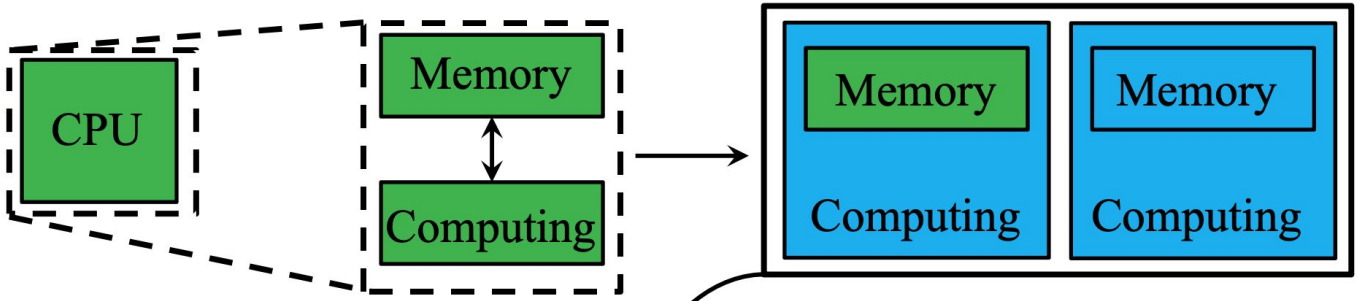
Compute-In-Memory SRAM Comparative Study for Accelerating Low-Precision Neural Network

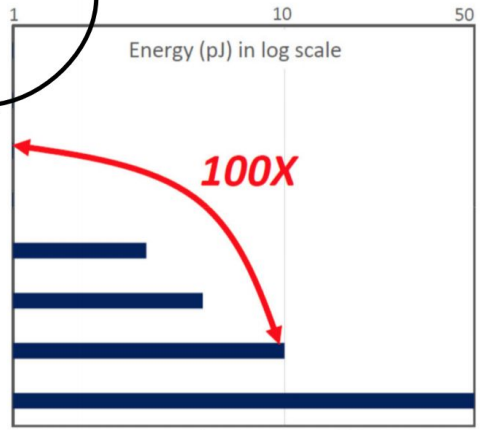
Chuyao Cheng, Xuan Su

Motivation

- **Heavy** computation
- Memory access is frequent and **expensive**



Operation	Energy (pJ)
Integer ADD (8b)	0.03
Integer ADD (16b)	0.05
Integer ADD (32b)	0.1
Integer MULT (8b)	0.2
Integer MULT (32b)	3.1
8KB SRAM Read (32b)	5
32KB SRAM Read (32b)	10
1MB SRAM Read (32b)	50



Peter Kogge,
University of Notre Dame

Motivation

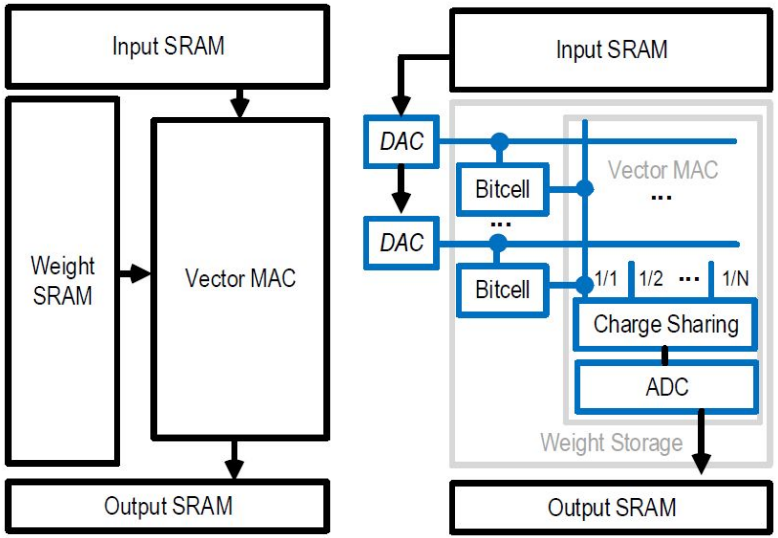
Compute-In-Memory (CIM)

Pros:

- less limited by on-chip BW
- higher energy efficiency and throughput

Cons:

- lower computation accuracy
- less flexibility
- limitations of ADC



Brian Zimmer, Nvidia

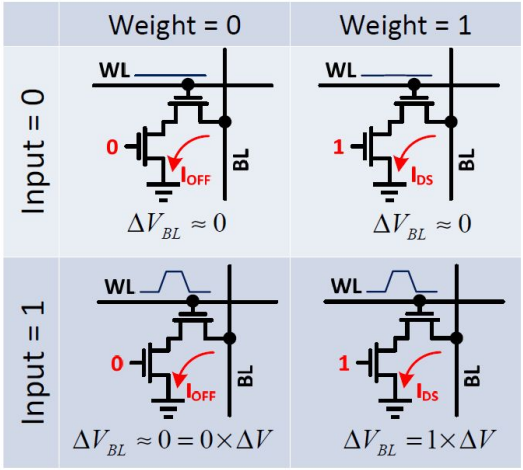
Twin 8T SRAM CIM [1] vs. FinFET 8T SRAM CIM [2]

Architecture & Techniques

- Macro Structure
- SRAM Cell
- Multibit Input and Weight Mapping
- Circuit Techniques (Operation and Output Readout)

Evaluation & Comparison

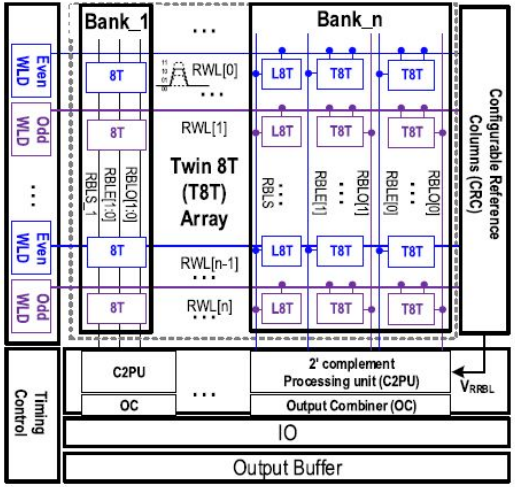
- Multi-Pulse Input vs. Multi-Voltage-Level Input
- Performance of SRAM CIM Macros



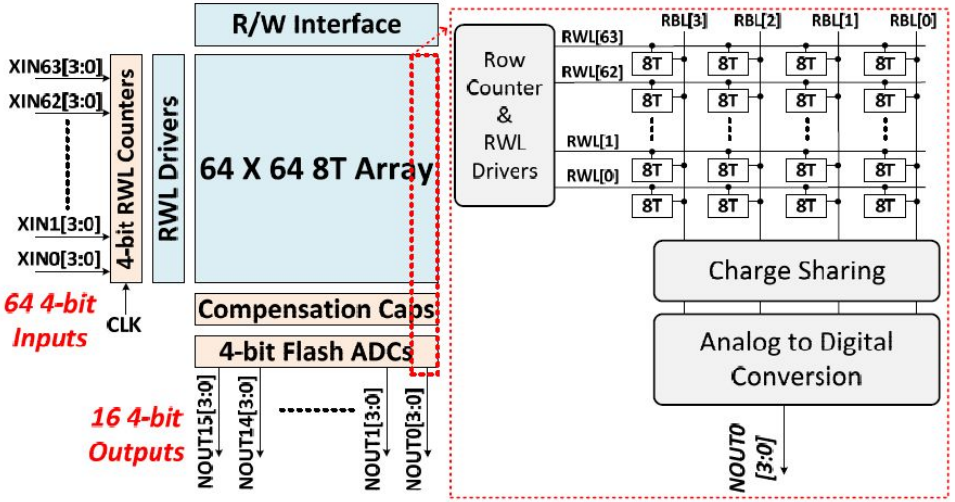
Qing Dong et al., TSMC

Architecture & Techniques

- Macro Structure



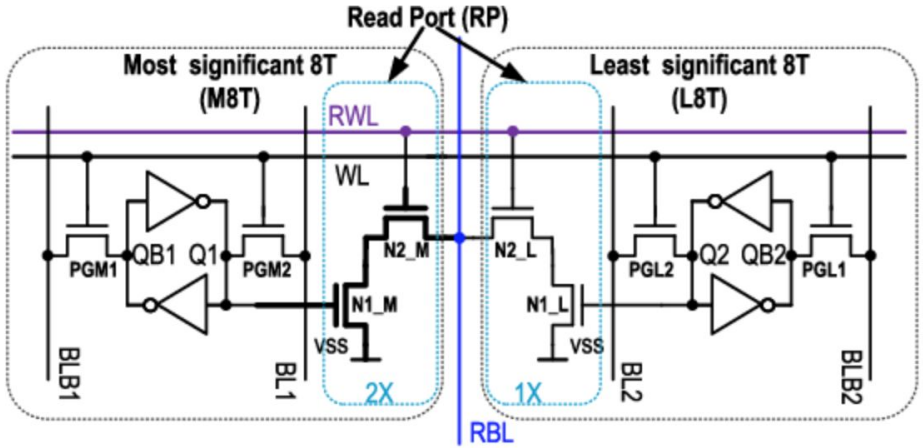
Macro Architecture of the 55nm T8T SRAM CIM



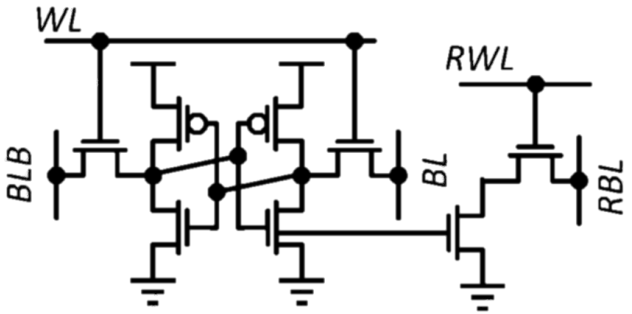
Macro Architecture of the 7nm 8T SRAM CIM

Architecture & Techniques

- SRAM Cell



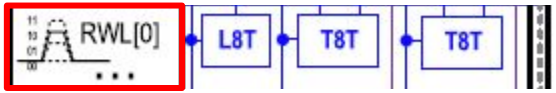
SRAM Cell of the 55nm T8T SRAM CIM



SRAM Cell of the 7nm 8T SRAM CIM

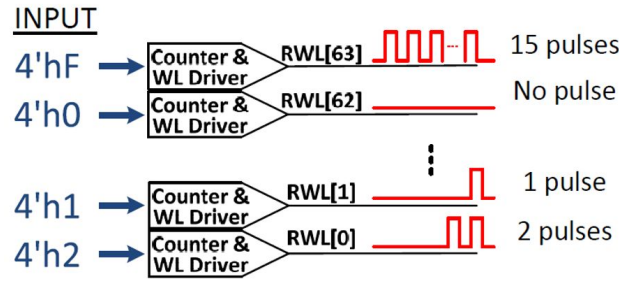
Architecture & Techniques

- Multibit Input - Multi-Voltage vs. Multi-Pulse



Input (2b)	RWL Voltage	Weight		RBL Current	IWP
		M8T	L8T		
11	WLL3	1	1	$3\times I_{MC}$	9
10	WLL2	1	1	$2\times I_{MC}$	6
01	WLL1	1	1	$1\times I_{MC}$	3
11	WLL3	1	0	$2\times I_{MC}$	6

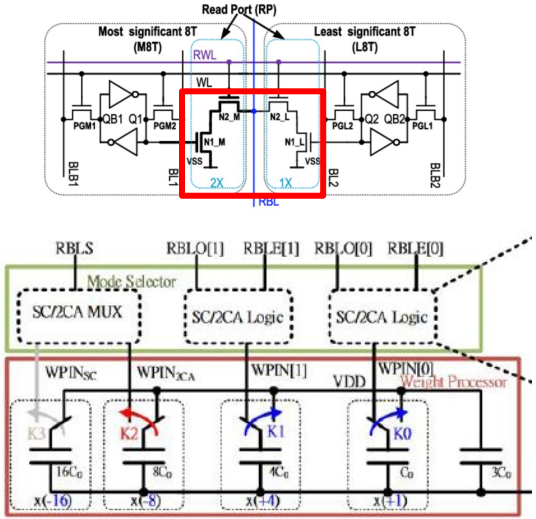
Multiple-voltage-level input mapping of the 55nm T8T SRAM Macro



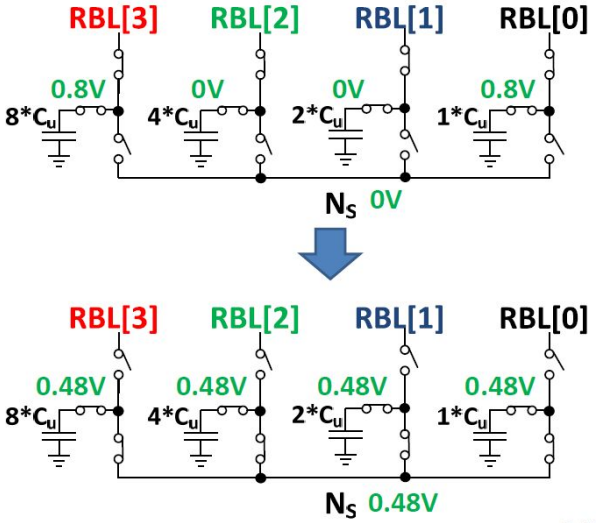
Multiple-pulse input mapping of the 7nm 8T SRAM Macro

Architecture & Techniques

- Multibit Weight - Charge Sharing Techniques



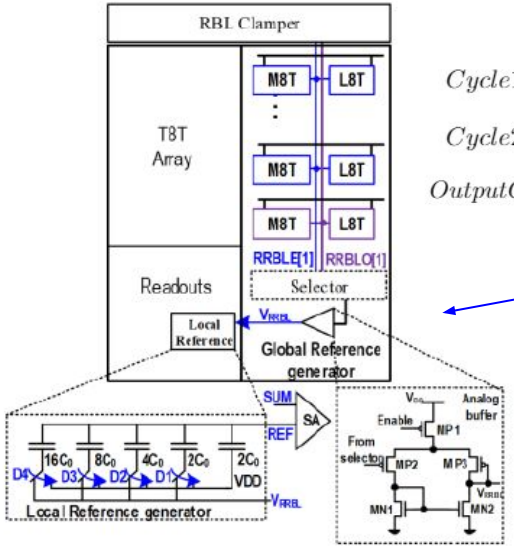
Input and Weight Truth Table of the 55nm T8T SRAM Macro



Input and Weight Mapping of the 7nm 8T SRAM Macro

Architecture & Techniques

- Circuit Techniques (Operation and Output Readout)

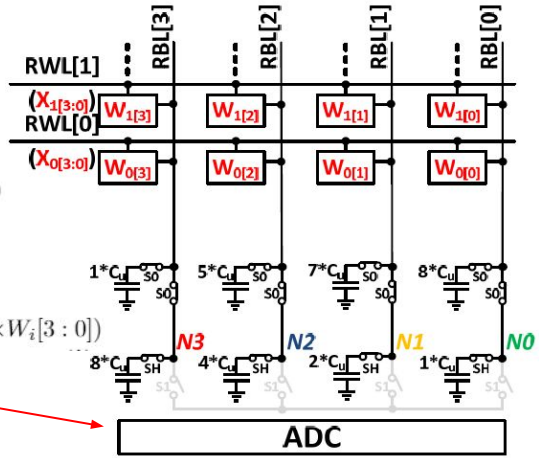


$$Cycle1 : MACV1 = \sum(IN_i[3:2] \times W_i[4:0])$$

$$Cycle2 : MACV2 = \sum(IN_i[1:0] \times W_i[4:0])$$

$$OutputCombiner : FMAC = 4 \times MACV1 + MACV2 = \sum(IN_i[3:0] \times W_i[4:0])$$

$$NOUT[3:0] = 4\text{-}bit\text{-}quantized(\sum(IN_i[3:0] \times W_i[3:0]))$$

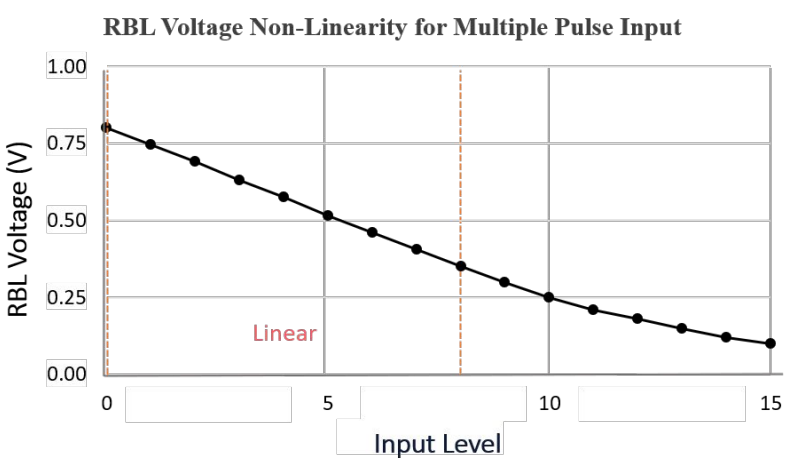


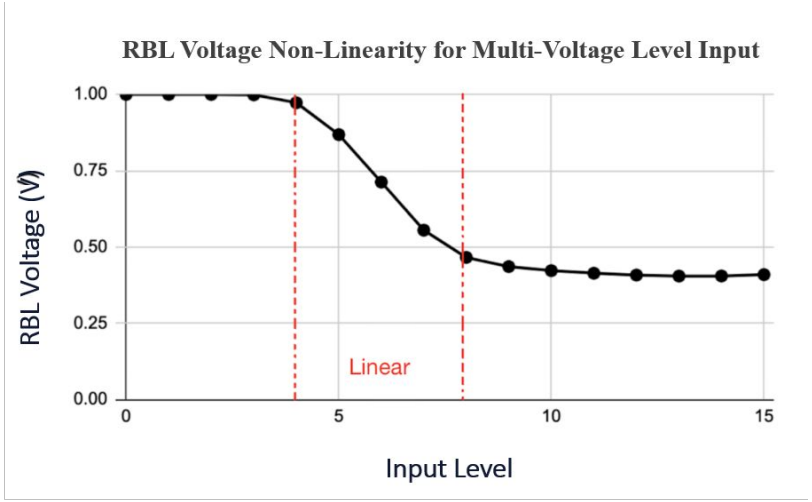
Circuits and operations of the 55nm T8T SRAM Macro

Circuits and operations of the 7nm 8T SRAM Macro

Evaluation & Comparison

- Row Circuitry Linearity



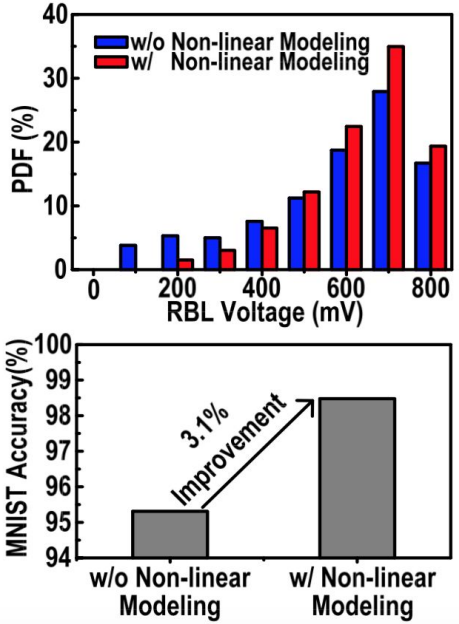


Non-linearity for multiple-pulse input and multi-voltage level input

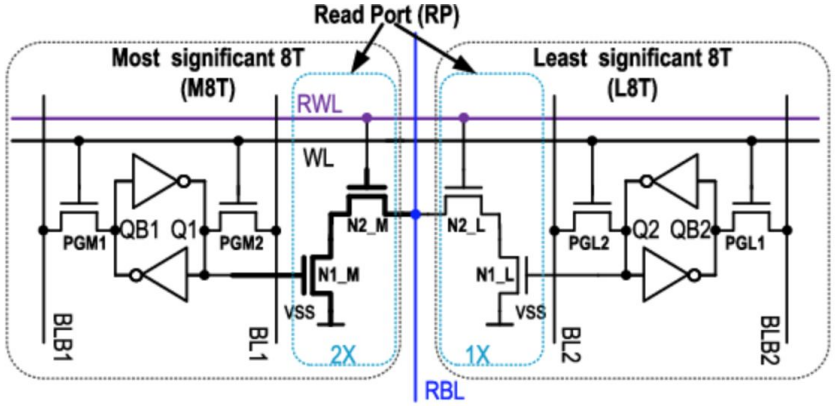
Evaluation & Comparison

- Row Circuitry Linearity Compensation

Incorporates non-linear modeling into activation function

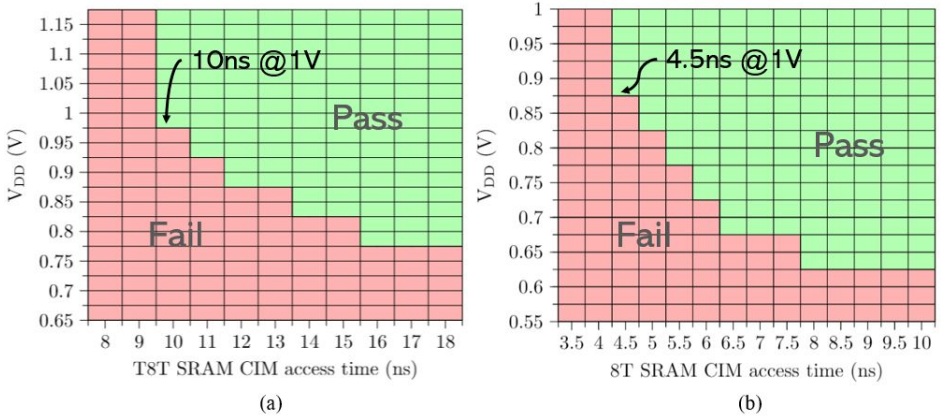


Limit unit cell bit size to 2-bit



Evaluation & Comparison

- SRAM CIM Macros Features



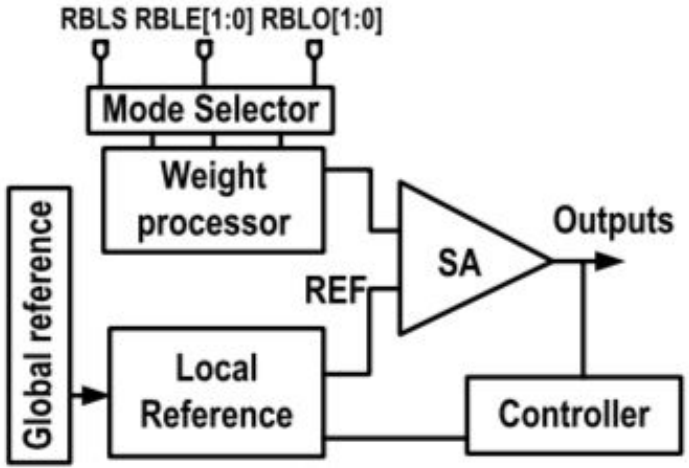
Access time of the T8T SRAM Macro and 8T SRAM Macro

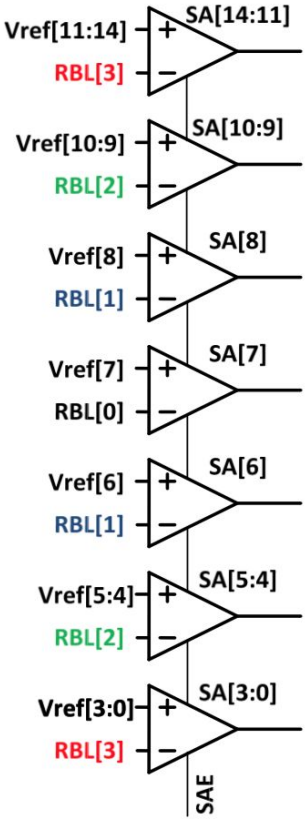
	T8T SRAM CIM	8T SRAM CIM
Technology	55nm	7nm
Array Size	64x60 b	64x64 b
Cell Type	T8T	8T
Bitcell Area (μm^2)	0.865	0.053
Input Precision (bit)	4	4
Weight Precision (bit)	5	4
Output Precision (bit)	7	4
Power Supply (V)	1.0	1.0
Access Time (ns)	10	4.5
Macro Energy (pJ)	11.7	13.1
Energy Efficiency (TOPS/W)	18.4	321 in average
Throughput (GOPS)	21.2	455.1
Accuracy of MNIST	99.52%	98.5%

Features summary and comparison of two SRAM Macros

Evaluation & Comparison

- ADC Overhead





Discussion & Conclusion

CIM Limitation

- ADC overhead and non-linearity scales with number of bits
- Amdahl's Law

Future Direction

- Only use at critical part, in combination with other digital logic
- Take advantage of sparsity; optimized software algorithm
- How to integrate with heterogeneous memories or architecture?

References

[1] X. Si et al., *A Twin-8T SRAM Computation-In-Memory Macro for MultipleBit CNN-Based Machine Learning.* ISSCC, pp. 396-397, Feb. 2019.

[2] Dong, q.et al. *A 351 TOPS/W and 372.4 GOPS Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine Learning Applications.* In: International Solid-State Circuits Conference. IEEE.

[3] Jia, H. et al. *A microprocessor implemented in 65nm CMOS with configurable and bit-scalable accelerator for programmable in-memory computing.*arXiv preprint arXiv:1811.04047.

[4] A. Biswas and A. P. Chandrakasan, *Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNNbased machine learning applications.* In IEEE ISSCC Dig. Tech. Papers, Feb. 2018, pp. 488–489.

[5] H. Valavi et al., *A mixed-signal binarized convolutional-neural-network accelerator integrating dense weight storage and multiplication for reduced data movement.* In Proc. IEEE Symp. VLSI Circuits, Jun. 2018, pp. 141–142.

[6] W.-S. Khwa et al., *A 65nm 4Kb algorithm-dependent computing-inmemory SRAM unit-macro with 2.3ns and 55.8TOPS/W fully parallel product-sum operation for binary DNN edge processors.* In IEEE ISSCC Dig. Tech. Papers, Feb. 2018, pp. 496–497.

[7] J. Zhang et al., *In-Memory Computation of a Machine-Learning Classifier in a Standard 6T SRAM Array.* In JSSC, vol. 52, no. 4, pp. 915-924, 2017.

Thank you!