July 5, 2017
MPSoC2017@Annecy, France

Energy-Efficient In-Memory Neural Network Processor

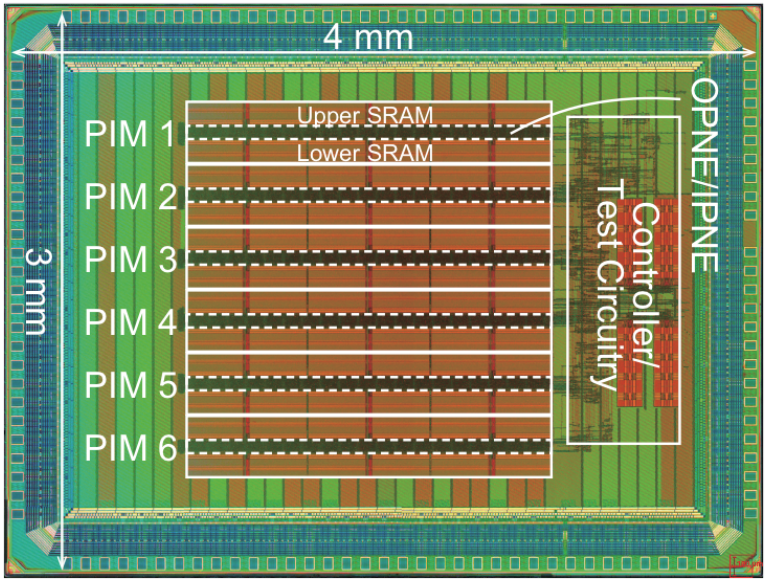
Shinya Takamaeda-Yamazaki

Hokkaido University, Japan

E-mail: takamaeda_at_ist_hokudai_ac_jp

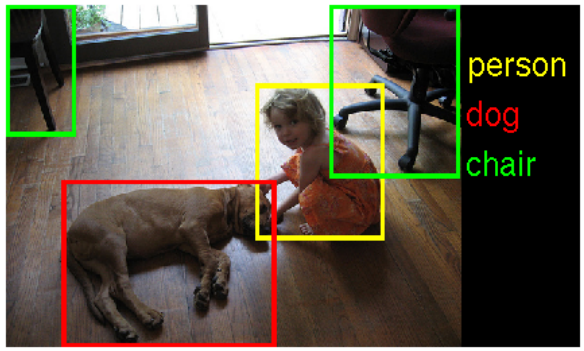
Agenda

- BRein Memory: a binary/ternary DNN accelerator
  - Employing Processing in Memory (PIM) architecture
- 1st test chip with 6 PIMs
  - Peak 1.38 TOPS @ 400 MHz, Efficiency 2.3 TOPS/W
  - First accelerator chip for binary & ternary DNNs



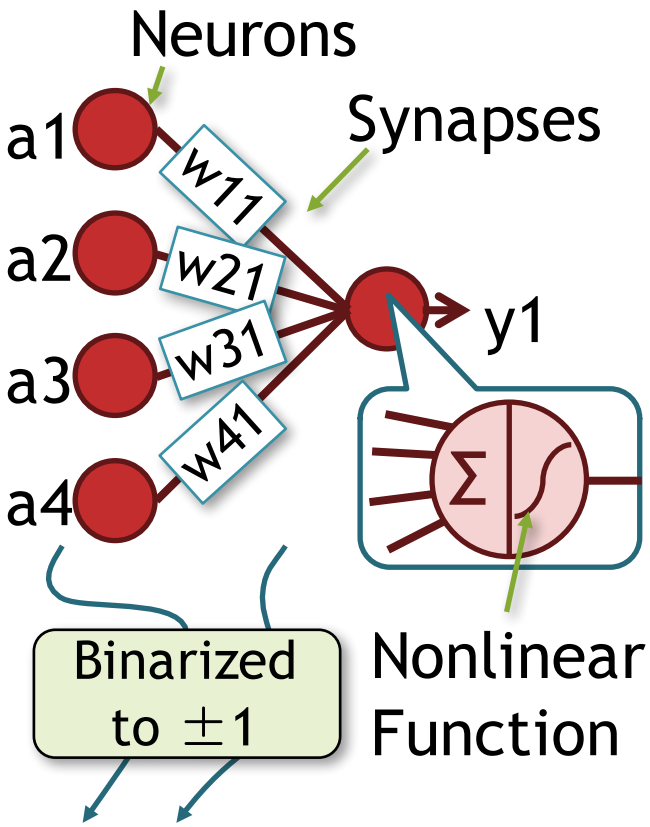
Deep Neural Network (DNN)

- Major AI technology in various applications
  - Ex) Recognition for autopilot car, robotics, …
- Performance and energy-efficiency demands
  - Low-power processing for embedded IoTs

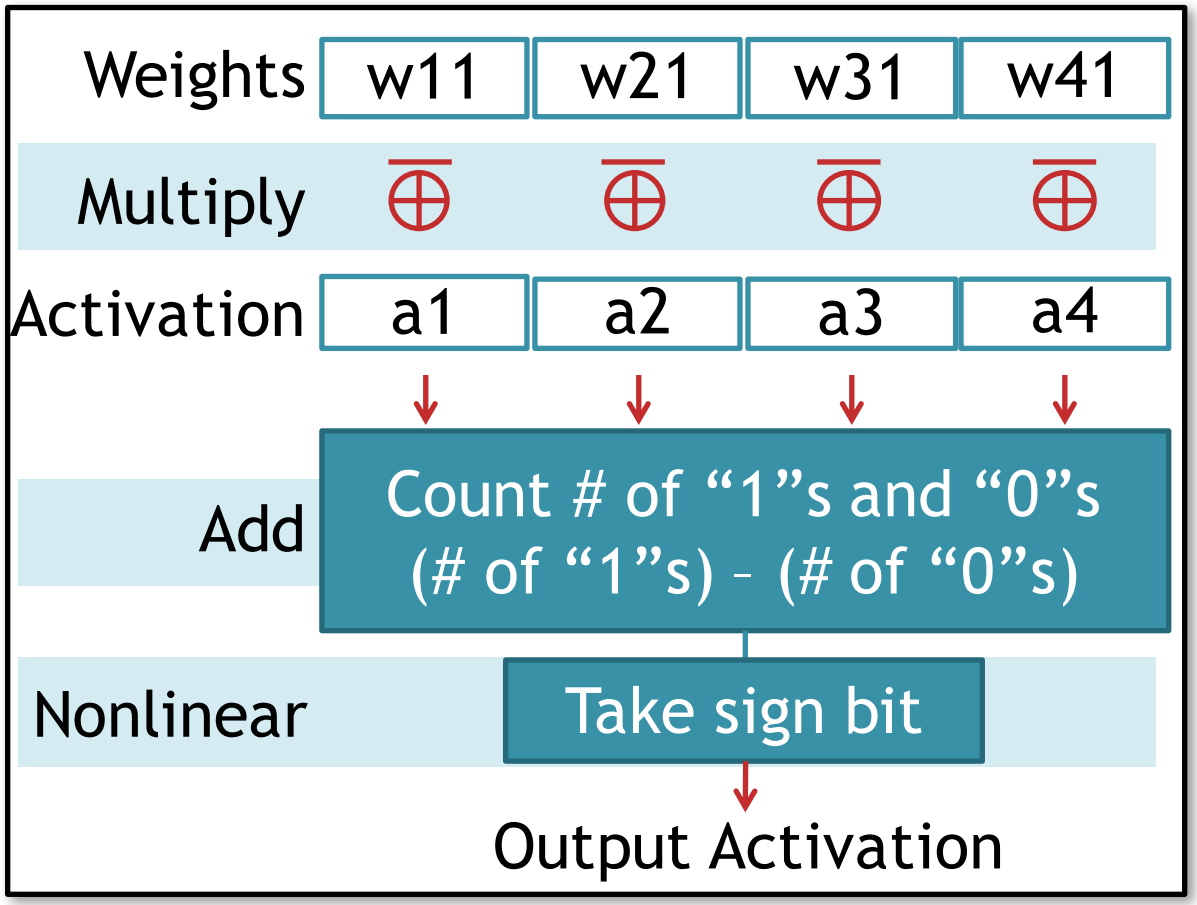


- Target: Energy-efficient binary DNN inference accelerator
- Approach: "Binary DNN" and "Processing in Memory (PIM)"
  - Binary DNN uses low-precision arithmetic operations instead of regular but complex operations, such as BinaryNet, XNOR-net
    - ✓ "XNOR" instead of "multiplication
  - PIM can reduce the data movement energy

Binary Neural Network (BNN)



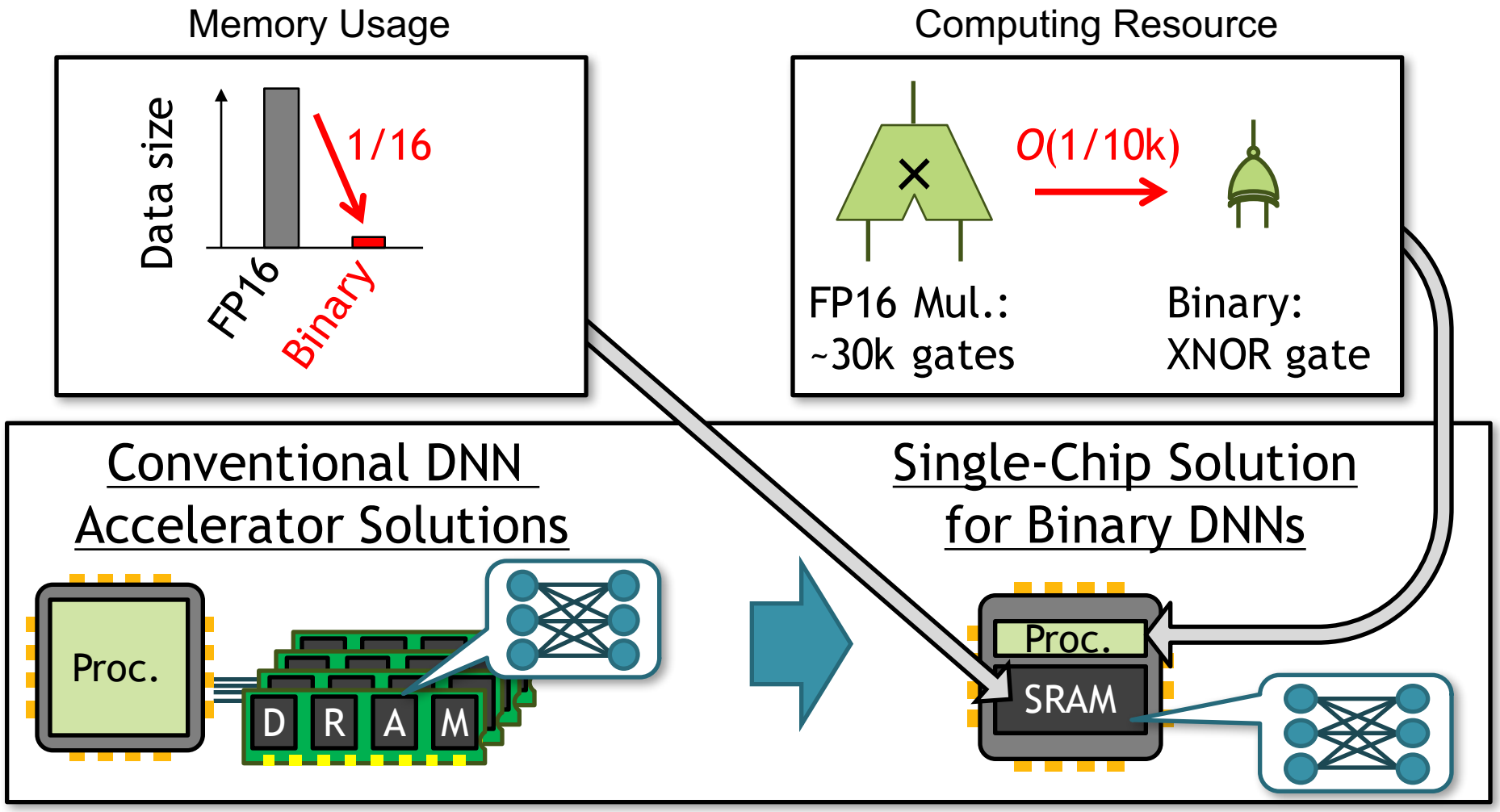
A	W	A×W
1	1	1
1	0	0
0	1	0
0	0	1



Restricts weights and activations to +1 or -1

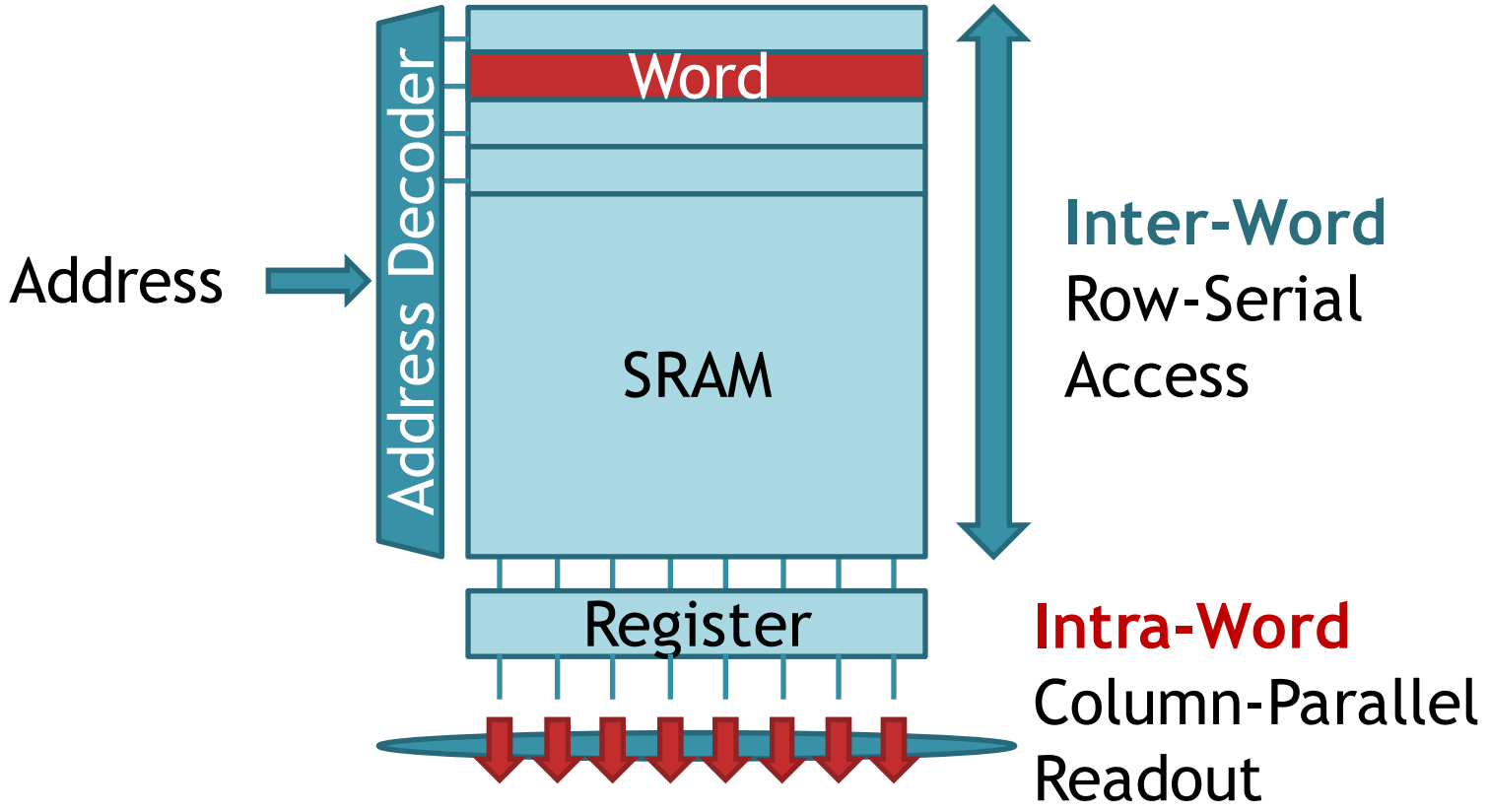
Multiply => Bit-wise XNOR
Add => Counting numbers
Nonlinear Func => Taking sign

Concept: Binary DNN by PIM



- On-chip synaptic weights for excluding costly DRAM accesses
- Light-weight neural processing near SRAMs for enabling massively parallel computation

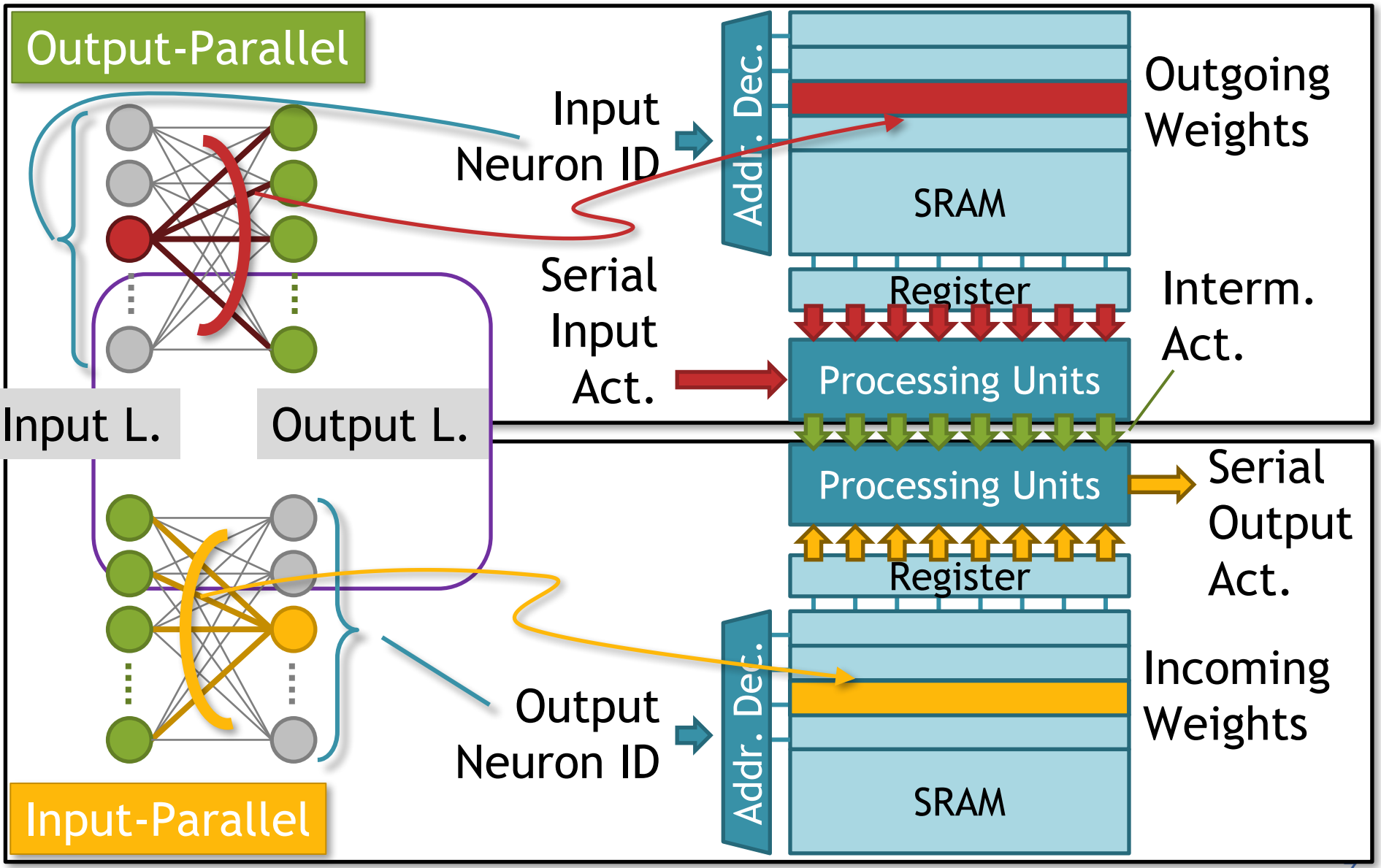
Processing in Memory



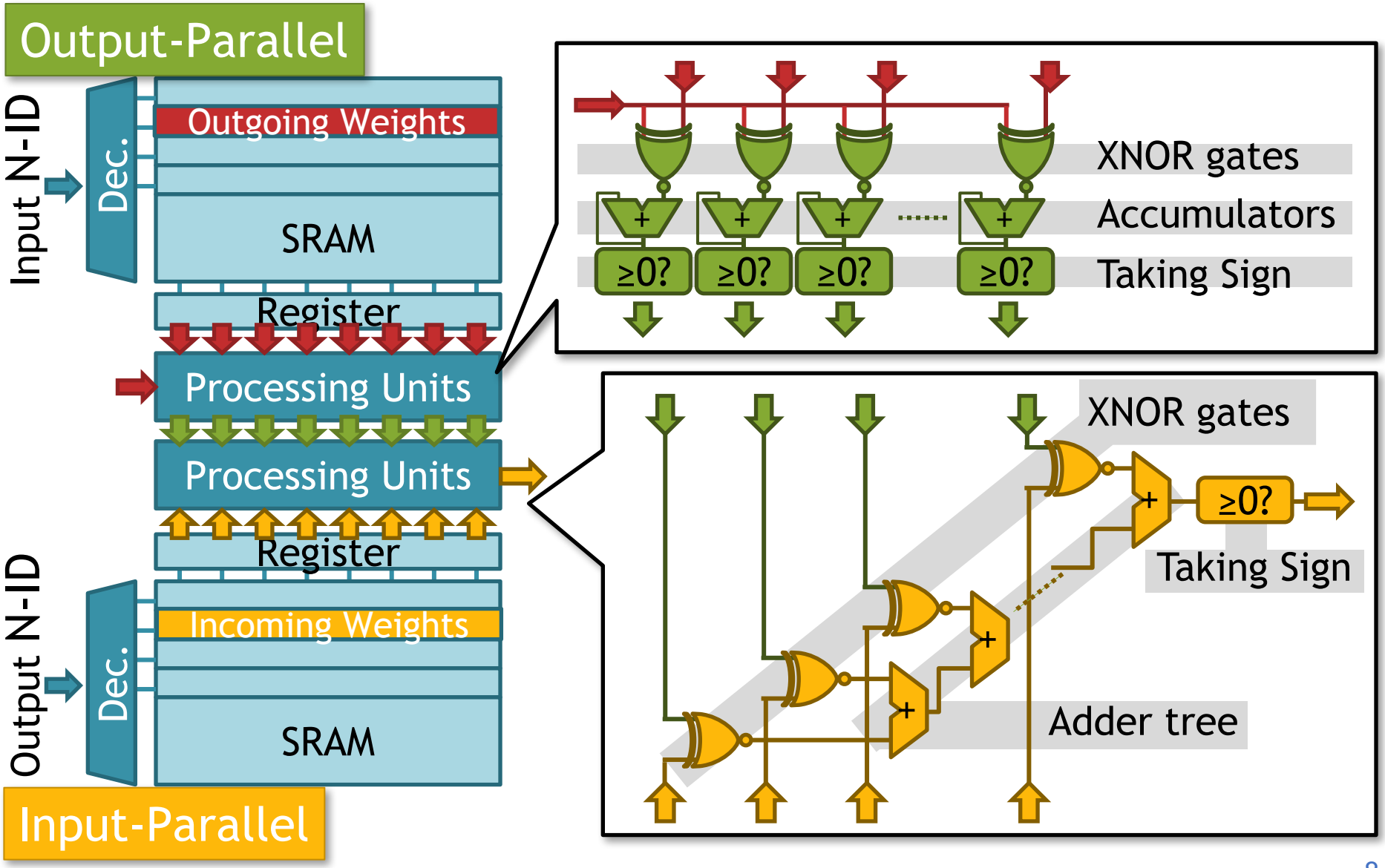
- Fully exploiting intra-word parallelism
- Hiding/tolerating inter-word serial behavior

How can we do that for accelerating binary DNNs?

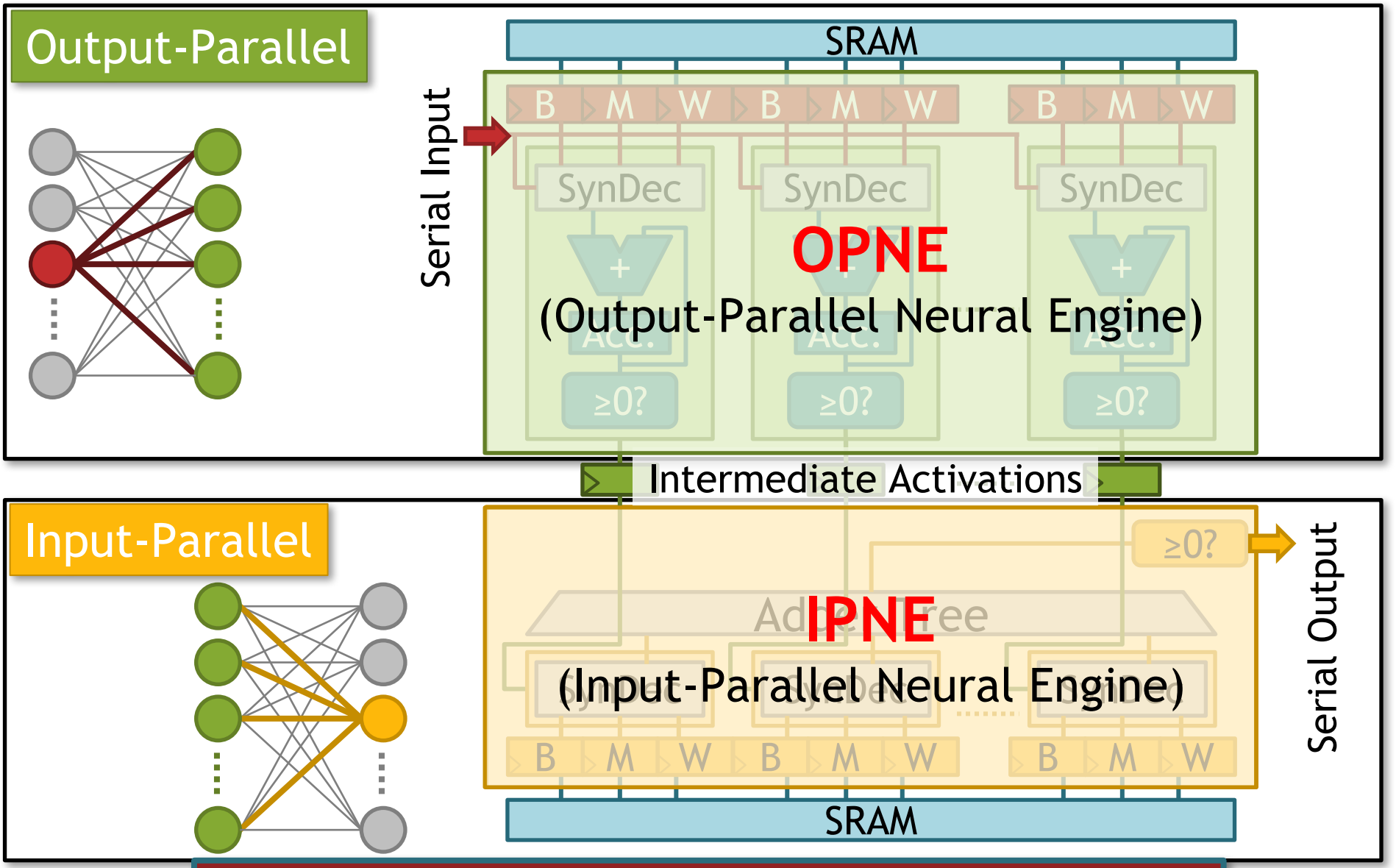
3-layer NN In-Memory Module



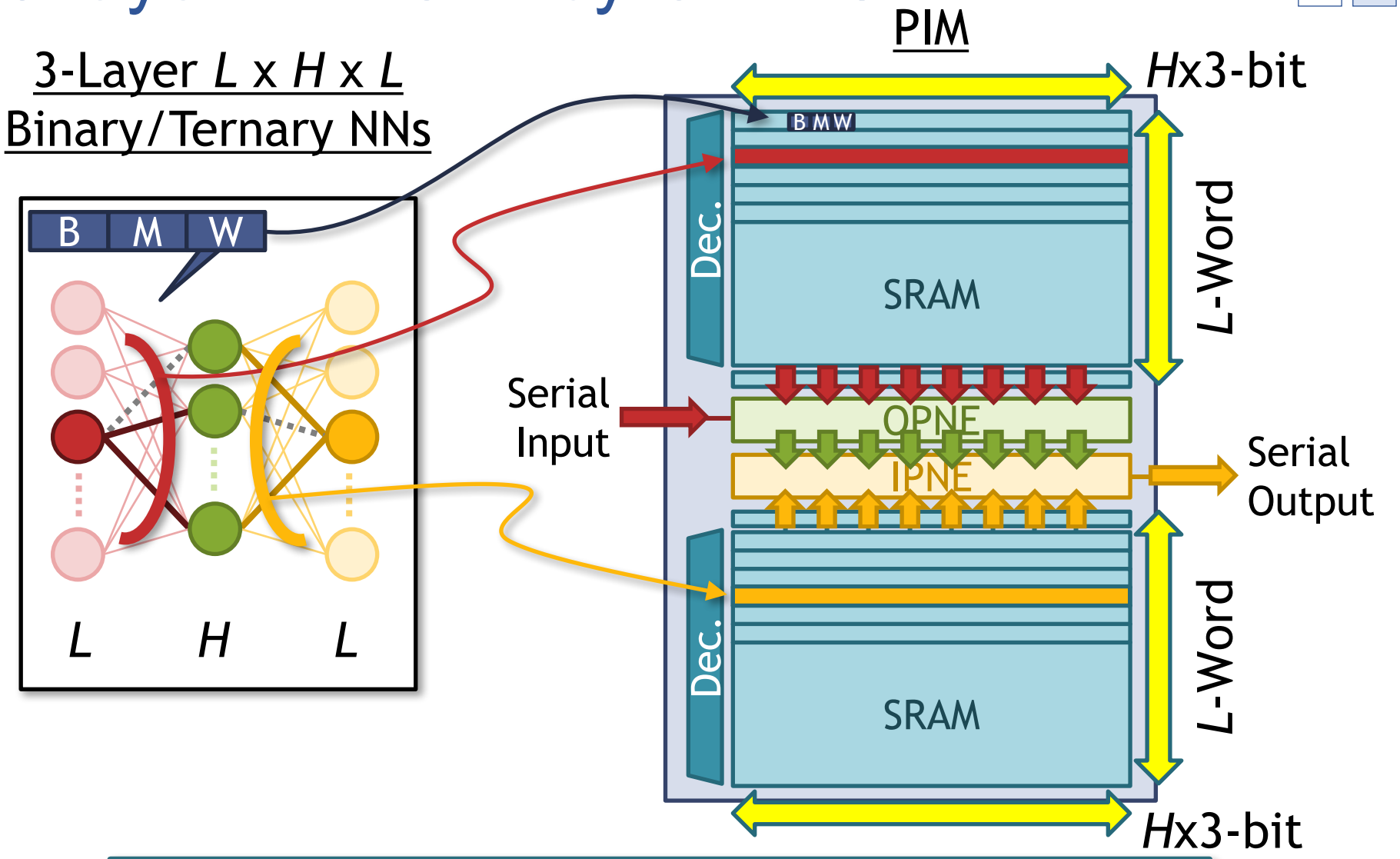
Binary Operation in 3-layer Module



Serial-Parallel-Serial Structure



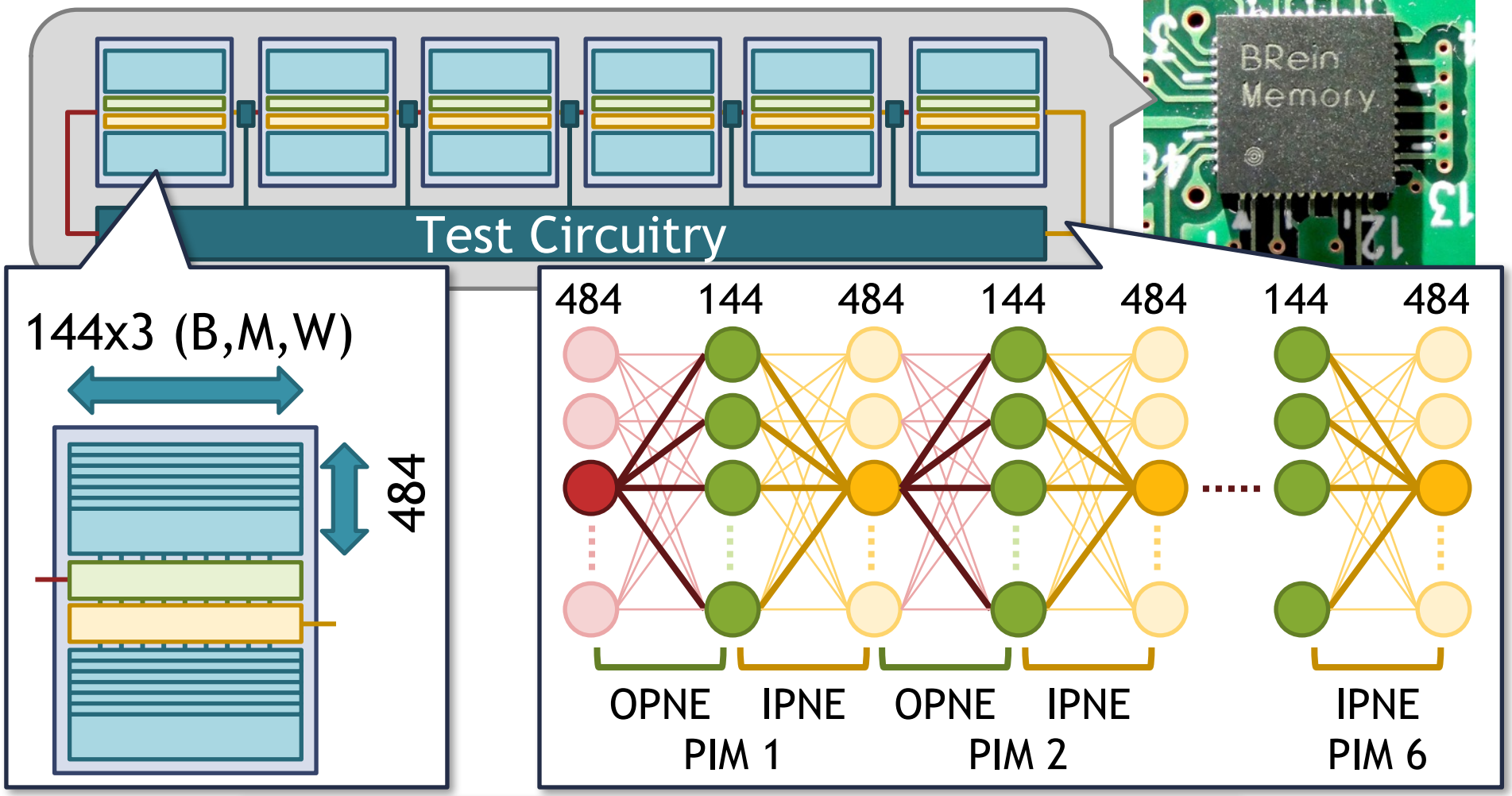
3-layer PIM Unit by OPNE/IPNE



Can Host Versatile Binary/Ternary DNNs under the Maximum Size in a Reconfigurable Manner

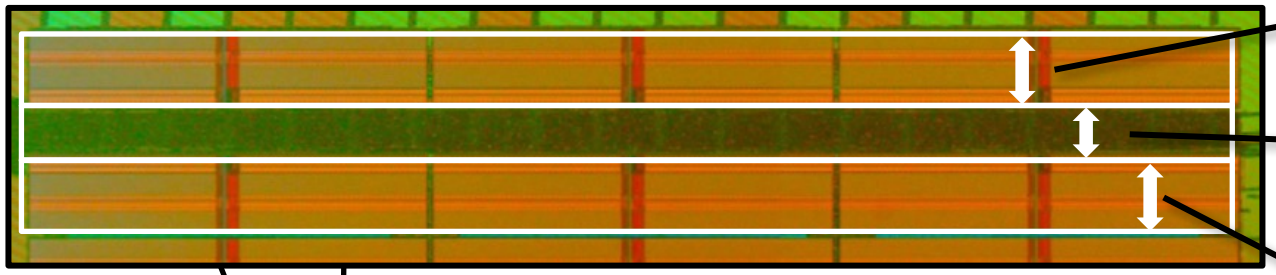
BRein Memory: Overall Architecture

Prototype Chip with 6 PIMs

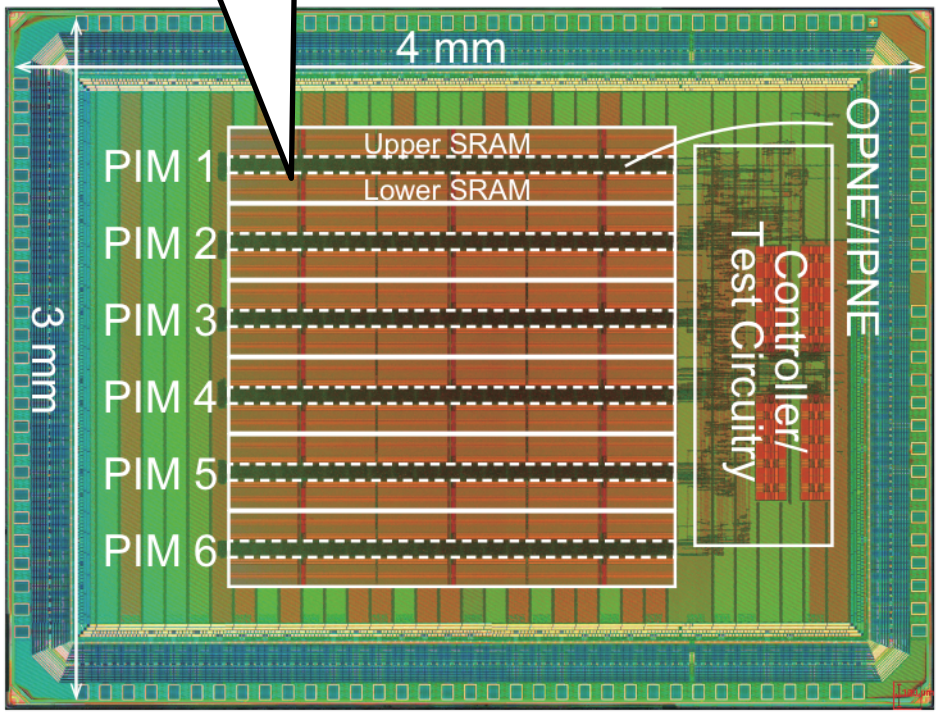


13-Layer, 4.2K-Neuron, 0.8M-Synapses, Binary/Ternary DNN in a Small Chip

BRin Memory: Chip Specs



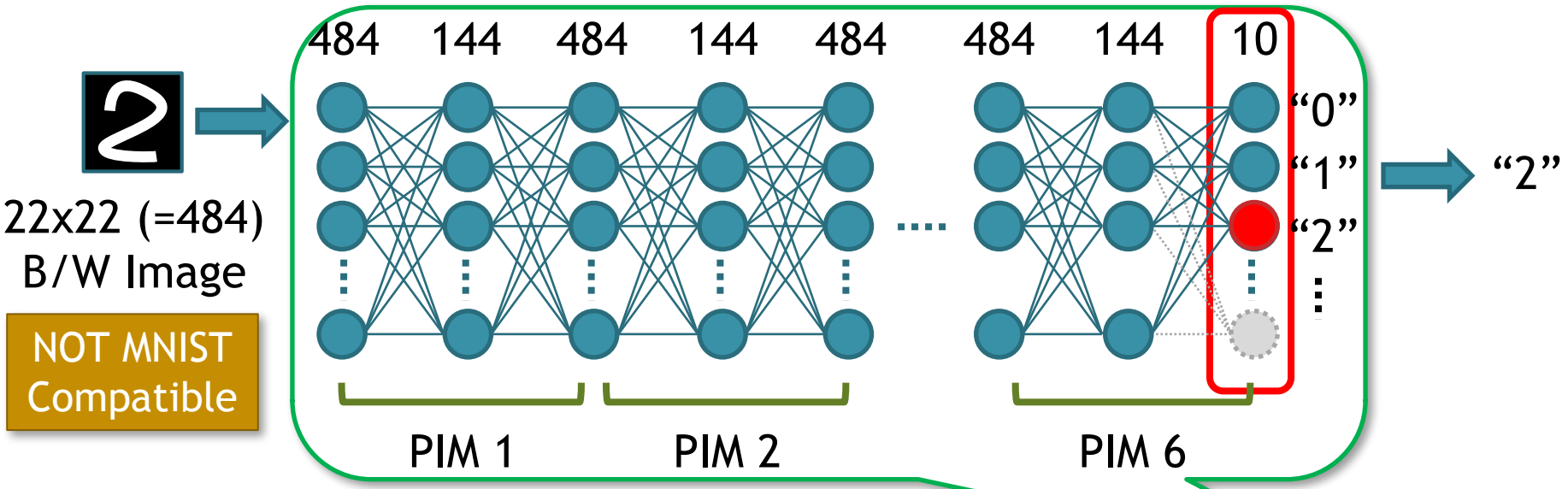
PIM

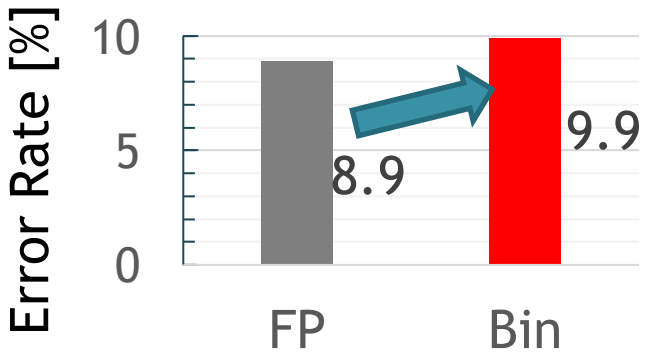


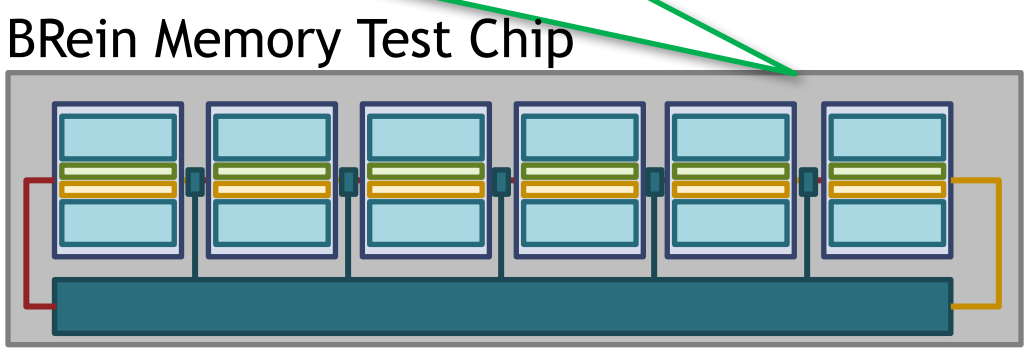
Technology	65nm GP
Die Size	4 x 3 mm^2
Core Size	3.9 mm^2
Operating Frequency	400 MHz
Voltage	1.0 V
Power Consumption	0.58 W

Test Chip Evaluation: Setup

Test Application: Fully-connected 13-layer Binary DNN for Handwritten Digit Recognition



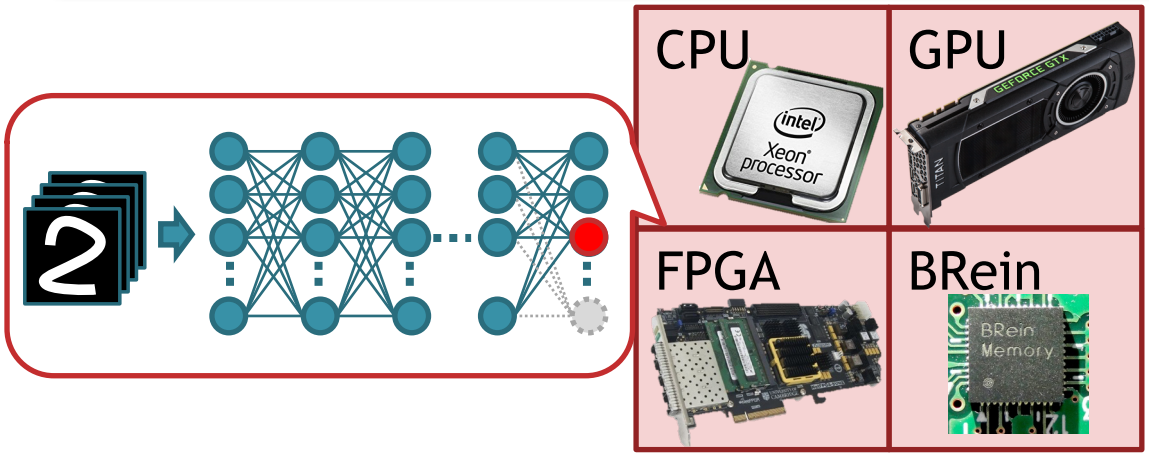




Only 1% degradation by Binarization

Test Chip Evaluation: vs. CPU/GPU/FPGA

Test Application: Fully-connected 13-layer Binary DNN for Handwritten Digit Recognition

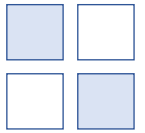


Measured using 1M transactions

CPU: Xeon® E5-2650 v4 ×2
GPU: GeForce® GTX TITAN X
FPGA: Virtex®-7 XC7V690T

	Clock [GHz]	Exec. Time [s]	Power [W]	Energy [J]	Perf. [GOPS]	Efficiency [GOPS/W]	Ex.Time Ratio	Energy Ratio
CPU	2.2	124	155	19.2 k	12.4	0.08	102	27.1 k
GPU	1.0	13.8	152	2098	111.3	0.73	11	2966
FPGA	0.2	53	11	583	29.0	2.64	44	825
BRein	0.4	1.22	0.58	0.73	1264.4	2172.4	1	1

Test Chip Evaluation: vs. Existing Chips



	ISSCC 2016 MIT [1]	ISSCC 2017 KAIST [2]	ISSCC 2017 KU Leuven [3]	Our Work
Technology	65nm LP	65nm	28nm FD-SOI	65nm GP
Target	CNN	CNN + RNN	CNN	Versatile DNN
Precision (bits)	16	4 – 16 for CNN 4 – 7 for FC	4 - 16	Binary/Ternary
Frequency [MHz]	200	50 - 200	Typ. 200	100 – 400
Voltage [V]	1.0	0.77 – 1.1	0.6 – 1.1	0.55 – 1.0
Power [W]	0.3	0.03 – 0.28	0.075 – 0.3	0.05 – 0.6
Core Size [mm^2]	12.3	7.4	0.95	3.9
Peak Perf. [TOPS]	0.07	1.25	0.41	1.38
Energy Efficiency [TOPS/W]	0.2	1.0 - 8.1	0.26 – 10	2.3 - 6.0
Area Efficiency [TOPS/mm^2]	0.006	0.010 – 0.169	0.106 – 0.425	0.089 - 0.365

[1] Y. H. Chen, *et al.*, ISSCC 2016
[2] D. Shin, *et al.*, ISSCC 2017
[3] B. Moons, *et al.*, ISSCC 2017

Conclusion

- BRein Memory: a binary/ternary DNN accelerator
  - Employing Processing in Memory (PIM) architecture
- 1st test chip with 6 PIMs
  - Peak 1.38 TOPS @ 400 MHz, Efficiency 2.3 TOPS/W
  - First accelerator chip for binary & ternary DNNs
- The original work is presented at VLSI Circuit 2017:
  - BRein Memory: A 13-Layer 4.2 K Neuron 0.8 M Synapse Binary/Ternary Reconfigurable in-Memory Deep Neural Network Accelerator in 65 nm CMOS
    - ✓ Kota Ando, Kodai Ueyoshi, Kentaro Orimo, Haruyoshi Yonekawa, Shimpei Sato, Hiroki Nakahara, Masayuki Ikebe, Tetsuya Asai, Shinya Takamaeda-Yamazaki, Tadahiro Kuroda, and Masato Motomura