A 13.7μJ/prediction 88% Accuracy CIFAR-10 Single-Chip Wired-logic Processor in 16-nm FPGA using Non-Linear Neural Network

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

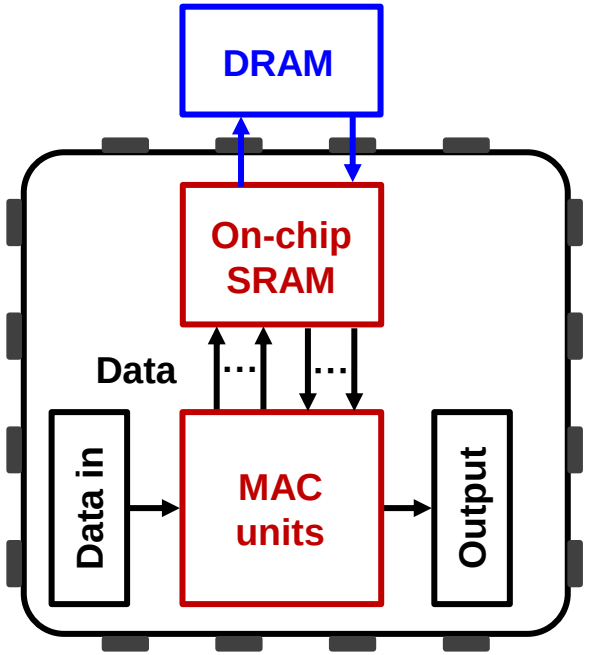
The University of Tokyo

Abstract

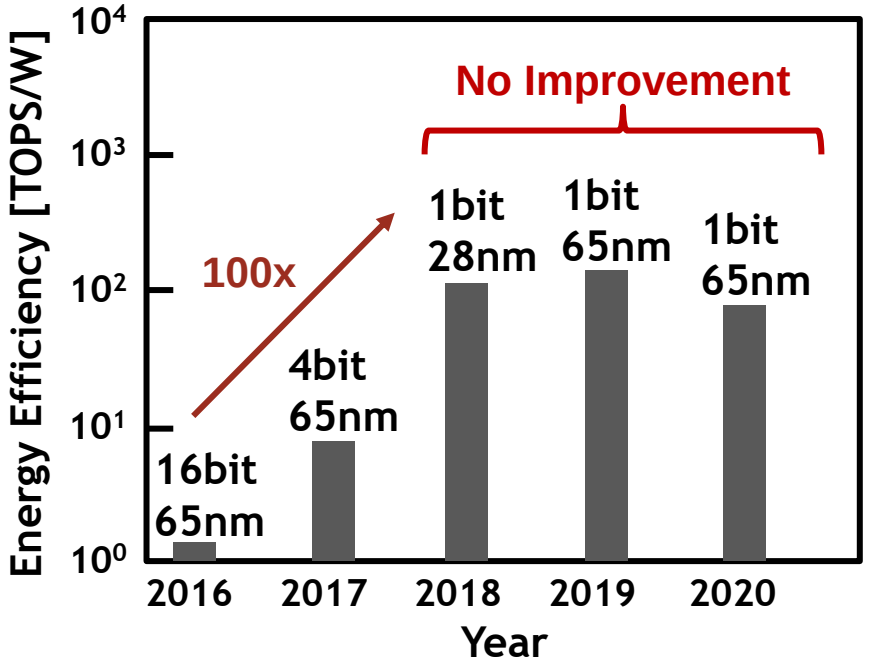
- In this study, we propose a 13.7μJ/prediction 88% accuracy CIFAR-10 single-chip wired-logic processor in 16-nm FPGA by utilizing a newly developed 98%-pruned ultra-sparse, binary-weight nonlinear neural network (NNN) and a shift-register based pipelined wired-logic architecture. Compared with the state-of-the-art FPGA-based processor, 2,036 times better energy efficiency is achieved.

Introduction

- Pace of Energy Efficiency Improvement Slowing
  - Processor Element (PE) Bit Width Already Reduced to 1b
  - Processors Using Only On-chip SRAM Already Realized
    ⇒ Power-Hungry SRAM Access also should be eliminated



Conventional von-Neumann AI Processor



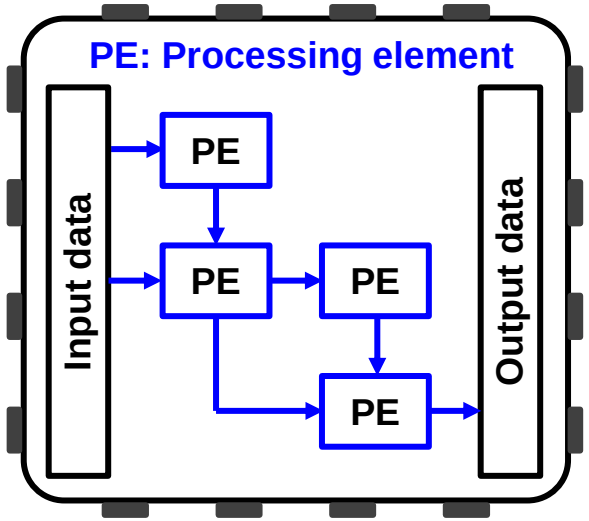
Energy Efficiency Trend in ISSCC

Operation	Energy [pJ]
32-bit int ADD	0.1
32-bit int MULT	3.1
32-bit 32KB SRAM	10
32-bit DRAM	650-1300

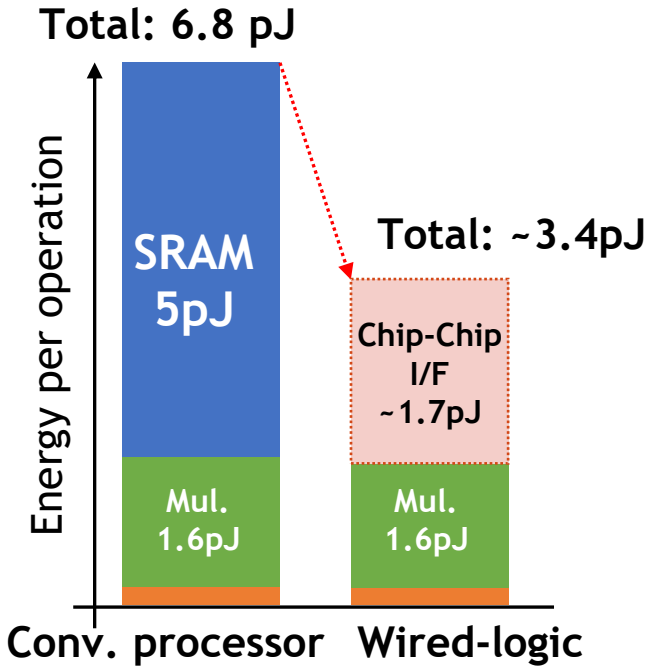
Energy table for 45nm CMOS

Wired-logic Architecture

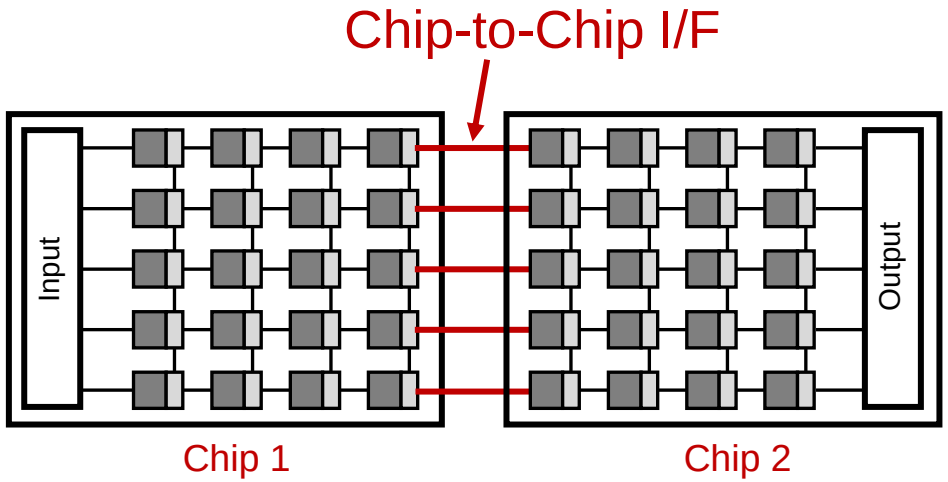
- Goal: Energy-efficient AI Processor by eliminating the memory access.
  - Ex. Implementing 88% Acc. CIFAR-10 SNN requires 3,080mm^2 in 28nm, resulting in 8 TrueNorth chips [4].
  - It requires power-hungry chip-to-chip I/F, resulting in poor energy efficiency.



Wired-logic AI Processor



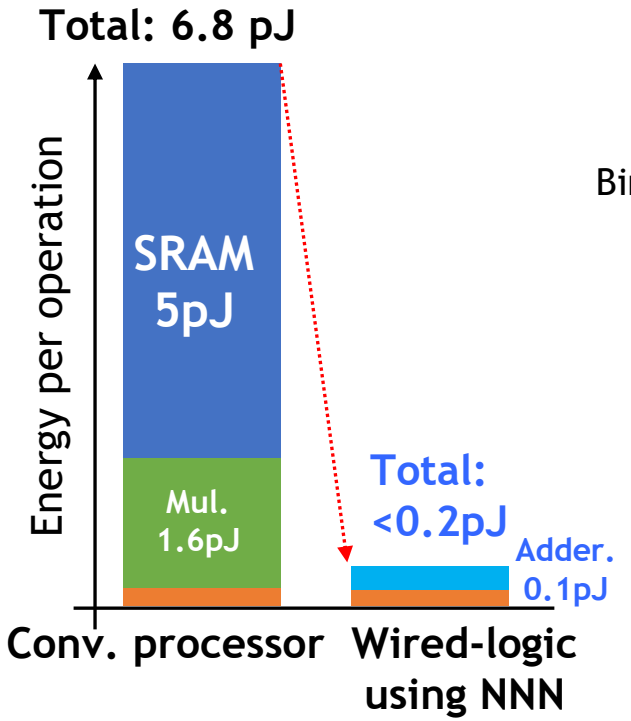
Energy Consumption Comparison



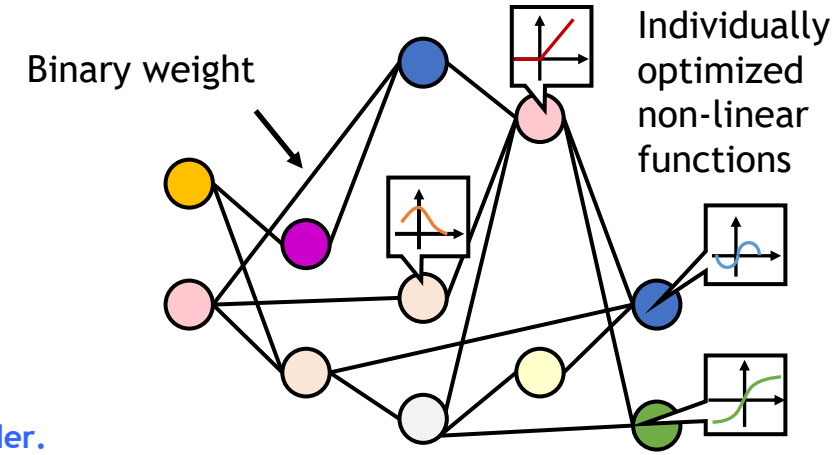
Conventional wired-logic architecture

Our Research Goal

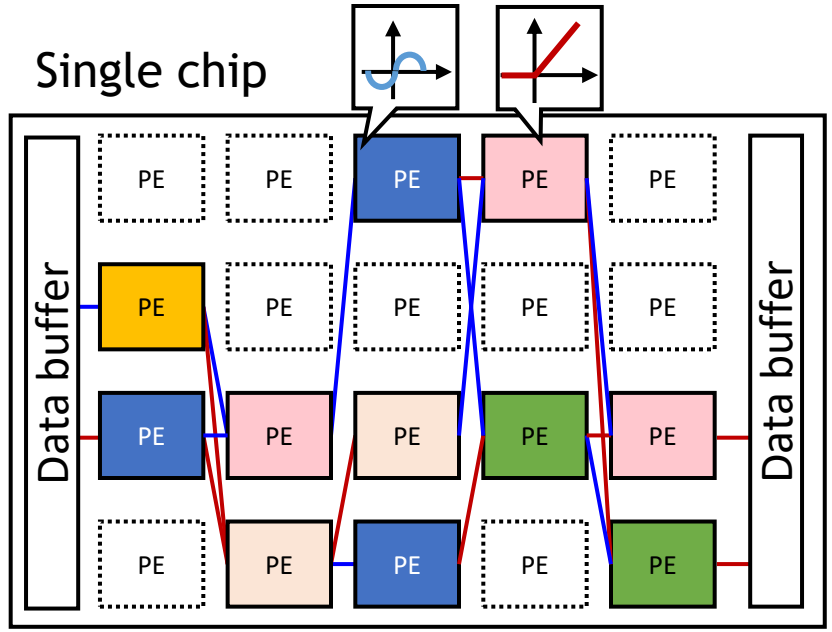
- Goal: Single Chip AI Processor with 2,036x Higher Energy Efficiency Using
  - (A) ~98% pruned ultra-sparse, binary-weight nonlinear neural network (NNN).
  - (B) Shift-reg. based wired-logic architecture saves chip area by a factor of 14.
  - (C) Agile (5 min.) synthesis from Python.



Energy Consumption Comparison



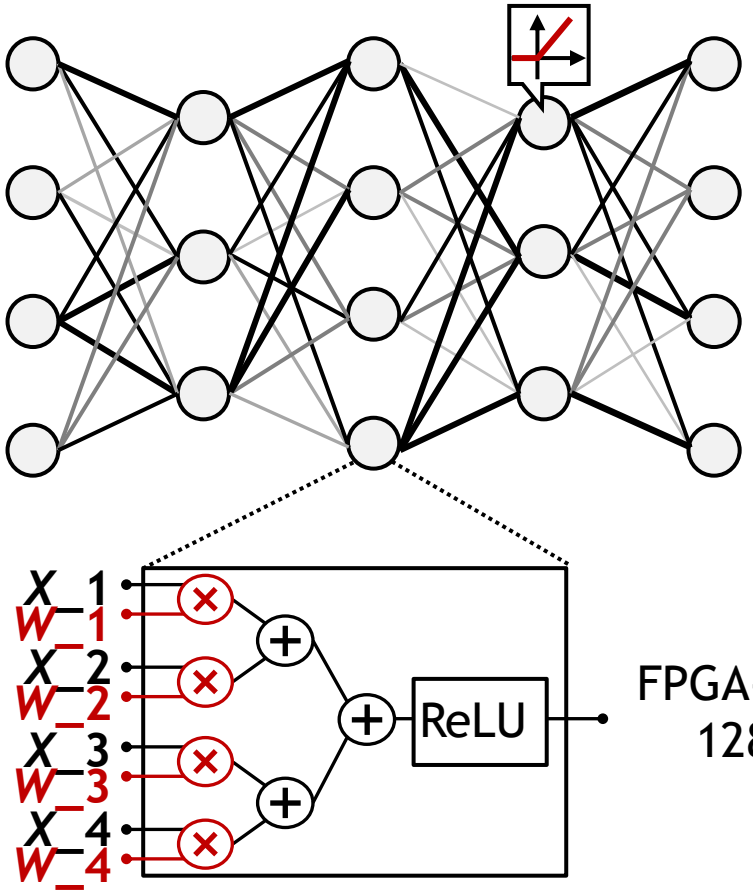
Proposed ultra-sparse network: NNN



Proposed wired-logic architecture

Nonlinear Neural Network (NNN)

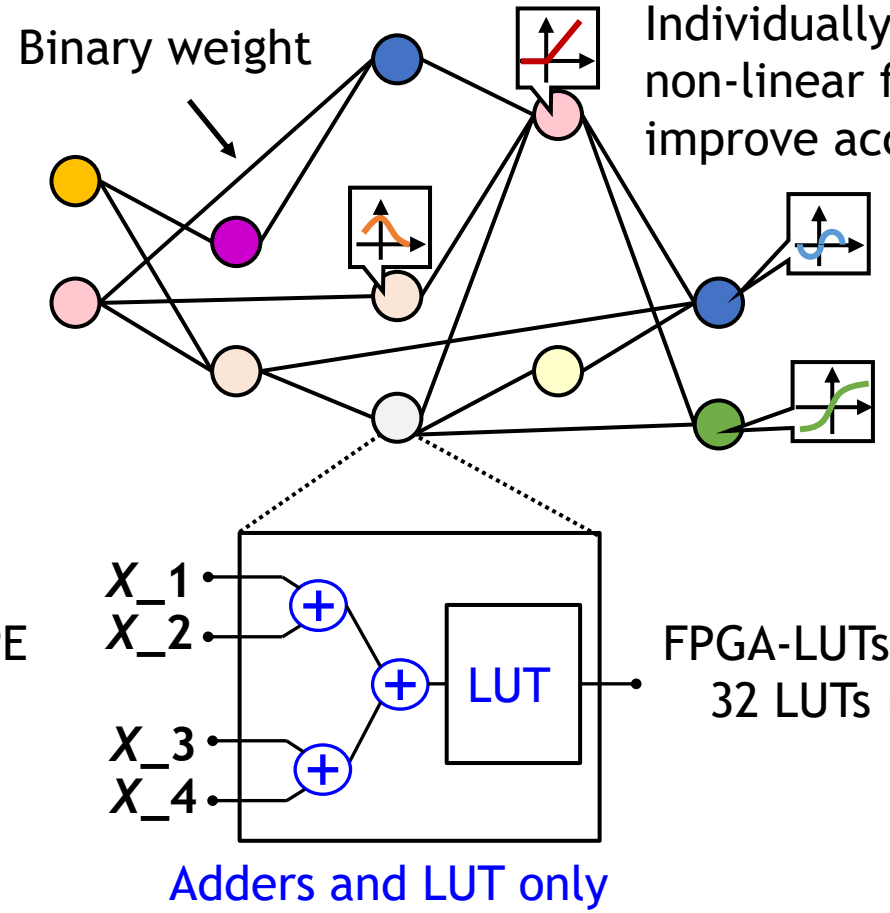
Deep Neural Network (DNN)



Multipliers and adders

FPGA-LUTs per PE
128 LUTs (1)

Nonlinear Neural Network (NNN)

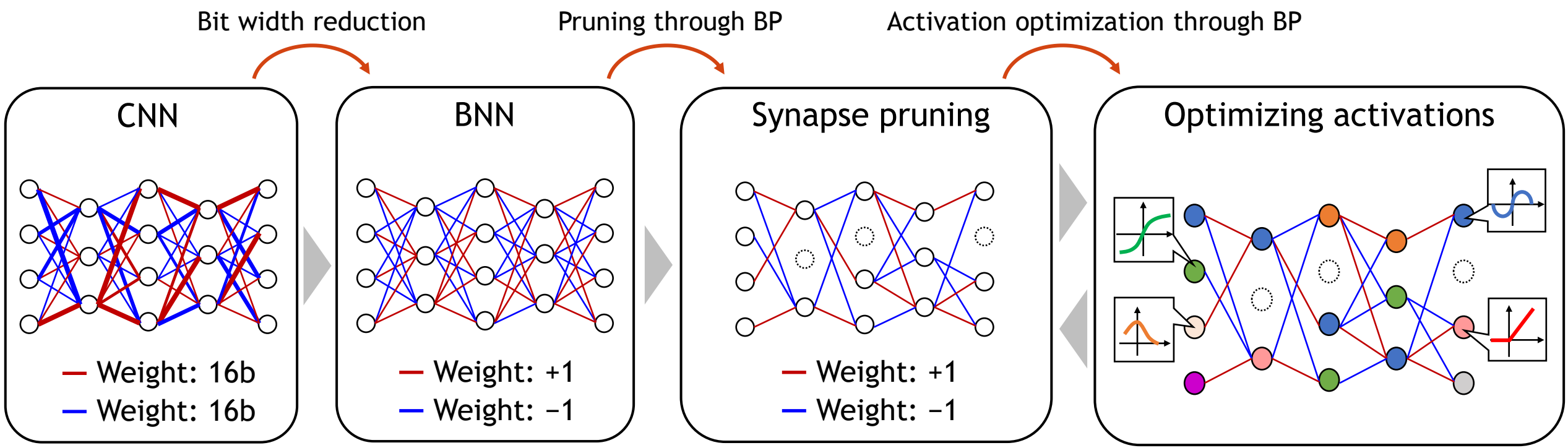


Adders and LUT only

FPGA-LUTs per PE
32 LUTs (1/4)

Back Propagation Based Training

- An CNN is given as an initial structure to limit search space
- Both synapse pruning and activation function optimization are updated by back propagation



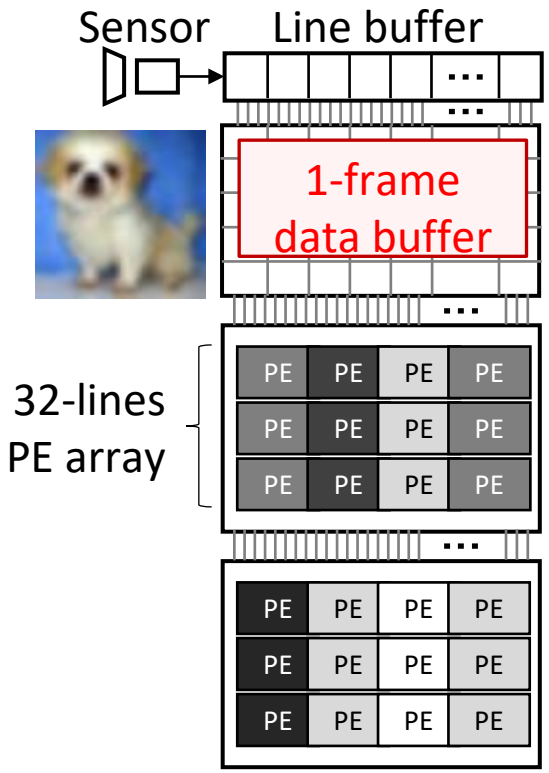
Comparison with Other Neural Networks

- ~98% pruning is achieved while maintaining high accuracy
- Hardware resource utilization is reduced by a factor of 468

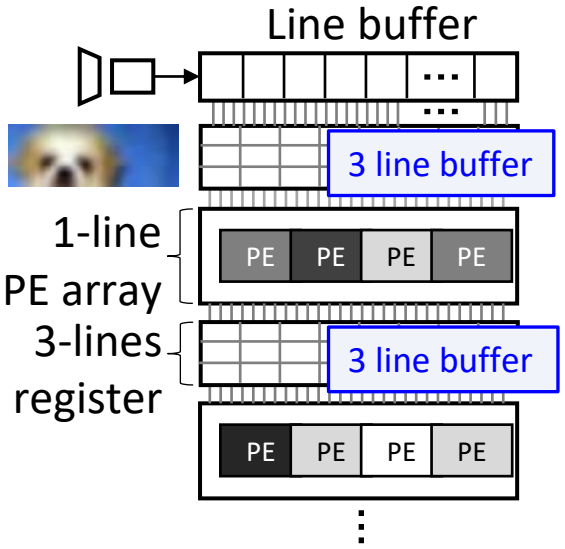
	Conv. CNN	Binarized CNN	Pruned BNN	Proposed NNN
Data set	CIFAR-10			
# of CNN layers	8 convolution , 2 dense, 4 pooling layers			
Weight bit width	INT8b	Binary		
Pruning rate	0%	50%	97.8%	97.8%
Activation function	ReLU	ReLU	ReLU	Various functions
Accuracy	84 %	85%	67%	88%
# of FPGA-LUTs	7.0×10^9 (1)	6.3×10^8	1.5×10^7	1.5×10^7 (1/468)

Pipelined Wired-Logic

- Conventional wired-logic processor processes all data at the same time, resulting in large circuit size
- In pipelined wired-logic processor, only a portion of the data are processed at the same time

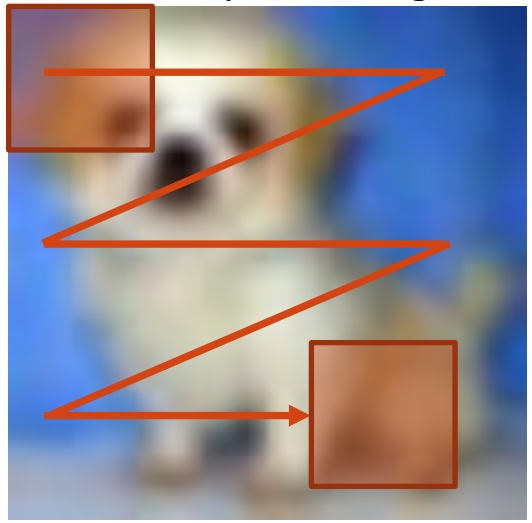


Conv. wired-logic processor



Proposed 3line buffer-based wired-logic processor

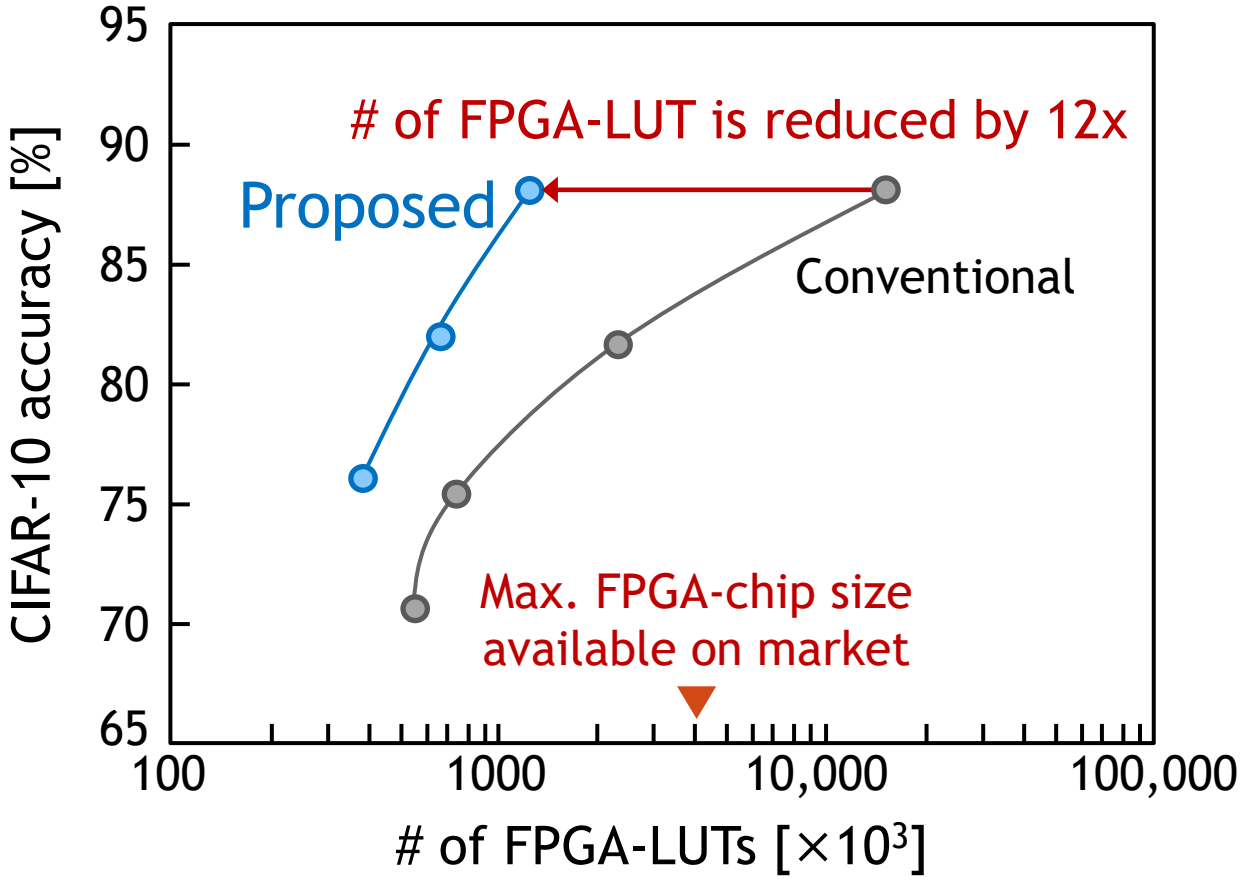
Filter operation in NNN processing



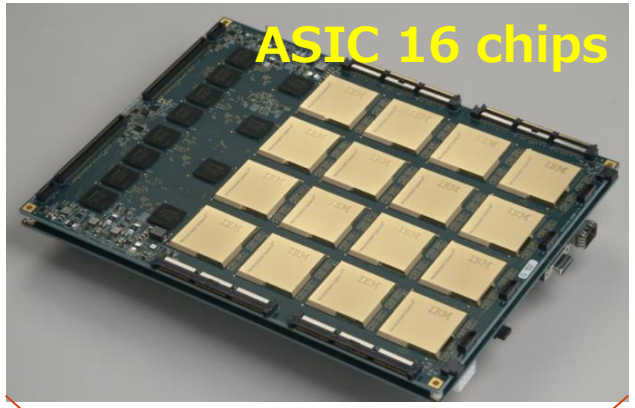
NNN filter

Hardware Resource Reduction

- Pipelined wired-logic reduces hardware usage by 12x
- 14-layer NNN can be implemented with single FPGA



IBM TrueNorth 28nm

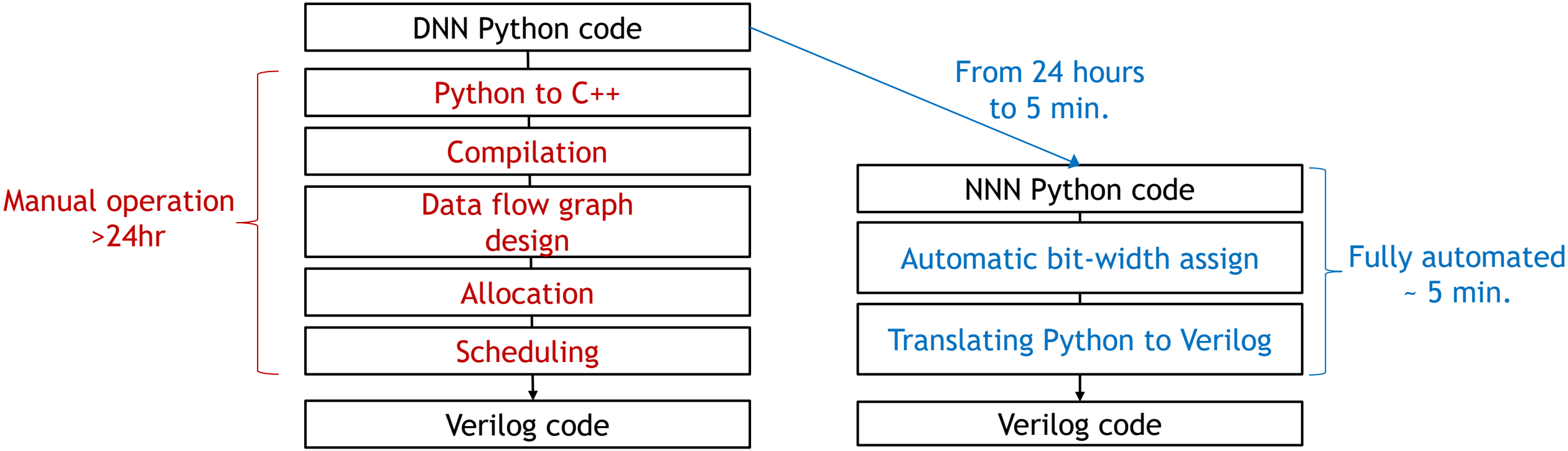


CNN 15 layer
Acc. 86%

Single 16nm-FPGA
NNN 14 layer, Acc. 88%

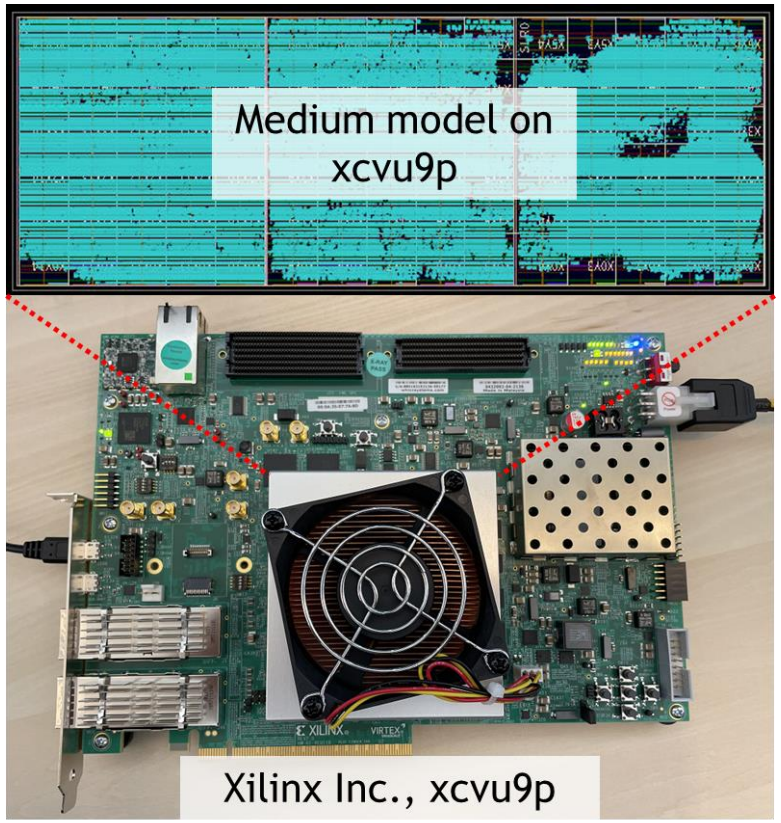
Agile Implementation

- Verilog code for NNN pipelined wired-logic processor can be generated by Python code agiely



Implementation Results

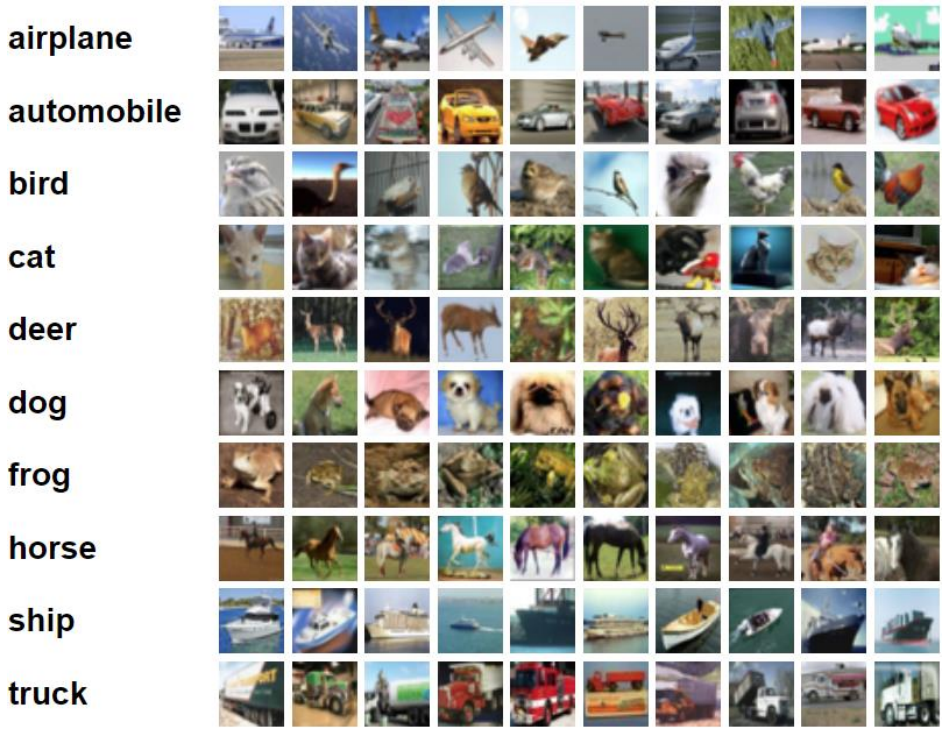
- 3 models, with different size, are implemented
- None of them uses memory



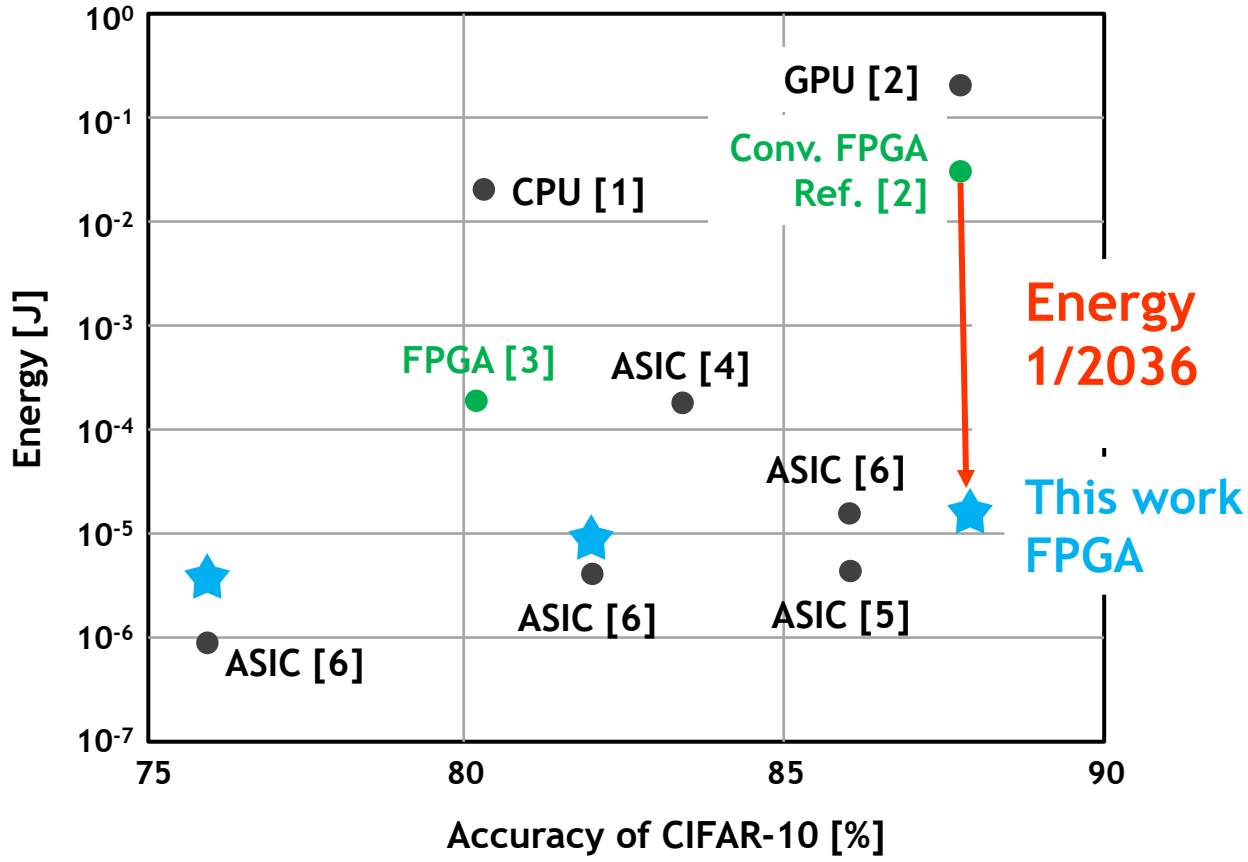
Chip Size	Small	Medium	Large
Pruning rate [%]	99.55	99.11	97.78
Accuracy [%]	76	82	88
LUTs	396,341	667,353	1,250,075
FFs	69,697	96,838	106,723
BRAMs	0	0	0
DSPs	0	0	0
Throughput [Mfps]	0.94	0.94	1.03
Clock [MHz]	30	30	33
Static power [W]	2.5	2.7	8.1
Dynamic power [W]	1.7	5.2	6.1
Energy efficiency [μJ/frame]	**4.0**	**7.6**	**13.7**

Comparison with Previous Works

- Energy efficiency is improved by 2036x compared with SOTA 28nm FPGA work[2]



10 classes and examples of CIFAR-10



[1] L. Lai, *et al.*, *arXiv 1801.06601*, pp. 1-10, Jan. 2018. [2] R. Zhao, *et al.*, *Int' Symp. on FPGA*, pp. 15-24, Feb. 2017.
[3] Y. Umuroglu, *et al.*, *Int' Symp. on FPGA*, pp. 65-74, Feb. 2017.
[4] Steven K. Esser et al., "Convolutional networks for fast, energy-efficient neuromorphic computing,"
in PNAS, vol. 113, no. 41, Oct. 2016.
[5] D. Bankman, *et al.*, *ISSCC*, pp. 222-224, Feb. 2018
[6] B. Moons, *et al.*, *CICC*, pp. 1-4, Apr. 2018.
[7]: A. Kosuge et al., IEEE OJCAS, vol. 3, pp. 4-14, Jan. 2022.

Conclusion

- A pipelined wired-logic AI processor using NNN is proposed
  - NNN can achieve high accuracy even with binarized weight and aggressive pruning
  - By using NNN, hardware resource usage can be reduced by 468x
  - By using pipelined wired-logic, hardware resource usage can be reduced by 12x
  - Verilog code can be generated by Python code agiley
- This architecture is implemented with Xilinx xcvu9p
  - 2036x energy efficiency improvement compared with SOTA FPGA implementation