Enabling Energy-Efficient and High-Density eDRAM-LUT-Based Computing-in-Memory Using Anti-Ferroelectric Transistors

Hongtao Zhong[1†], Zijie Zheng[2†], Zhonghao Chen[1†], Taixin Li[1], Zuopu Zhou[2], Leming Jiao[2], Xiaoyang Ma[3],
Hongyang Jia[1], Huazhong Yang[1], Chen Jiang[1], Thomas Kämpfe[4], Kai Ni[5*], Xiao Gong[2*], Xueqing Li[1*]
[1]Department of Electronic Engineering, LFET/BNRist, Tsinghua University; [2]Electrical and Computer Engineering, National University of Singapore, Singapore; [3]Princeton University; [4]Fraunhofer IPMS, Germany; [5]University of Notre Dame, USA
[†]Equal Contributions; [*]Corresponding Email: xueqingli@tsinghua.edu.cn, elegong@nus.edu.sg, kni@nd.edu

***Abstract*—This paper reports a novel energy-efficient and high-memory/compute-density eDRAM Computing-in-Memory (CiM) by introducing AFeFET-based eDRAM cells into look-up tables (LUTs). The highlights include: (1) Fabricated 1-transistor-1-AFeFET (1T1AF) eDRAM cell with short Amorphous-Indium-Gallium-Zinc-Oxide (*a*-IGZO) channel length of 25 nm and high cell density; (2) Measured high endurance over 2×10^9 cycles for AFeFETs and long retention time over 10^3 s for 1T1AF eDRAM cell; (3) Proposed computation mode combining digital LUT and analog capacitor coupling for the first time, enabling high energy efficiency, high compute density, and high accuracy. Evaluation results show that the proposed 1T1AF eDRAM-LUT-based CiM achieves a high memory density of 5.52 Mb/mm^2, a high 8b peak compute density of 11.3 TOPS/mm^2, and a high 8b peak energy efficiency of 107.2 TOPS/W, showing great potential for energy-efficient and high-performance CiM designs.**

***Index Terms*—eDRAM, Computing-in-Memory, Look-up Table, Anti-ferroelectric FET**

I. Introduction

Computing-in-Memory (CiM) can integrate computation within the memory, showcasing strong capability in accelerating data-intensive applications, especially for deep neural networks (DNNs). With the exponential growth of DNN model size, high-density CiM designs are becoming more necessary with more parameters stored in memory and further reduced data movement. Among various CiM designs, eDRAM CiM [1]–[6] is a promising approach towards high-density CiM with fewer transistors than SRAM CiM.

However, most existing CMOS-based eDRAM CiM designs [1]–[3] fail to achieve high memory density due to the large capacitors in each cell for longer retention time. As a competitive alternative, IGZO-based eDRAM CiM [5], [6] utilizes the low-leakage IGZO channel to enhance retention time. However, existing IGZO-based eDRAM CiM designs still suffer from low memory density due to their long channel length (L_{CH}) and the large capacitor in each cell for charge-domain computing [5] or multi-bit storage [6]. Moreover, large capacitors also heavily degrade compute density.

Unlike the above eDRAM-based analog CiM, [4] proposes an eDRAM-LUT-based digital CiM based on capacitor-less 3T eDRAM with high memory density. Besides, by storing the pre-computed accumulation results in the eDRAM LUTs, high compute density is achieved. However, its short retention time involves frequent refresh operations, which induces performance degradation and limits energy efficiency. As shown in Fig. 1(a), there is a trade-off between energy efficiency and memory/compute density for existing eDRAM CiM designs.

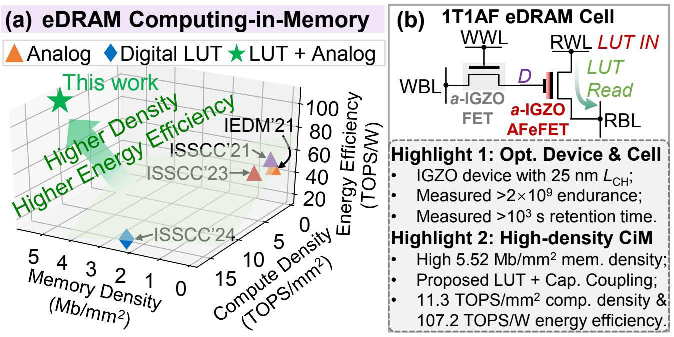

Fig. 1. (a) This work breaks the impossible trinity of memory density, compute density and energy efficiency in prior eDRAM CiM designs; (b) Schematic of an 1T1AF eDRAM cell and highlights of this work.

To break this impossible trinity, this work proposes a novel eDRAM-LUT-based CiM based on the anti-ferroelectric FETs (AFeFETs) with the 1-transistor-1-AFeFET (1T1AF) cell design and the BEOL-compatible *a*-IGZO channel, as presented in Fig. 1(b) with the highlights of this work. AFeFETs are volatile devices with much higher endurance than their FeFET counterparts [7], [8]. Besides, the unique pinched hysteresis of AFeFET can be used to eliminate subthreshold leakage, which enables ultra-long retention time with the IGZO channel, further improving performance and energy efficiency. 1T1AF eDRAM cells with down to 25 nm L_{CH} were also successfully fabricated. Moreover, this work proposes a computation mode, called *LUT-CC*, that combines digital LUTs and analog capacitor coupling in charge-domain computing with high compute density, high energy efficiency and variation resilience.

The contributions of this work are itemized as follows:

- Fabricated capacitor-less 1T1AF eDRAM with down to 25 nm IGZO channel length and further high cell density;
- **Measured** $>2\times10^9$ high endurance and $>10^3$ s long retention time for fabricated 1T1AF eDRAM cells;
- Proposed *LUT-CC* enabling 5.52 Mb/mm^2 high memory density, 11.3 TOPS/mm^2 8b peak compute density, and 107.2 TOPS/W 8b peak energy efficiency.

II. Experimental Results of 1T1AF eDRAM Cell

A. Device Fabrication and Characterization

Fig. 2(a) shows the cross-sectional view and the schematic of the fabricated *a*-IGZO AFeFET with the key process

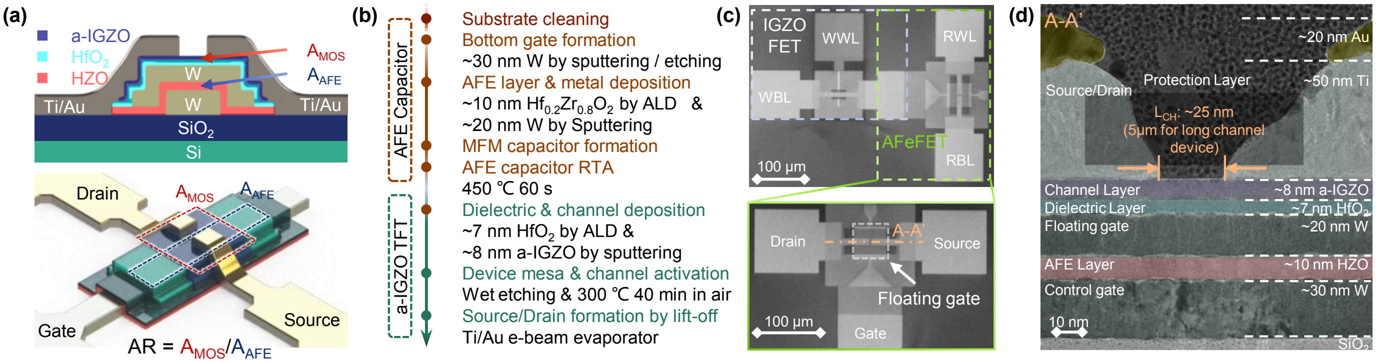

Fig. 2. (a) Device structure and (b) Key process steps of the *a*-IGZO AFeFETs; (c) Top-view SEM image of the *a*-IGZO AFeFET and 1T1AF eDRAM cell; (d) HRTEM image in the channel region (A-A') with the thickness of each layer, highlighting the fabrication precision and 25 nm L_{CH}.

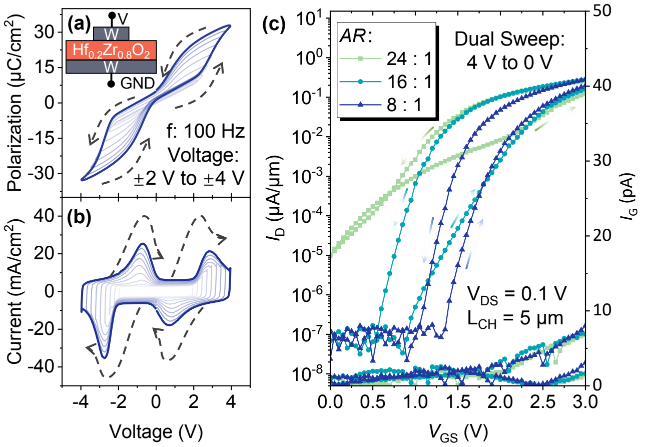

Fig. 3. (a) *P-V* and (b) *I-V* loops of the AFE capacitor with different RTA temperatures; (c) I_D-V_{GS} curves of the *a*-IGZO AFeFETs with various ARs.

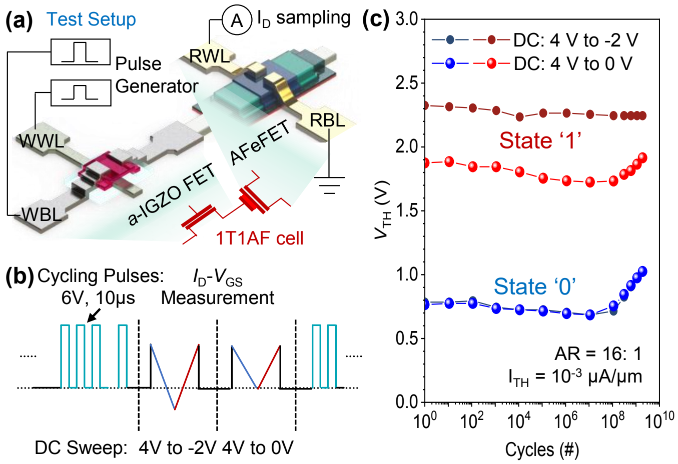

Fig. 4. (a) Test setup and (b) Waveform with cycling pulses of the endurance measurement for 1T1AF eDRAM cell; (c) **Measured** endurance of the AFeFETs with AR of 16:1, highlighting a high endurance over 2×10^9 cycles.

steps in Fig. 2(b). The AFeFETs use the MFMIS structure with the flexibility to engineer the area ratio (AR) of the MOSFET area (A_{MOS}) to the AFE area (A_{AFE}) to tune the overall device characteristics. Meanwhile, the rapid thermal annealing (RTA) with 450 °C 60 s can provide strong AFE behavior, indicating the BEOL compatibility of HZO for AFE applications. Fig. 2(c) shows the SEM image of the fabricated AFeFET device and the 1T1AF eDRAM cell, and Fig. 2(d) shows the HRTEM image cut from the channel region (A-A') that confirms L_{CH}, thickness and quality of each layer, highlighting the precision of the fabrication and 25 nm L_{CH}.

Fig. 3(a) and (b) illustrate the *P-V* and *I-V* characteristics of the AFE capacitors with a MIM structure to show the anti-ferroelectricity, respectively. Note that, the negligible remnant polarization (P_r) at 0 V in the *P-V* curves implies the AFE volatility. Fig. 3(c) shows the I_D-V_{GS} curves of the *a*-IGZO AFeFETs with various ARs under dual sweep measurement from 4 V to 0 V, exhibiting volatile anti-clockwise hysteresis of AFeFETs. Besides, It is seen that the fabricated AFeFETs with proper ARs enjoy a large memory window (MW) and a high ON/OFF ratio (e.g., ~1 V MW and ~700 ON/OFF ratio with an AR of 16:1), enabling reliable memory applications.

B. Measurements of the 1T1AF eDRAM Cell

Fig. 4(a) shows the test setup for the measurement of 1T1AF eDRAM cell. Fig. 4(b) illustrates the waveform applied to the AFeFET gate with the unipolar cycling pulses of 6 V and 10 μs, followed by DC sweep ranging from 4 V to -2 V and from 4 V to 0 V to obtain the I_D-V_{GS} curves. -2 V is used for more saturated switching of the AFE layer and further larger MW. Fig. 4(c) presents the measured endurance characteristic of the AFeFETs with an AR of 16:1 through the extracted V_{TH}. Our fabricated devices maintain proper function after >2×10^9 cycles without dielectric breakdown or significant MW degradation. This proves that using the AFE material with unipolar stress for the devices can significantly improve the endurance compared with its FE counterpart [9].

The unique feature of AFeFET is the pinched hysteresis behavior with the intrinsic volatile feature. Prior works [8] usually engineer the built-in field into the AFE layer to obtain non-volatile AFeFETs. However, this method suffers from a much more complex process and a smaller MW & ON/OFF ratio [7]. To overcome this challenge, we propose a novel data maintaining operation, called almost-leakage-free (ALF) operation, that can achieve the AFeFETs with the almost non-volatile feature. As shown in Fig. 5(a), given an external voltage V_m within the hysteresis window of AFeFET I_D-V_{GS}, ALF applies V_m to the drain of the turned-off IGZO FET in the 1T1AF eDRAM cell. It is seen that both states '0' and '1' can be maintained using the same voltage source with ultra-low hardware cost. Besides, thanks to the voltage balance between the drain and source of IGZO FET, retention time can be significantly extended with eliminated subthreshold leakage.

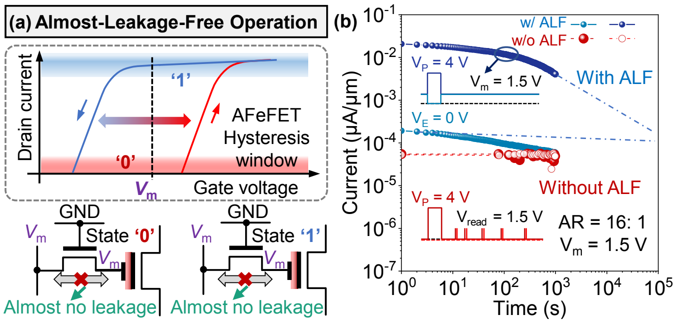

Fig. 5. (a) Proposed Almost-Leakage-Free (ALF) operation to achieve much longer retention time; (b) **Measured** retention time of the fabricated 1T1AF eDRAM cell, highlighting an ultra-long retention time over 10^3 s with ALF.

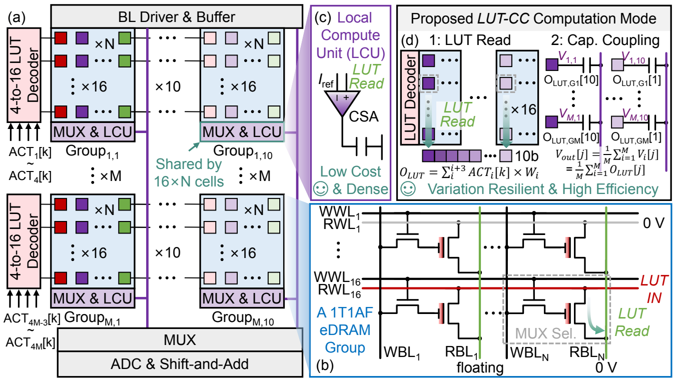

Fig. 6. (a) Overall architecture of the proposed eDRAM-LUT-based 1T1AF CiM with $M\times 10$ groups; (b) Schematic of a 1T1AF eDRAM group with $16\times N$ cells and voltage configuration to read out the selected cell through the LUT input determined by activations (ACTs); (c) Schematic of a local compute unit (LCU) shared by a group with low cost and high density; (d) Detailed operations of the proposed *LUT-CC* computation mode with high energy efficiency, high compute density and variation resilience.

(a) Mapping Table of eDRAM LUT for MAC

ACT	LUT-Read Data	ACT	LUT-Read Data
0000	0	1000	W_4
0001	W_1	1001	$W_4 + W_1$
0010	W_2	1010	$W_4 + W_2$
0011	$W_2 + W_1$	1011	$W_4 + W_2 + W_1$
0100	W_3	1100	$W_4 + W_3$
0101	$W_3 + W_1$	1101	$W_4 + W_3 + W_1$
0110	$W_3 + W_2$	1110	$W_4 + W_3 + W_2$
0111	$W_3 + W_2 + W_1$	1111	$W_4 + W_3 + W_2 + W_1$

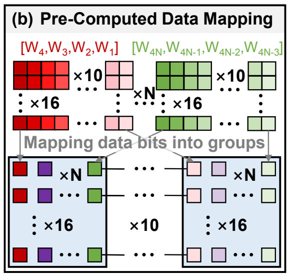

Fig. 7. (a) Mapping Table of the 1T1AF eDRAM LUT that stores the pre-computed weight combinations and performs MAC by the ACT input; (b) Mapping method of pre-computed data into different groups in the same row.

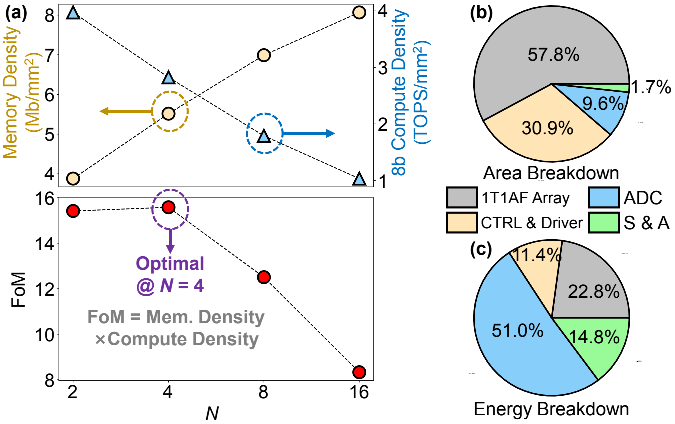

Fig. 8. (a) Group size optimization. N=4 is selected as an optimal group size with the maximum FoM; (b) Area and (c) Energy breakdown of the proposed eDRAM-LUT-based 1T1AF CiM with optimized M=128 and N=4.

The test setup for probing the retention characteristics of 1T1AF eDRAM cell is shown before in Fig. 4(a). Fig. 5(b) depicts the retention characteristics with test waveforms. A 4 V/10 ms program pulse is initially applied to ensure the saturated write of AFeFET. Then, for the proposed ALF operation, a V_m of 1.5 V is applied to the drain of the turned-off IGZO FET to maintain the memory state. It is evident that, without V_m, the AFeFET loses its memory state, as the merged red curves suggested. Instead, with V_m, the memory states can be held in AFeFETs, and two distinct states can be observed with a decent ON/OFF ratio and a **measured** long retention time over 10^3 s. The extrapolated results (blue dash line) can even reach a retention time of $\sim 10^5$ s, which is fundamental for energy-efficient eDRAM CiM designs.

III. Proposed eDRAM-LUT-based 1T1AF CiM

A. Proposed LUT-CC Computation Mode

Fig. 6(a) presents the architecture of the proposed eDRAM-LUT-based 1T1AF CiM with $M\times 10$ groups. The 10 groups in the same row share one 4-to-16 LUT decoder whose output is determined by 4 aligned bits from 4 activations (ACTs). The M groups in the same column share the BL drivers & buffers and one ADC & Shift-and-Add (S & A). Fig. 6(b) shows the schematic of a 1T1AF eDRAM group with $16\times N$ cells, and Fig. 6(c) shows the schematic of a local compute unit (LCU) module with one current sense amplifier (CSA) that converts the dynamic current by LUT read into a static voltage and a MOM capacitor that performs analog accumulation by capacitor coupling (CC). Note that an LCU is shared by a group so the area overhead is greatly amortized, enabling high density with the compact capacitor-less 1T1AF eDRAM.

Fig. 6(d) illustrates the detailed operations of the proposed *LUT-CC* computation mode with two steps. Before the multiply-accumulate (MAC) operation, following the mapping rules in Fig. 7(a), N 16×10b weight combinations from 4 8b weights are pre-computed and interleaved into 10 $16\times N$ groups, as shown in Fig. 7(b). In the first step, for the 10 groups in each row, a 10b LUT result is read-out according to the 4b ACT input. Fig. 6(b) depicts the voltage configuration for LUT read. Note that this step is essentially memory read with variation resilience to the AFeFET variation. In the second step, M obtained LUT results from M rows are accumulated by CC. The proposed *LUT-CC* has the advantages of both digital LUTs and analog computing, enabling high energy efficiency, high density, and variation resilience.

B. Parameter Optimization

Selecting the specific value of M and N involves a big design space. In this work, 5b ADCs are utilized and M is set to 128 as a suitable design considering the typical 50% activation and weight sparsity. Fig. 8(a) shows the memory density and 8b compute density of the proposed eDRAM-LUT-based 1T1AF CiM in different N values, and it is seen that larger N can increase memory density but decrease compute density. In this work, N=4 is selected as a sweet spot with the

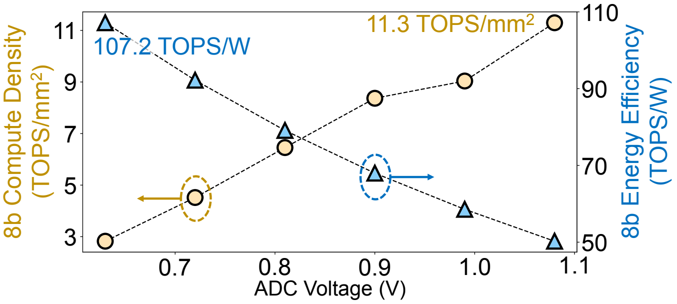

Fig. 9. 8b compute density and energy efficiency in different ADC voltages.

maximum figure-of-merit (FoM, defined by memory density × compute density), and corresponding memory density and 8b compute density are 5.52 Mb/mm^2 and 2.82 TOPS/mm^2, respectively. Fig. 8(b) and (c) illustrate the area and energy breakdown of the proposed eDRAM-LUT-based 1T1AF CiM with 128×10 groups and the 16×4 group size, respectively. The ADC voltage (V_{ADC}) here is 0.63 V, enabling the 8b peak energy efficiency of 107.2 TOPS/W.

Fig. 9 shows the 8b compute density and energy efficiency of the proposed eDRAM-LUT-based 1T1AF CiM in different V_{ADC}. It is seen that higher V_{ADC} increases compute density but decreases energy efficiency. At 0.63 V V_{ADC}, the 8b peak energy efficiency of 107.2 TOPS/W is achieved as aforementioned. At 1.08 V V_{ADC}, the 1T1AF CiM macro achieves the 8b peak compute density of 11.3 TOPS/mm^2 with maintained-high 8b energy efficiency of 50.3 TOPS/W.

C. System-Level Benchmarking

Table I presents the performance comparison between the proposed eDRAM-LUT-based 1T1AF CiM and the state-of-the-art (SOTA) eDRAM CiM designs. With the compact capacitor-less 1T1AF cell and locally shared LCU, this work achieves the record-high memory density (for storage) of 5.52 Mb/mm^2, which is 5.46× and 2.40× higher than the SOTA capacitor-based eDRAM CiM [3] and eDRAM-LUT-based CiM [4], respectively. Besides, this work also achieves the record-high 8b peak energy efficiency of 107.2 TOPS/W (V_{ADC}=0.63 V), which is at least 1.96× higher than existing eDRAM CiM designs. Such a big improvement results from (1) Proposed ALF operation with long retention time over 10^3 s and further ultra-low refresh overhead; (2) Proposed *LUT-CC* computation mode with the pre-computed partial-sum results stored in LUT and energy-efficient analog capacitor coupling. This work can also achieve a very high 8b peak compute density of 11.3 TOPS/mm^2 (V_{ADC}=1.08 V), which is much higher than existing capacitor-based eDRAM CiM designs.

Benchmarking with ResNet-20/18 on CIFAR-10/100, with the measured 2% device variation of AFeFET, this work can achieve 91.63%/71.84% accuracy with only 0.39%/0.36% loss. Such variation resilience is thanks to the calibration capability of LUT and the reliable output of capacitor coupling.

TABLE I
Performance comparison with the SOTA eDRAM CiM designs

	ISSCC'21 [1]	ISSCC'23 [2]	VLSI'23 [3]	ISSCC'24 [4]	IEDM'21 [5]	**This Work**
Technology	65 nm Si	28 nm Si	28 nm Si	28 nm Si	45 nm IGZO	**25 nm IGZO**
Cell Structure	3T1C eDRAM	3T2C eDRAM	3T2C eDRAM	3T eDRAM	4T1C eDRAM	**1T1AF eDRAM**
Computation Mode	Analog Current Summation	Analog Capacitor Coupling	Analog Capacitor Coupling	Digital LUT	Analog Capacitor Coupling	**Digital LUT + Capacitor Coupling**
Capacitor-Less?	No	No	No	**Yes**	No	**Yes**
Weight Storage	**4b/cell**	1b/cell	1b/cell	1b/cell	1b/cell	**1b/cell**
Memory Density (Mb/mm^2)	0.15	0.95	1.01	2.30 [a] 0.46 [b]	0.32 [c]	**5.52 [a]** **1.10 [b]**
8b Peak Compute Density (TOPS/mm^2) [d]	2.20	0.63	0.81	**16.2**	0.58	**11.3**
8b Peak Energy Efficiency (TOPS/W) [d]	54.8	33.9	45.7	19.7	43.1	**107.2**
Retention Time (s)	3.6×10^{-4}	4×10^{-5}	4×10^{-5}	1.3×10^{-6} [e]	13	**$>10^3$**
Cifar-10 Accuracy (%)	91.2 [f]	89.5 [g]	-	**92.02 [g]**	89.0 [h]	**91.63 [g]**
Cifar-100 Accuracy (%)	-	-	70.7 [f]	-	-	**71.84 [f]**

[a] For storage; [b] For computation; [c] C_{par}= 2 fF, C_C = 10 fF, cell area estimated by Si layout; [d] 1 OP is 1 multiplication or 1 addition, normalized to 8b input & weight; [e] @25 °C; [f] ResNet-18; [g] ResNet-20; [h] VGG-8.

IV. Conclusion

This paper presents a novel energy-efficient and high-memory/compute-density eDRAM-LUT-based 1T1AF CiM with a series of device-circuit co-optimizations including: (1) Fabricated highly-scaled AFeFET devices and 1T1AF cells with down to 25 nm L_{CH} and **measured** high endurance over 10^9 cycles; (2) Proposed ALF operation with **measured** long retention time over 10^3 s; (3) Proposed *LUT-CC* computation mode enabling record-high memory density of 5.52 Mb/mm^2, record-high 8b peak energy efficiency of 107.2 TOPS/W and high 8b peak compute density of 11.3 TOPS/mm^2. These remarkable results evidence the great potential of the proposed eDRAM-LUT-based 1T1AF CiM for energy-efficient and high-density neural network acceleration.

Acknowledgment

This work is supported in part by NSFC (92264204, U24B6-015, U21B2030) and in part by Singapore Ministry of Education MOE Tier 1 (A-8001168-00-00 and A-8002027-00-00).

References

[1] Z. Chen *et al.*, "15.3 A 65nm 3T Dynamic Analog RAM-Based Computing-in-Memory Macro and CNN Accelerator with Retention Enhancement, Adaptive Analog Sparsity and 44TOPS/W System Energy Efficiency," in *ISSCC*, vol. 64, 2021, pp. 240–242.

[2] S. Kim *et al.*, "16.5 DynaPlasia: An eDRAM In-Memory-Computing-Based Reconfigurable Spatial Accelerator with Triple-Mode Cell for Dynamic Resource Switching," in *ISSCC*, 2023, pp. 256–258.

[3] S. Kim *et al.*, "Scaling-CIM: An eDRAM-based In-Memory-Computing Accelerator with Dynamic-Scaling ADC for SQNR-Boosting and Layer-wise Adaptive Bit-Truncation," in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2023, pp. 1–2.

[4] Y. He *et al.*, "34.7 A 28nm 2.4 Mb/mm^2 6.9-16.3 TOPS/mm^2 eDRAM-LUT-Based Digital-Computing-in-Memory Macro with In-Memory Encoding and Refreshing," in *ISSCC*, vol. 67, 2024, pp. 578–580.

[5] J. Liu *et al.*, "Low-Power and Scalable Retention-Enhanced IGZO TFT eDRAM-Based Charge-Domain Computing," in *2021 IEEE International Electron Devices Meeting (IEDM)*, 2021, pp. 21.1.1–21.1.4.

[6] M.-C. Chen *et al.*, "A > 64 Multiple States and > 210 TOPS/W High Efficient Computing by Monolithic Si/CAAC-IGZO+ Super-Lattice $ZrO_2/Al_2O_3/ZrO_2$ for Ultra-Low Power Edge AI Application," in *2022 International Electron Devices Meeting (IEDM)*, 2022, p. 18.2.1–18.2.4.

[7] Z. Liang *et al.*, "A Novel High-Endurance FeFET Memory Device Based on ZrO_2 Anti-Ferroelectric and IGZO Channel," in *2021 IEEE International Electron Devices Meeting (IEDM)*, 2021, pp. 17.3.1–17.3.4.

[8] Z. Zheng *et al.*, "First Demonstration of Work Function-Engineered BEOL-Compatible IGZO Non-Volatile MFMIS AFeFETs and Their Co-Integration with Volatile-AFeFETs," in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2023, pp. 1–2.

[9] C. Sun *et al.*, "First Demonstration of BEOL-Compatible Ferroelectric TCAM Featuring a-IGZO Fe-TFTs with Large Memory Window of 2.9 V, Scaled Channel Length of 40 nm, and High Endurance of 10^8 cycles," in *2021 Symposium on VLSI Technology*, 2021, pp. 1–2.