15.4 A 22nm 2Mb ReRAM Compute-in-Memory Macro with 121-28TOPS/W for Multibit MAC Computing for Tiny AI Edge Devices

Cheng-Xin Xue, Tsung-Yuan Huang, Je-Syu Liu, Ting-Wei Chang,
Hui-Yao Kao, Jing-Hong Wang, Ta-Wei Liu, Shih-Ying Wei,
Sheng-Po Huang, Wei-Chen Wei, Yi-Ren Chen, Tzu-Hsiang Hsu,
Yen-Kai Chen, Yun-Chen Lo, Tai-Hsing Wen, Chung-Chuan Lo,
Ren-Shuo Liu, Chih-Cheng Hsieh, Kea-Tiong Tang, Meng-Fan Chang

National Tsing Hua University, Hsinchu, Taiwan

Nonvolatile computing-in-memory (nvCIM) can improve the latency (t_{AC}) and energy-efficiency (EF_{MAC}) of tiny AI edge devices performing multiply-and-accumulate (MAC) computing after system wake-up. Prior nvCIMs have proven effective for binary input (IN) and weight (W), and 3b output (OUT) [1], 1-8-1b IN-W-OUT [2], and 2-3-4b IN-W-OUT [3] neural networks; however, the higher precision (4-4b IN-W) for MAC operations is needed for multi-bit CNNs to achieved high-inference accuracy [4]. As Fig.15.4.1 shows, improving the precision of nvCIM macros involves various challenges. (1) A large number of activated WLs provides a wide range of BL current (I_{BL}) resulting in an inaccurate BL-clamping voltage (V_{BLC}); as well as a large I_{BL} requiring a large array area due to the need for wide metal lines to support high-current density. (2) Previous "WL = input" approaches suffer from: (a) few parallel inputs (IN#) due to (1), and (b) long t_{AC} in multiple cycles of binary WL inputs on 1T1R cells for multibit inputs. (3) Previous positive-negative-split weight-mapping consumes high total I_{BL} and area overhead (needing 2×(m-1) cells for a signed m-bit weight) for cell arrays with high-weight precision. (4) Long t_{AC} and a large number of reference currents (IREF#) for high-precision outputs. To overcome these challenges, this work proposes: (1) a BL-IN-OUT multibit computing (BLIOMC) scheme using a single WL-on and input-aware multibit BL clamping (IA-MBC) to shorten t_{AC} for multibit inputs, increase IN#, and reduce the I_{BL} range/size for accurate V_{BLC} and a compact array area. (2) Scrambled 2's complement (S2C) weight mapping (S2CWM), input-aware source-line (SL) voltage biasing (IA-SLVB), and an S2C value combiner (S2CVC) to reduce area overhead and I_{BL} in the cell array. (3) A dual-bit small-offset current-mode sense amplifier (DbSO-CSA) to reduce IREF# and t_{AC}. A fabricated 22nm 2Mb ReRAM-CIM macro presents the first 4b-input nvCIM macro, featuring a 9.8-18.3ns t_{AC} and an EF_{MAC} of 121.3-28.9TOPS/W from binary to 4bIN-4bW-11bOUT compute precisions.

Figure 15.4.2 shows the BLIOMC and S2CWM schemes. Unlike positive-negative-split weight mapping [1-3], the proposed scheme using the 2's complement approach requires only four 1T1R cells to store each signed 4b-weight (W[3:0]). Unlike [4], the even-bits (W[2] and W[0]) are placed in the half-value group (HVG), whereas the odd bits (W[3] and W[1]) are placed in the full-value group (FVG). The 8-to-1 column-mux selects one of the 8 BLs to connect to a dataline (DL). A 4b-input (IN[3:0]) is split into two sequential 2b signals (IN[1:0] and IN[3:2]) as the inputs (IN-BC) of IA-MBC and IA-SLVB. In a 2bIN-4bW MAC operation, the 2b IN-BC selects a BL-clamping reference voltage ($V_{BLC\text{-}REF}$) that provides the BL clamper output 4 levels of V_{BLC} (V_{RD}, 2/3×V_{RD}, 1/3×V_{RD}, or 0V). Each IA-MBC and IA-SLVB is shared by the 4 accessed BLs/DLs, including 2 from FVG (BL[0]=DL[3] and BL[8] = DL[1]) and 2 from HVG (BL[16] = DL[2] and BL[24] = DL[0]). The SL voltage (V_{SL}) for FVG is at 0V, whereas the V_{SL} of HVG is biased at 1/2×V_{BLC}. When a row is selected (WL = 1), I_{BL} is determined by the partial-product (pPD) of input-aware V_{BLC} and the data (e.g. W[0]) stored in the memory cell (pPD = IN[1:0]×W[0]). For example, if IN[1:0] = 3 and W[3:0] = 7, then $V_{BLC} = V_{RD}$ for DL[3:0], $V_{SL\text{-}FVG}$ = 0V for DL[3, 1], and $V_{SL\text{-}HVG}$ = 1/2×V_{RD} for DL[2, 0], where $I_{DL[3]} = I_{HRS}$, $I_{DL[2]}$ = 1/2×I_{LRS}, $I_{DL[1]} = I_{LRS,}$ $I_{DL[0]}$ = 1/2×I_{LRS}. The I_{LRS} and I_{HRS} refer to cell currents at $V_{BL} = V_{RD}$ and V_{SL} = 0V for reading LRS and HRS cells, respectively.

Figure 15.4.3 shows the operations of S2CVC comprising 1 sign-detector logic (SDL), 1 sign-bit mirror-transistor (NP0), 4 source-current mirror-transistors (P0-P3), 3 positive place-value mirror-transistors (PP1-PP3), 3 negative place-value mirror-transistors (NN1-NN3), and 3 negative mid-stage current-mirror pairs (MP1-MN1, MP2-MN2, MP3-MN3). PP1, PP2, and PP3 respectively represent the place-values of (+4), (+2), and (+1), whereas NP0, NN1, NN2, and NN3 respectively represent the place-values of (+8), (-4), (-2), and (-1). At the initiation of a MAC operation, SDL detects the $I_{DL[3]}$ and generates POSEN and NEGEN. If a weight is positive (POSEN = 1 and NEGEN = 0), then NN1-NN3 are off and PP1-PP3 mirror each DL current ($I_{DL[2:0]}$), as follows: $I_{WDLP[2]}$ = (1/2)×$I_{DL[2]}$ (for 4×IN[1:0]×W[2]); $I_{WDLP[1]}$ = 1/8×$I_{DL[1]}$ (for 2×IN[1:0]×W[1]); $I_{WDLP[0]}$ = 1/8×$I_{DL[0]}$ (for IN[1:0]×W[0]), then the current at node C2SUM is I_{C2SUM} = 1/2×$I_{DL[2]}$+1/8×$I_{DL[1]}$+1/8×$I_{DL[0]}$. If a weight is negative (POSEN = 0 and NEGEN = 1), then PP1-PP3 are off, and NP0, NN1-NN3 mirror each DL current ($I_{DL[2:0]}$) as follows: $I_{WDLN[3]}$ = 1/2×$I_{DL[3]}$ (for 8×IN[1:0]×W[3]); $I_{WDLN[2]}$ = -1/2×$I_{DL[2]}$ (for -4×IN[1:0]×W[2]); $I_{WDLN[1]}$ = -1/8×$I_{DL[1]}$ (for -2×IN[1:0]×W[1]); $I_{WDLN[0]}$ = -1/8×$I_{DL[0]}$ (for -IN[1:0]×W[0]), then I_{C2SUM} = 1/2×$I_{DL[3]}$-1/2×$I_{DL[2]}$-1/8×$I_{DL[1]}$-1/8×$I_{DL[0]}$. Note that we used a (1/2) down-scaling current-mirror ratio to suppress read-path current. The low-power serial-input weighted combiner (LP-SIWC) combines the I_{C2SUM} of the two input phases (IN[1:0] and IN[3:2]) and outputs the combined partial-MAC current (I_{PSUM}, for Σ(INi [3:0]×Wi [3:0]) to a DbSO-CSA.

Figure 15.4.4 shows the operation of DbSO-CSA using 2 I_{REF} (I_{REF_H} and I_{REF_L}) to detect I_{PSUM} and simultaneously output 2b to reduce t_{AC} and IREF#. In standby mode, SW1~SW4 = N0 = N1 = on, PRE = S1 = 0, and the voltages ($V_Q/V_{QB}/V_{Q2}/V_{QB2}$) of node Q/QB/Q2/QB2 are at 0V. In phase-1 (PH1, V_{TH} and I-sampling), SW5 = SW6 = S1 = off, and PRE = SW1~SW4 = P3 = on. The threshold voltage ($V_{TH\text{-}P4}/V_{TH\text{-}P5}$) of P4/P5 is stored at GP4/GP5 ($V_{GP4} = V_{Q2} = V_{DD}\text{-}V_{TH\text{-}P4,}$ $V_{GP5} = V_{QB2} = V_{DD}\text{-}V_{TH\text{-}P5}$), Q2/QB2, and C2/C3. I_{PSUM}/I_{REF_H} is sampled by storing the gate-source voltage ($V_{GS\text{-}P1}/V_{GS\text{-}P2}$) of P1/P2 on C0/C1. In phase-2 (PH2, I-subtraction and V-coupling), PRE = SW1~SW4 = off and S1 = 1. For period T_{PH2}, node Q/QB has voltage swing ($\Delta V_Q/\Delta V_{QB}$) due to $T_{PH2}\times(I_{PSUM}\text{-}I_{REF\_L})/T_{PH2}\times(I_{REF\_H}\text{-}I_{PSUM})$. C2 couples ΔV_Q to GP4 ($V_{GP4} = V_{DD}\text{-}V_{TH\text{-}P4}+\Delta V_{GP4}$, ΔV_{GP4} ~= ΔV_Q), and C3 couples ΔV_{QB} to GP5 ($V_{GP5} = V_{DD}\text{-}V_{TH\text{-}P5}+\Delta V_{GP5}$, ΔV_{GP5} ~= ΔV_{QB}). If ΔV_{GP4}>0 and ΔV_{GP5}>0, then P4 and P5 are both off and there is no voltage swing at node Q2 and QB2 ($\Delta V_{Q2} = \Delta V_{QB2}$ = 0). If ΔV_{GP4}<0 and ΔV_{GP5}>0, then P4 is on to raise V_{Q2}, while $V_{QB2} = V_{DD}\text{-}V_{TH\text{-}P5}$. In phase-3 (PH3, inside/outside detection), the N-amplifier (P3-P5, N2-N4) is on to amplify V_{Q2}/V_{QB2}. If $V_{Q2\text{-}PH3}$ and $V_{QB2\text{-}PH3}$ are both below the trip-point (V_{TRIP}) of the inverter (INV) in SADEC, then SA1 = SA2 = 1 and the Flag-latch (FL) = 0 ("inside" case). If $V_{Q2\text{-}PH3}>V_{TRIP}$ and $V_{QB2\text{-}PH3}<V_{TRIP}$, then SA1≠SA2 (FL = 1, "outside" case). In phase-4 (PH4), after resetting ($V_{Q2} = V_{QB2}$ = 0V), SW5 = SW6 = on, and CD = 0 to make $V_{Q2} = V_Q$ and $V_{QB2} = V_{QB}$. In phase-5 (PH5, MSB detection), SAEN = 1, S1 = 0, and N2-N4 = on to detect $\Delta V_{Q2\text{-}QB2}$ and generate the 2nd output (SA1/SA2). SAOUT[1] = Q (MSB value), when SAOUT[0] is the XNOR of FL value generated in PH3 and SAOUT[1] of PH5. After repeating the DbSO-CSA operation 3 times using 3 pairs of I_{REF} ($I_{REF\_L[2:0]}$ and $I_{REF\_H[2:0]}$), the DbSO-CSA outputs 6b of unsigned data. Finally, the digital-combiner (DC) accumulates all 7b MACV from the 16 IOs (1b-sign, 6b-data) and outputs 11b MACV (1b-sign, 10b-data).

Figure 15.4.5 shows the performance of the proposed scheme. The proposed DbSO-CSA reduced access time by 1.3× at 6b CSA output, before the DC. Under a 4bIN-4bW configuration, the proposed schemes (in combination) outperformed conventional scheme (with WL = input and conventional CSA) in terms of EF_{MAC} (2.4-3.9× improvement across various IN-W-OUT precisions).

Figure 15.4.6 shows the measurement results from a fabricated 22nm 2Mb ReRAM nvCIM macro with typical DFF-based path-delay excluding scheme. The captured waveform confirms that the macro achieved t_{AC} = 9.8ns for 1bIN-2bW-6bOUT. The shmoo test for 4bIN-4bW-11bOUT confirmed that t_{AC} = 18.5ns at V_{DD} = 0.8V. The measured EF_{MAC} was 121.3TOPS/W at 1bIN-2bW-6bOUT and 28.9TOPS/W at 4bIN-4bW-11bOUT. Compared to a previous work (2bIN-3bW-4bOUT) [3], this work achieved a 2× improvement in EF_{MAC} and a 1.11× shorter in t_{AC} with higher precision (2bIN-4bW-10bOUT). The system-level measurement results using our 2Mb ReRAM-CIM testchip for the ResNet-20 model. Using the CIFAR-100 dataset, this work achieved results that were only 0.93% degraded from those obtained using the pure-software approach (4bIN-4bW-11bOUT). Fig. 15.4.7 represents a die photo.

Acknowledgements:
The authors would like to thank MOST-Taiwan, TSRI, NTHU-TSMC JDP for financial and manufacturing support.

References:

[1] W.-H. Chen et al., "A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processors," *ISSCC*, pp. 494-495, Feb. 2018.
[2] R. Mochida et al., "A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture," *VLSI*, pp. 175-176, 2018.
[3] C.-X. Xue et al., "A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN Based AI Edge Processors," *ISSCC*, pp. 388-389, Feb. 2019.
[4] X. Si et al., "A Twin-8T SRAM Computation-In-Memory Macro for Multiple-bits CNN-Based Machine Learning," *ISSCC*, pp. 396-398, Feb. 2019.

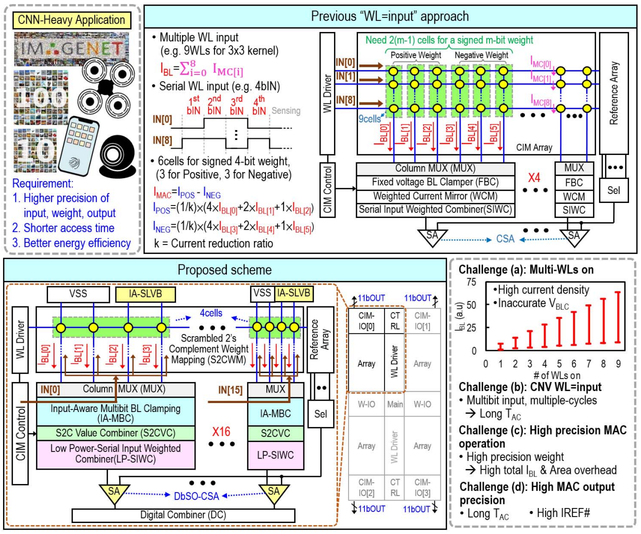

CNN-Heavy Application
Requirement:
1. Higher precision of input, weight, output
2. Shorter access time
3. Better energy efficiency
Previous "WL=input" approach
• Multiple WL input (e.g. 9WLs for 3x3 kernel)
• Serial WL input (e.g. 4bIN)
• 6cells for signed 4-bit weight, (3 for Positive, 3 for Negative)
k = Current reduction ratio
Need 2(m-1) cells for a signed m-bit weight
CIM Array
Proposed scheme
Input-Aware Multibit BL Clamping (IA-MBC)
S2C Value Combiner (S2CVC)
Low Power-Serial Input Weighted Combiner(LP-SIWC)
DbSO-CSA
Digital Combiner (DC)
Challenge (a): Multi-WLs on
•High current density
•Inaccurate V_BLC
Challenge (b): CNV WL=input
• Multibit input, multiple-cycles
→ Long T_AC
Challenge (c): High precision MAC operation
• High precision weight
→ High total I_BL & Area overhead
Challenge (d): High MAC output precision
• Long T_AC • High IREF#

Figure 15.4.1: Challenges in improving nvCIM precision and comparison between conventional approches and the proposed scheme.

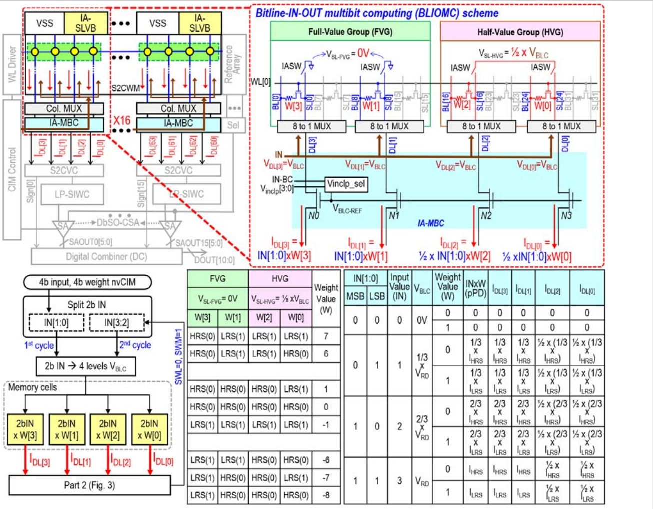


Figure 15.4.2: Proposed multibit MAC operations in cell array using the BL-IN-OUT multibit computing (BLIOMC) and the scrambled 2's complement weight mapping (S2CWM) schemes.

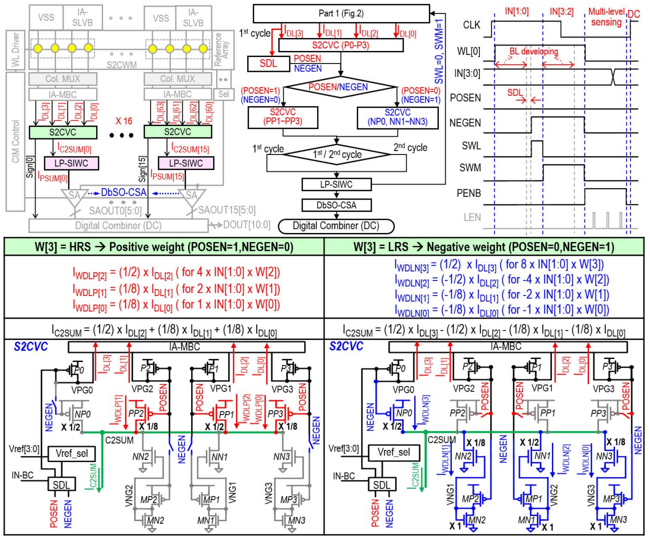


Figure 15.4.3: Operations of the proposed scrambled 2's complement value combiner (S2CVC).

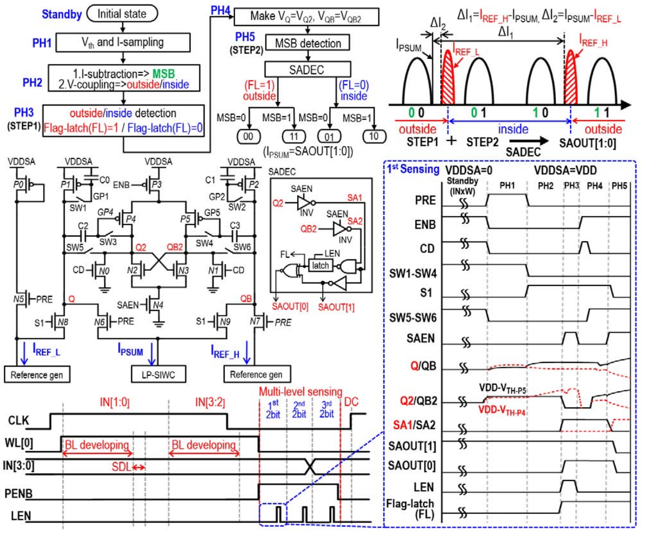

Standby
Initial state
PH1 V_th and I-sampling
PH2 1.I-subtraction=> MSB 2.V-coupling=>outside/inside
PH3 (STEP1) outside/inside detection Flag-latch(FL)=1 / Flag-latch(FL)=0
PH4 Make V_Q=V_Q2, V_QB=V_QB2
PH5 (STEP2) MSB detection
SADEC
STEP1 + STEP2 SADEC SAOUT[1:0]
Reference gen
LP-SIWC
Multi-level sensing

Figure 15.4.4: Structure and operations of dual-bit small-offset current-mode sense amplifier (DbSO-CSA).

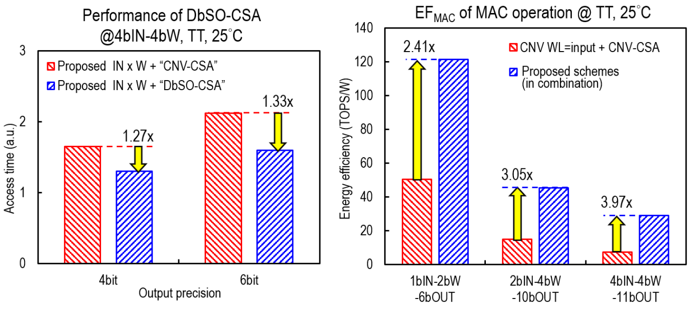


Figure 15.4.5: Simulated performance of the proposed DbSO-CSA (left) and energy efficiency comparison between a conventional scheme and this work (right).

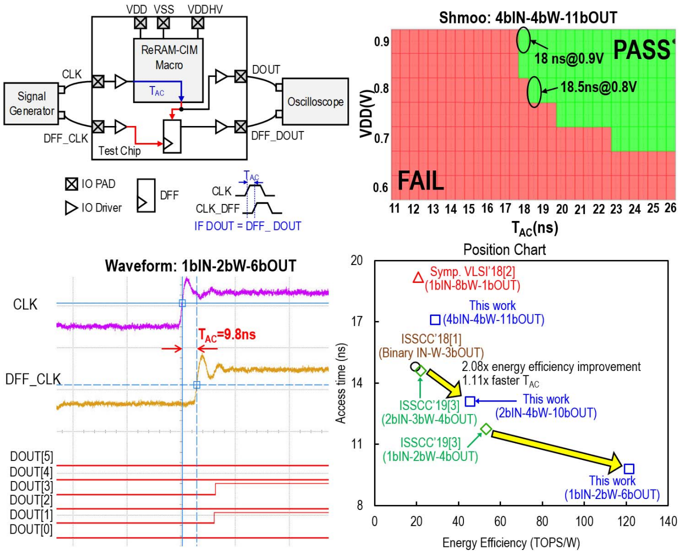
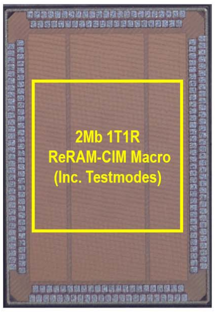


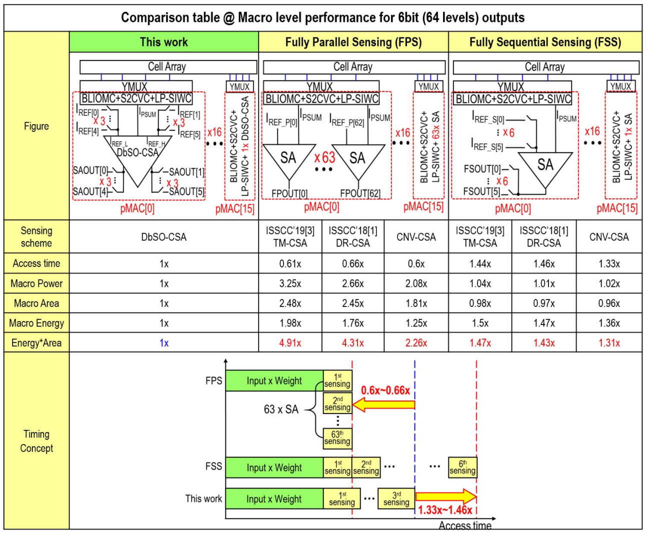
Figure 15.4.6: Measurement results.

15



Technology	22nm CMOS Logic Process	
ReRAM	Foundry 1T1R SLC ReRAM	
Testchip Size (Inc. IO pad and testmodes)	2mm x 3mm	
Macro Mode	(1) Memory (2) CIM (MAC)	
Capacity	2Mb (8 Sub-bank)	
Sub-bank	512rows x 512columns	
Read Delay, T_{AC} @ VDD =0.8V	CIM Mode: 9.8 ns (1bIN-2bW-6bOUT)	
	CIM Mode: 13.1 ns (2bIN-4bW-10bOUT)	
	CIM Mode: 18.3 ns (4bIN-4bW-11bOUT)	
Energy Efficiency, EF_{MAC} @ VDD =0.8V	CIM Mode: 121.38 (TOPS/W) (1bIN-2bW-6bOUT)	
	CIM Mode: 45.52 (TOPS/W) (2bIN-4bW-10bOUT)	
	CIM Mode: 28.93 (TOPS/W) (4bIN-4bW-11bOUT)	

Figure 15.4.7: Die photo and chip summary.


Comparison table @ Macro level performance for 6bit (64 levels) outputs

	This work	Fully Parallel Sensing (FPS)			Fully Sequential Sensing (FSS)		
Sensing scheme	DbSO-CSA	ISSCC'19[3] TM-CSA	ISSCC'18[1] DR-CSA	CNV-CSA	ISSCC'19[3] TM-CSA	ISSCC'18[1] DR-CSA	CNV-CSA
Access time	1x	0.61x	0.66x	0.6x	1.44x	1.46x	1.33x
Macro Power	1x	3.25x	2.66x	2.08x	1.04x	1.01x	1.02x
Macro Area	1x	2.48x	2.45x	1.81x	0.98x	0.97x	0.96x
Macro Energy	1x	1.98x	1.76x	1.25x	1.5x	1.47x	1.36x
Energy*Area	1x	4.91x	4.31x	2.26x	1.47x	1.43x	1.31x

Figure 15.4.S1: Comparison of FPS, FSS and the DbSO-CSA for 6bOUT. The DbSO-CSA achieved 1.31x to 4.91x improvement in Energy × Area.

	This work			ISSCC'19 [3]		Symp. VLSI'18 [2]	ISSCC'18 [1]
Technology	22nm	22nm	22nm	55nm	55nm	180nm	65nm
Synapses	2Mb	2Mb	2Mb	1Mb	1Mb	2Mb	1Mb
Input precision	1b	2b	4b	1b	2b	1b	1b
Weight precision	2b	4b	4b	3b	3b	8b	Ternary
Input direction	BL clamping voltage	BL clamping voltage	BL clamping voltage	WL	WL	WL	WL
Weight (+/-)	Scrambled 2's complement	Scrambled 2's complement	Scrambled 2's complement	Different columns	Different columns	Different columns	Different blocks
Accuracy(MNIST)	N/A	N/A	N/A	98%	98.8%	90.8%	98%
Accuracy(CIFAR-10)	N/A	90.18%	N/A	81.83%	88.52%	N/A	N/A
Accuracy(CIFAR-100)	N/A	64.15%	66.46%	N/A	N/A	N/A	N/A
Sensing Scheme	DbSO-CSA	DbSO-CSA	DbSO-CSA	TM-CSA	TM-CSA	Current Comparator	DR-CSA
Accumulation #	16	16	16	9	9	196	9
MAC Output precision	6b	10b	11b	4b	4b	1b	3b
Full precision (ideal case)	6b	10b	12b	7b	9b	16b	5b
T_{AC} (ns)	9.8	13.1	18.3	11.75	14.6	19.18	14.8
Energy Efficiency, EF_{MAC} (TOPS/W)	121.38	45.52	28.93	53.17	21.9	20.7	19.2
*FoM (Normalize)	242.76 (23.46)	364.16 (35.18)	424.31 (41.04)	91.15 (8.81)	58.4 (5.64)	10.35 (1.00)	23.04 (2.23)

*FoM = Energy efficiency x input precision x weight precision x (output precision / full precision)

Figure 15.4.S2: Comparison the performance of recent silicon-proven ReRAM nvCIM works. This work outperformed all previous works in terms of FoM (EF_{MAC} × input precision × weight precision × (output precision / full precision), thanks to its superior t_{AC}, EF_{MAC}, and precision.

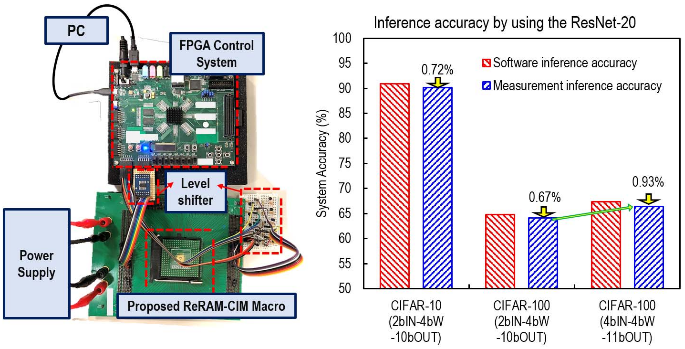


Figure 15.4.S3: The experimental FPGA-based platform and system-level measurement results using our 2Mb ReRAM-CIM testchip for the ResNet-20 model. Under 2bIN-4bW-10bOUT MAC operations, the proposed scheme achieved 90.18% inference accuracy when applied to the CIFAR-10 dataset, which is superior to the results in [3] (88.5% at 2bIN-3bW-4bOUT). Using the CIFAR-100 dataset, this work achieved results that were only 0.93% degraded from those obtained using the pure-software approach (4bIN-4bW-11bOUT MAC operation).