An In-Memory Computing SRAM Macro for Memory-Augmented Neural Network in 40nm CMOS

Sunghoon Kim

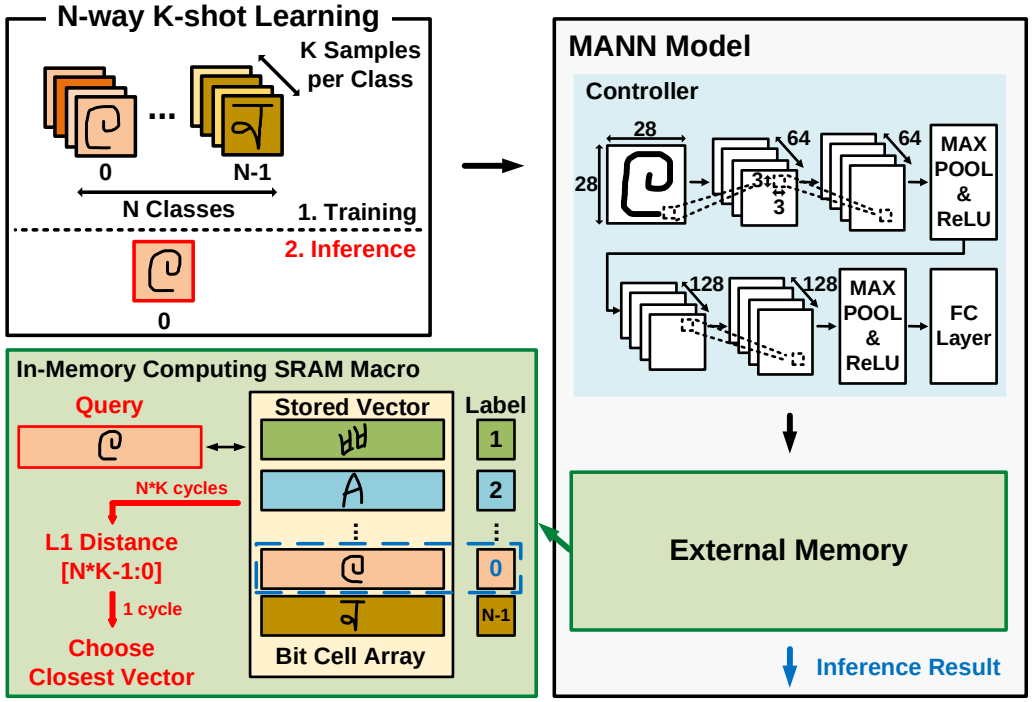
rnd810@snu.ac.kr

Graduate School of Convergence Science and Technology

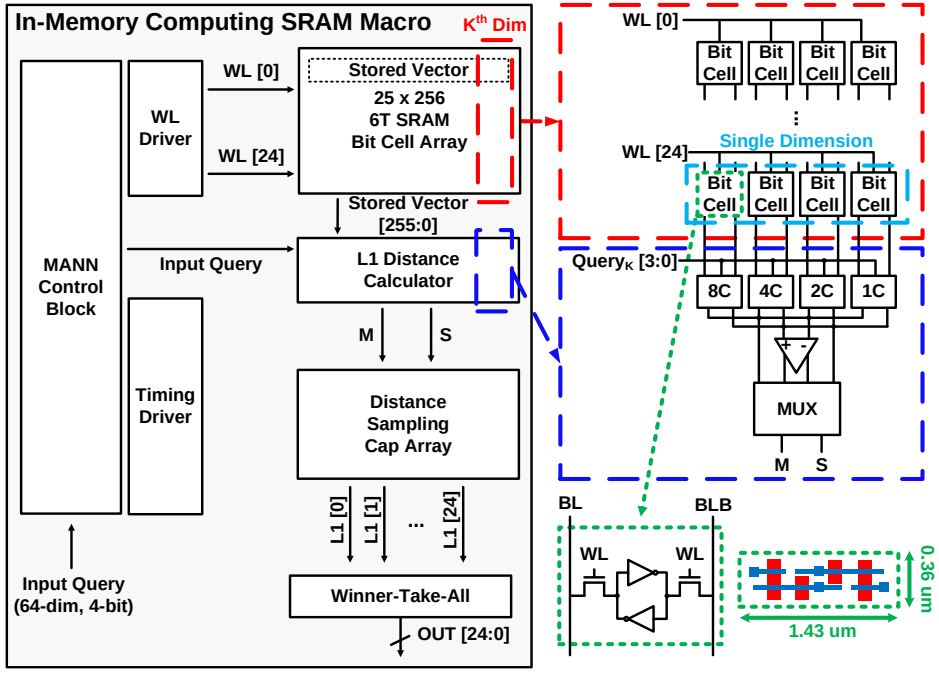
Seoul National University

Mobile Multimedia Systems Group

Motivation and Overall Structure



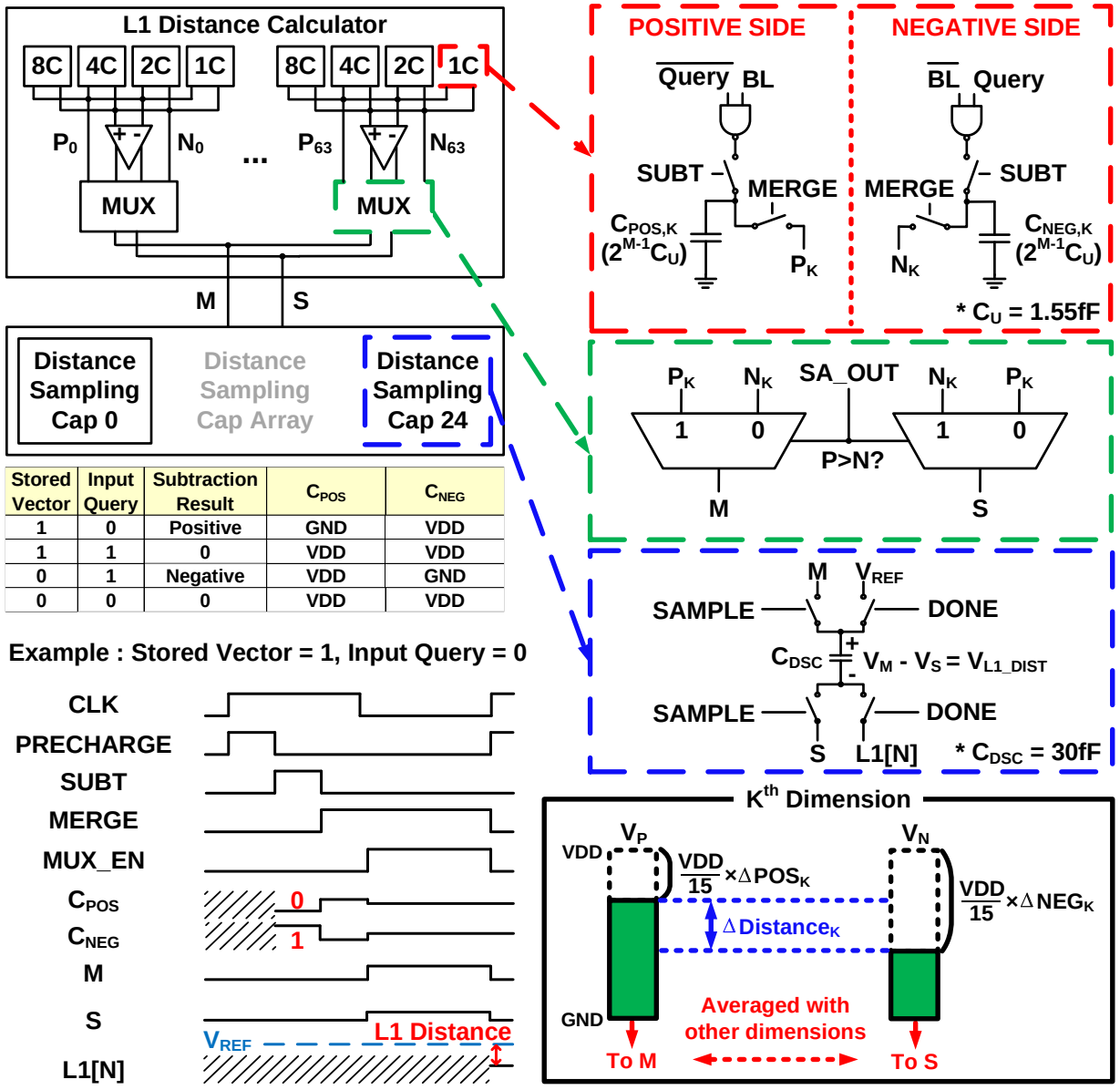
MANN-based Few-Shot Learning



Overall Structure of Proposed SRAM macro

- **Memory-Augmented Neural Network (MANN) introduces additional memory in training and inference**
  - **Similarity-based search → Huge memory access**
- **We propose an IMC SRAM macro for accelerating MANN-based few shot learning inference**

L1 Distance Calculation Circuit

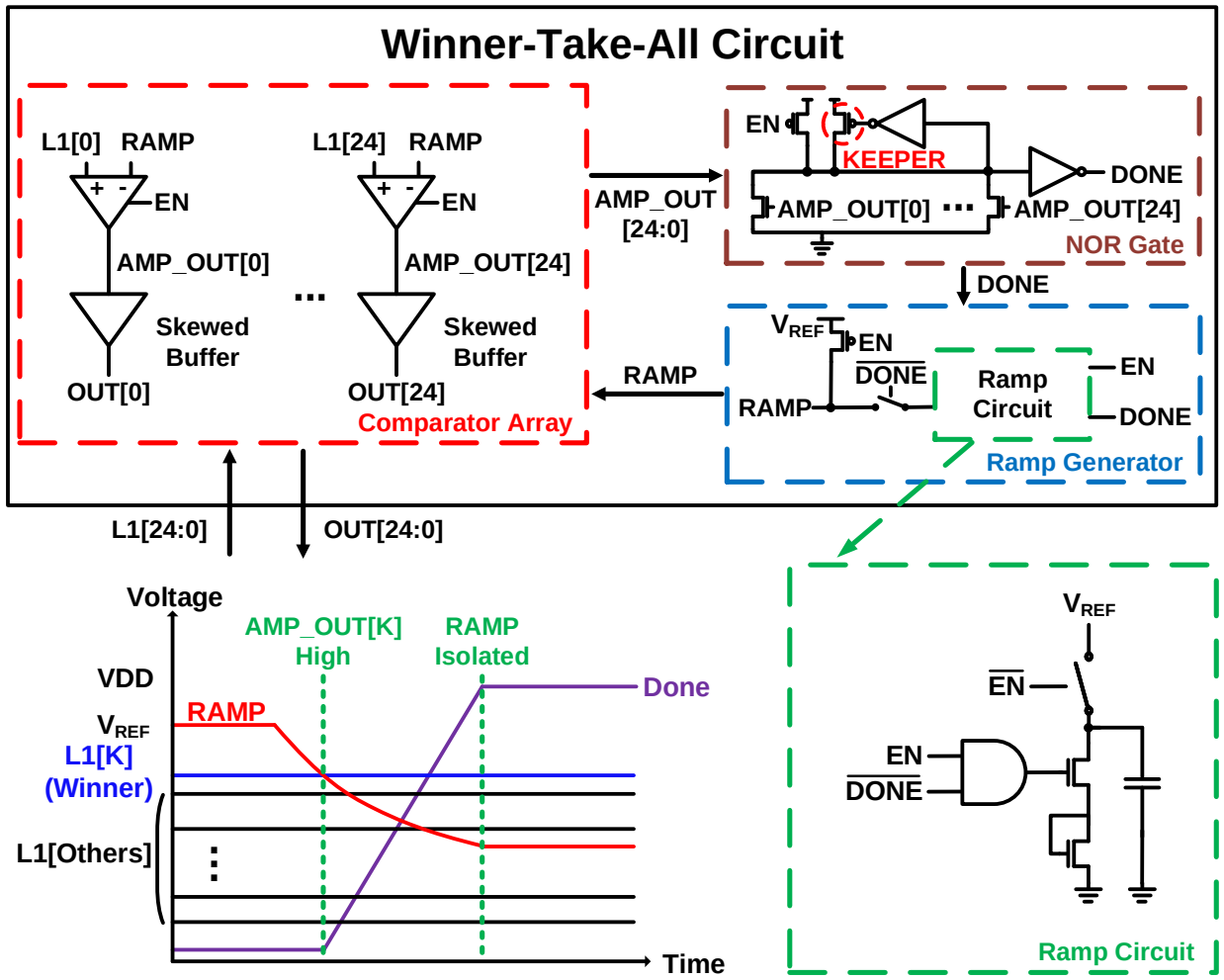


Stored Vector	Input Query	Subtraction Result	C_{POS}	C_{NEG}
1	0	Positive	GND	VDD
1	1	0	VDD	VDD
0	1	Negative	VDD	GND
0	0	0	VDD	VDD

Example : Stored Vector = 1, Input Query = 0

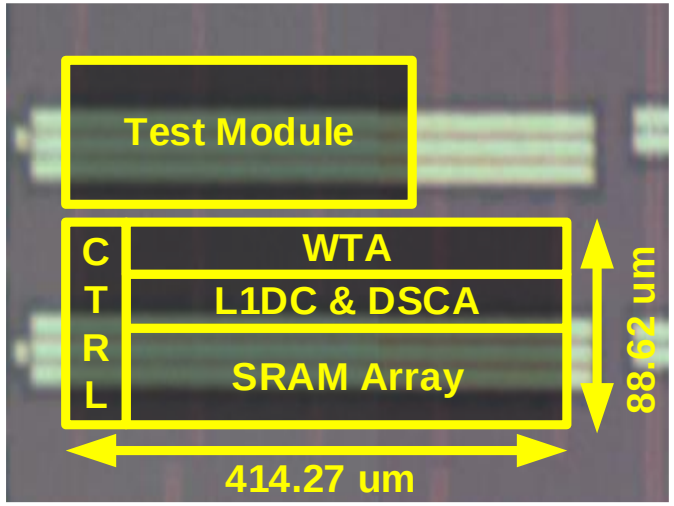
- Each weighted bits conditionally charge/discharge the capacitor on pos/neg side
- Bit-wise subtraction results are merged to global M, S node
- The top plate of every DSC is connected to V_{REF}
  - Voltage level of bottom plate becomes V_{REF}-$L1_{dist}$

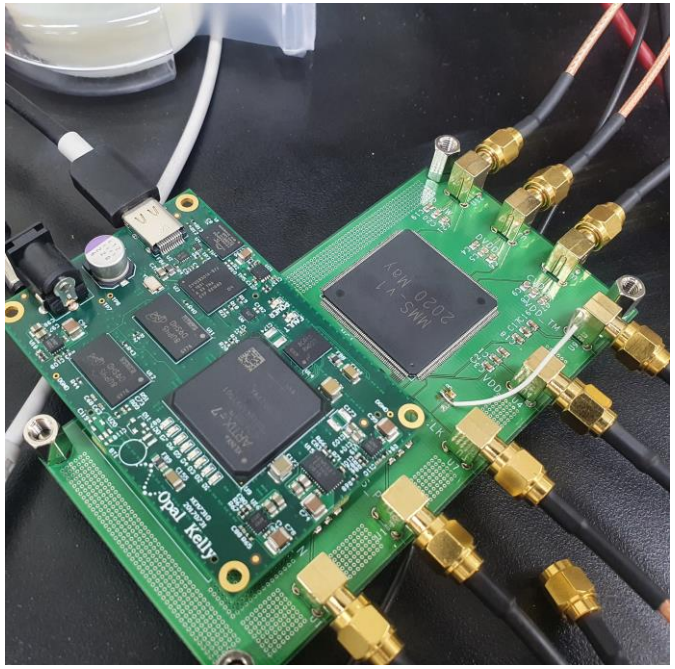
Winner-Take-All Circuit

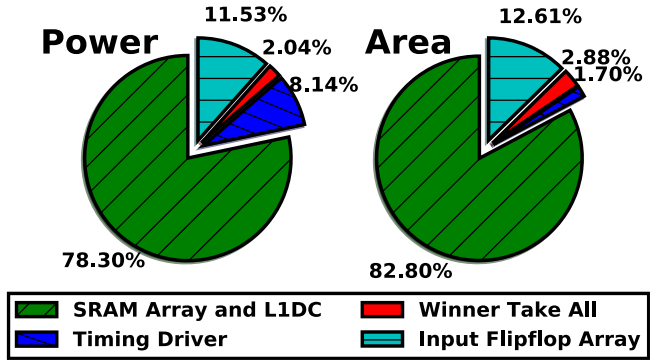


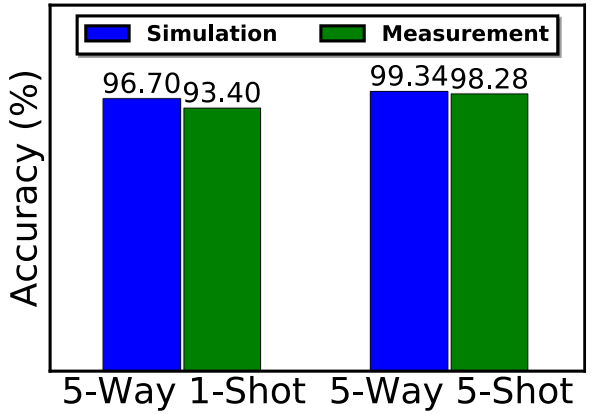
- WTA circuit compare the L1 distances with the global ramp signal in parallel using OTAs
- The feature with smallest L1 distance fires OUT signal and disable ramp circuit

Measurement Results









Comparison table with IMC SRAM macros

	[1]	[2]	[3]	[4]	[5]	This Work
Technology	28nm	7nm	65nm	28nm	55nm DDC	40nm LP
Cell Type	6T	8T	12T	8T	4+2T	6T
SRAM Bitcell Area (um^2)	1.68 x 6T	0.053	3.915	0.782	1.12 x 6T	0.527
Core Area (mm^2)	-	0.0032	0.1125	2.7	0.0233	0.0367
Array Size	64Kb	4Kb	16Kb	16Kb	16Kb	6.25Kb
Precision	2 - 8b Input 4 - 8b Weight 10 - 20b Output	4b Input 4b Weight 4b Output	1b Input 1b Weight 3.46b Output	Arbitrary	-	4b
Supported Operation	MAC	MAC	XOR	Logic/Add/Sub /Mult/Div/FP	SRAM/CAM/Logic	L1 Distance / Closest Vector Search
Supply Voltage (V)	0.85 - 1	0.8 - 1.0	0.6 - 1.0	0.6 - 1.1	0.25 - 0.8 (Read / Write) 0.35 - 0.8 (CAM / Logic)	0.53 @ 850KHz 1.1 @ 80MHz
Dataset	CIFAR-10 / CIFAR-100	MNIST	MNIST / CIFAR-10	-	-	Omniglot
Energy Efficiency (TOPS/W)	61.1 - 56.1 (Forward) 15.0 - 14.1 (Backward)	262.3 - 610.5	403	0.55 (Mult) 5.27 (Add)	-	27.7