A Fully Synthesized 13.7μJ/prediction 88% Accuracy CIFAR-10 Single-Chip Data-Reusing Wired-Logic Processor Using Non-Linear Neural Network

The University of Tokyo
Kuroda Laboratory

Yao-chung Hsu, Atsutake Kosuge,
Rei Sumikawa, Kota Shiba,
Mototsugu Hamada, Tadahiro Kuroda

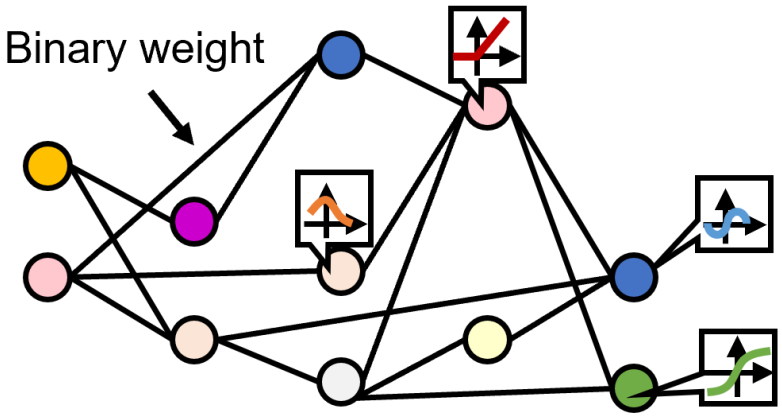
AI Processors on Edge Devices

- Energy consumption is the technical challenge for applying DNN to edge devices
- Memory access and multiplication are bottlenecks
- Wired-logic architecture doesn't need memory access but requires large area

Operation	Energy[pJ]
32-bit INT Add	0.1
32-bit INT Mul.	3.1
32-bit 32KB SRAM	10
32-bit DRAM	650-1300

Non-Linear Neural Network (NNN)

- Weights are binarized
- Unnecessary synapses are pruned
- Activation functions are optimized through training
- Accuracy can be remained

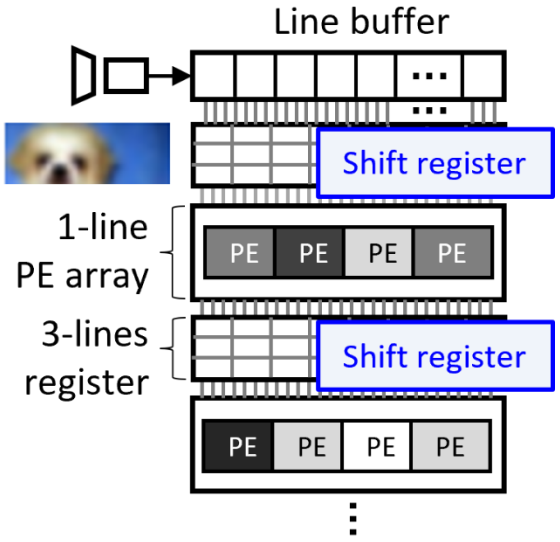


Non-linear Neural Network

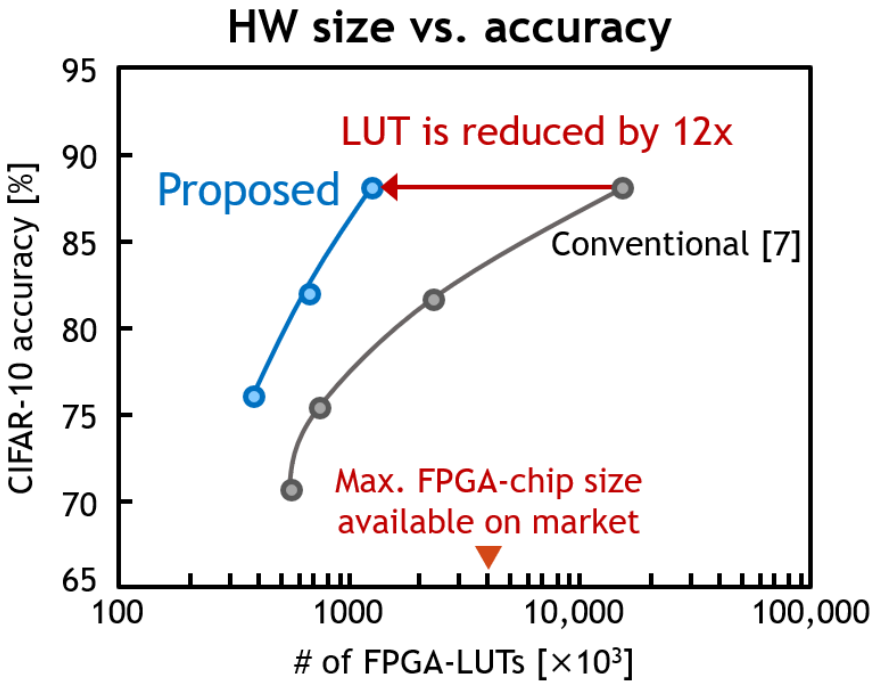
	Conv. CNN	Pruned BNN [7]	Proposed NNN
Data set	CIFAR-10		
# of CNN layers	8 convolution , 2 dense, 4 pooling layers		
Weight bit width	INT8b	Binary	
Pruning rate	0%	97.8%	97.8%
Activation	Relu	Relu	Various functions
Accuracy	84 %	67%	88%
# of FPGA-LUTs	7.0×10^9 (1)	1.5×10^7	1.5×10^7 (1/468)

Pipelined Wired-Logic Architecture

- No need for memory access and multipliers
- Data is reused and hardware utilization is reduced by using pipeline
- Verilog code can be generated by Python agilely

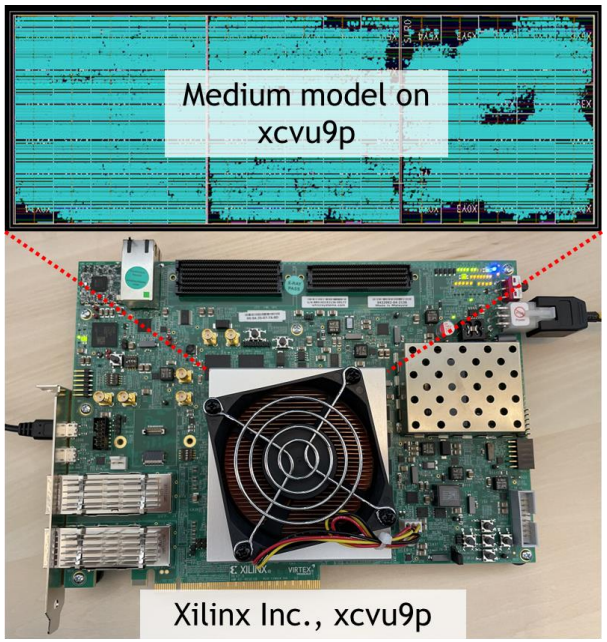


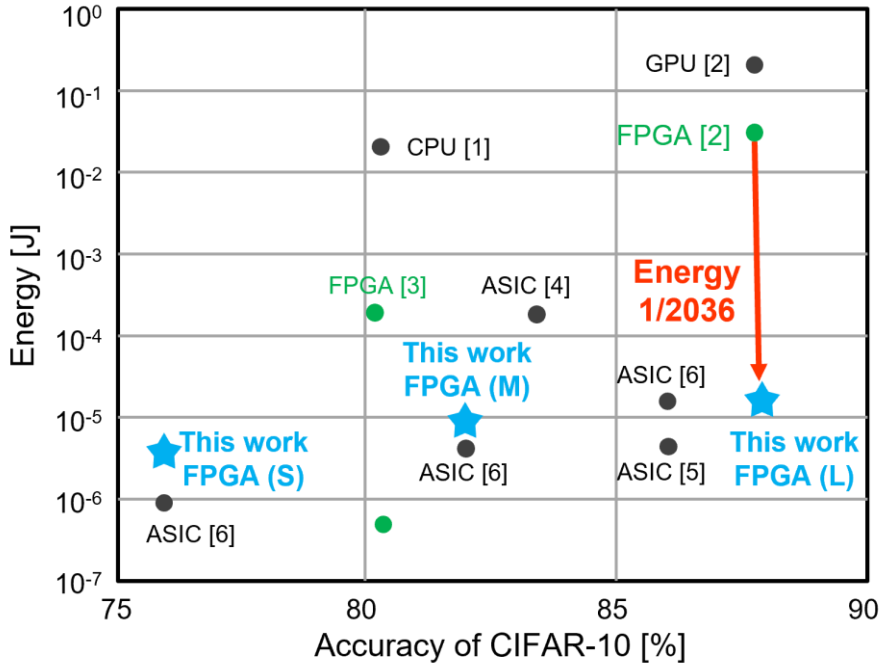
Proposed pipelined wired-logic processor



Implementation Results

- 3 models are implemented in terms of their accuracy
- Large model can reduce energy consumption by 2036 times compared to previous FPGA works
- The energy consumption of this work is closed to that of ASIC implementation even with FPGA





If you are interested in the contents, please come to my poster presentation.