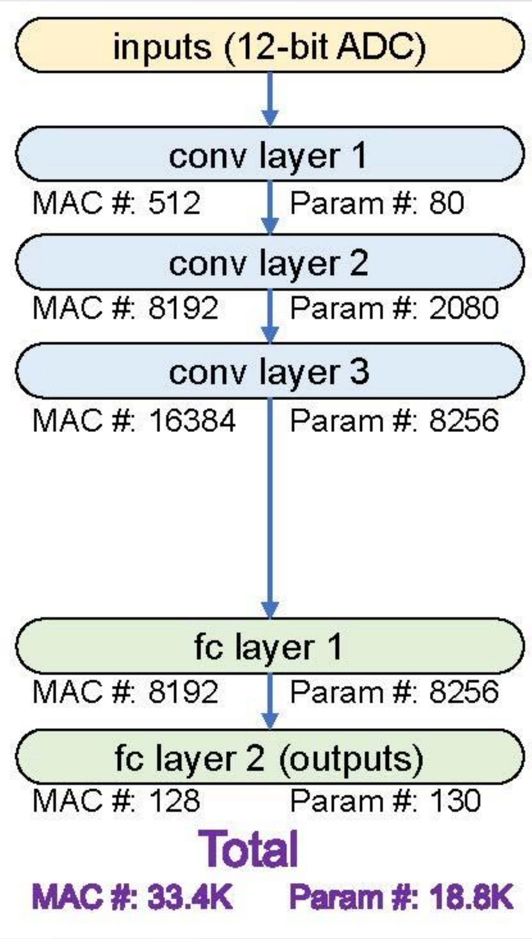
(*PulseDL-II* NN accelerator is isolated for fair comparison with *PulseDL*):



Performance
(time consumption)
1.83x less

Power
(energy consumption)
1.81x less

Area
(resource utilization)
comparable or less



inputs (12-bit ADC)

conv layer 1
MAC #: 512 Param #: 80

conv layer 2
MAC #: 8192 Param #: 2080

conv layer 3
MAC #: 16384 Param #: 8256

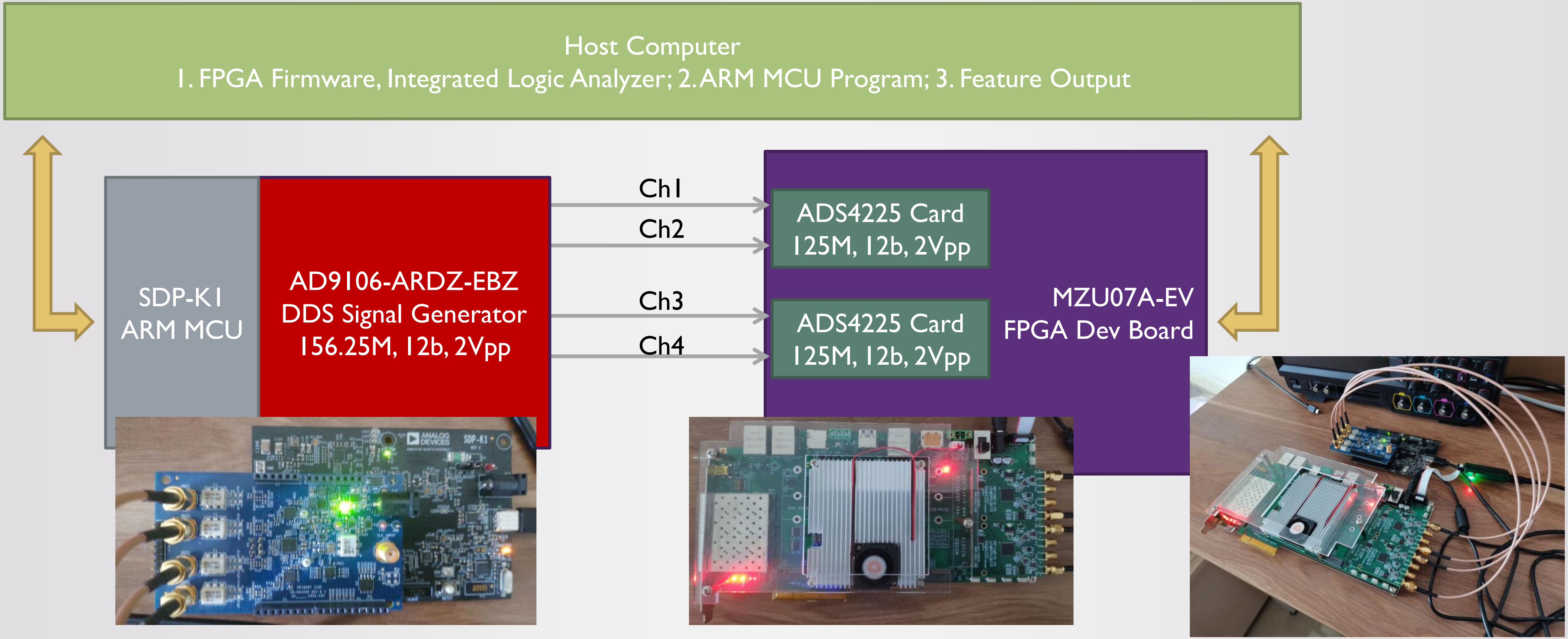
fc layer 1
MAC #: 8192 Param #: 8256

fc layer 2 (outputs)
MAC #: 128 Param #: 130

Total
MAC #: 33.4K Param #: 18.8K

Used NN workload

System Validation: Experimental Setup (ABC)



Host Computer

1. FPGA Firmware, Integrated Logic Analyzer; 2. ARM MCU Program; 3. Feature Output

SDP-K1
ARM MCU

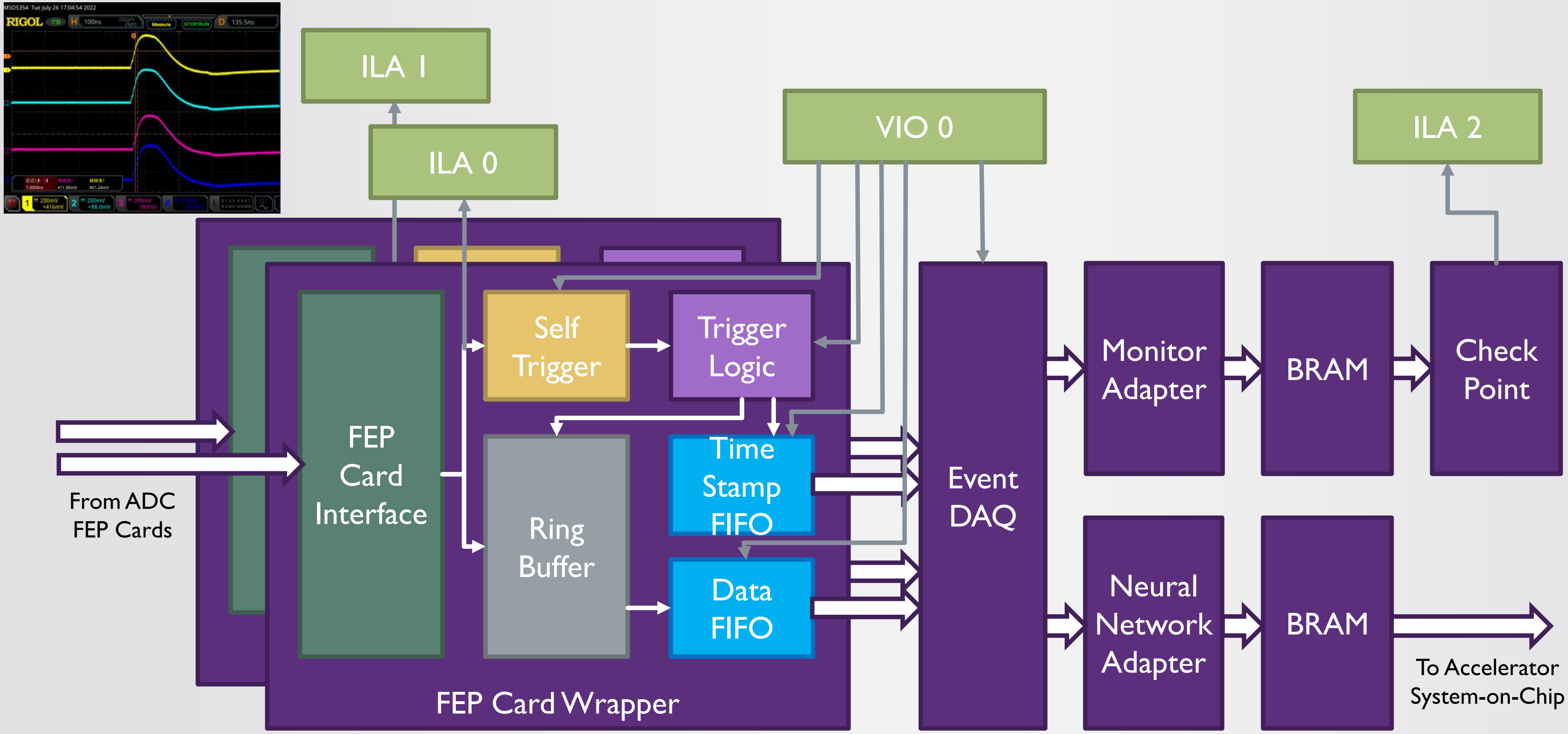
AD9106-ARDZ-EBZ
DDS Signal Generator
156.25M, 12b, 2Vpp

ADS4225 Card
125M, 12b, 2Vpp

ADS4225 Card
125M, 12b, 2Vpp

MZU07A-EV
FPGA Dev Board

System Validation: Digital Logic for Data Acquisition (ABC)

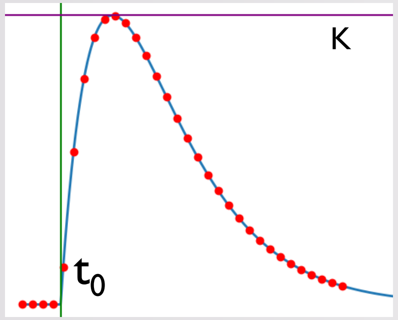


ILA:
Integrated Logic Analyzer

VIO:
Virtual Input/Output

System Validation: Experimental Results (ABC)

- Test waveform:



$$s(t) = K\left(\frac{t-t_0}{\tau}\right)e^{-\frac{t-t_0}{\tau}}u(t-t_0)$$

$$\tau = 40\,ns, \qquad K = K_1K_2$$

$$SNR = 20log_{10}\left(\frac{K_1}{\sigma_{base}}\right) = 47.4\,dB$$

$$K_2 \sim U(0.5, 2.0)$$

- Sample 32 points per event
- Dual-channel synchronous waveform input

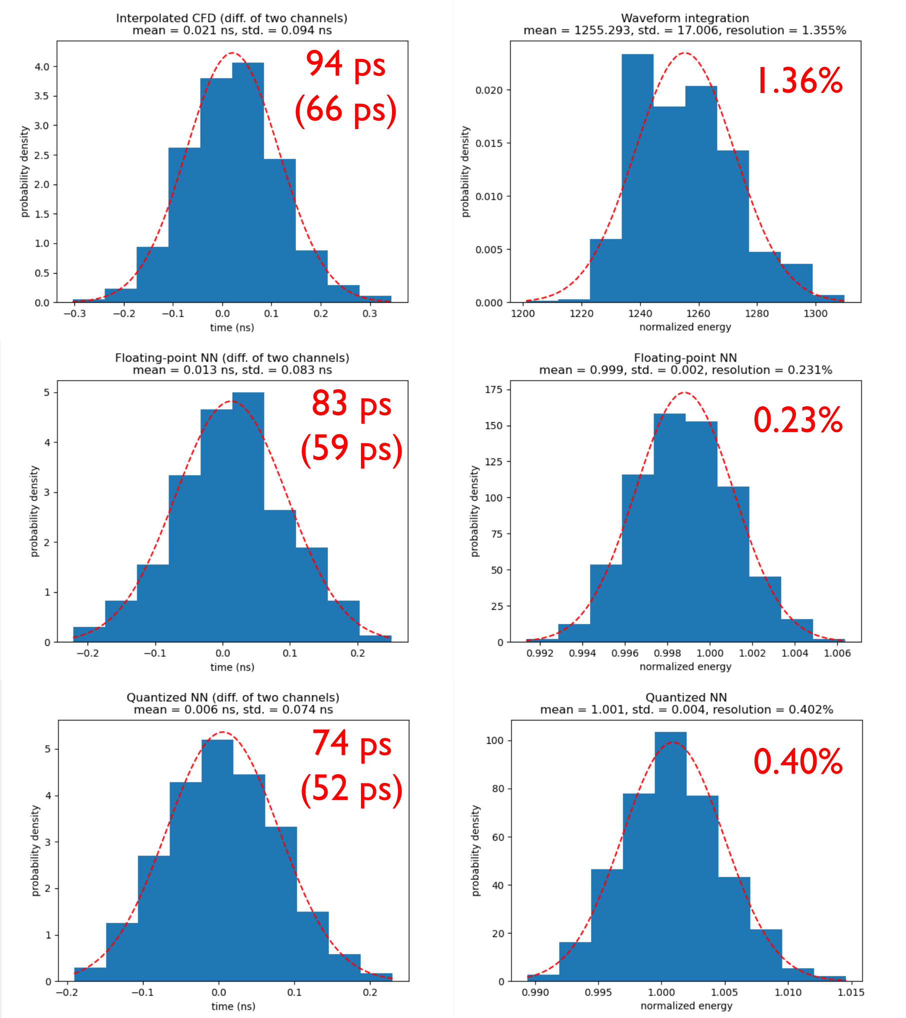
time

energy

traditional methods (CFD, integral)

floating-point NNs

quantized NNs



94 ps (66 ps)

1.36%

83 ps (59 ps)

0.23%

74 ps (52 ps)

0.40%

- Runtime statistics:
  - Zynq UltraScale+

Resources (area):	
LUT	2825 + 89540
FF	517 + 75028
BRAM	8.0 + 48.0
URAM	8 + 0
Power:	
Dynamic	(0.371 + 0.541) W
Static	0.594 W
Performance @ 100 MHz:	
Internal inf.	113.8 us
Throughput	8.3k events/sec

Summary

- The ability and potential of NNs in signal feature extraction are investigated
- Application-specific NN architectures are designed
- NN accelerator-based front-end electronics is prototyped
- System-on-Chip digital system with NN accelerator is developed
- System Validation on FPGA platform is done

What's next:

- Evaluate the whole system in real-world nuclear detector dataflows
- Design optimization, ASIC layout, tape-out with advanced technology

THANK YOU!

ANY QUESTIONS?