

Energy Efficient Dual Designs of FeFET-Based Analog In-Memory Computing with Inherent Shift-Add Capability

Zeyu Yang¹, Qingrong Huang¹, Yu Qian¹, Kai Ni², Thomas Kämpfe³ and Xunzhao Yin^{1,4*}

¹Zhejiang University, China ²University of Notre Dame, USA ³Fraunhofer IPMS, Germany

⁴Key Laboratory of CS&AUS of Zhejiang Province, China *Corresponding author email: xzyin1@zju.edu.cn

ABSTRACT

In-memory computing (IMC) architecture emerges as a promising paradigm, improving the energy efficiency of multiply-and-accumulate (MAC) operations within deep neural networks (DNNs) by integrating the parallel computations within the memory arrays. Various high-precision analog IMC array designs have been developed based on both SRAM and emerging non-volatile memories (NVMs). These designs perform MAC operations of partial input and weight, with the corresponding partial products then fed into shift-add circuitry to produce the final MAC results. However, existing works often present intricate shift-add process for weight. The traditional digital shift-add process is limited in throughput due to time-multiplexing of ADCs, and advancing the shift-add process to the analog domain necessitates customized circuit implementations, resulting in compromises in energy and area efficiency. Furthermore, the joint optimization of the partial MAC operations and the weight shift-add process is rarely explored. In this paper, we propose novel, energy efficient dual designs of ferroelectric FET (FeFET) based high precision analog IMC featuring inherent shift-add capability. We introduce a FeFET based IMC paradigm that performs partial MAC in each column, and inherently integrates the shift-add process for 4-bit weights by leveraging FeFET's analog storage characteristics. This paradigm supports both 2's complement mode (2CM) and non-2's complement mode (N2CM) MAC, thereby offering flexible support for 4-/8-bit weight data in 2's complement format. Building upon this paradigm, we propose novel FeFET based dual designs, CurFe for the current mode and ChgFe for the charge mode, to accommodate the high precision analog domain IMC architecture. Evaluation results at circuit and system levels indicate that the circuit/system-level energy efficiency of the proposed FeFET-based analog IMC is $1.56\times/1.37\times$ higher when compared to the state-of-the-art analog IMC designs.

1 INTRODUCTION

Over the past decade, deep neural networks (DNNs) have become crucial in artificial intelligence (AI), particularly in domains such as image recognition, speech recognition, and dynamic monitoring [1]. However, with technological advancements and exponentially growing data volumes, the computational and storage demands of DNNs have risen sharply. As a result, the significant data movement between memory and processing units has led to the "memory wall" bottleneck in the conventional Von Neumann architecture. In-memory computing (IMC) architecture, which integrates computational functions within memory, aims to address this by mitigating the extensive data movement. Its natural parallel processing

capabilities are well-suited for efficient execution of multiply-and-accumulate (MAC) operations, which are essential to DNNs [2–4].

In the analog IMC field, many SRAM-based designs have been proposed [5–10]. Nevertheless, the inherent binary storage nature of SRAM complicates multi-bit IMC circuitry design, and the large SRAM cell size and standby power limit area and energy efficiency. In contrast, emerging non-volatile memories (NVMs) like resistive RAM (ReRAM), magnetic RAM (MRAM), and ferroelectric FET (FeFET), are gaining attention due to their compact structure and near-zero standby power [11–23]. FeFET, in particular, is promising for building efficient IMC design to store weights and perform MAC operations in DNN inference due to multi-level cell (MLC), high ON/OFF ratio, and three-terminal read/write separation [24–29].

In high-precision IMC designs, multi-bit MAC dataflow typically represents each fixed-point n -bit weight value with n cells in adjacent columns. It thereby combines n outputs of weighted partial sums based on different significant weight bits using peripheral circuits, a process called the shift-add process for weight. Currently, there are two shift-add process categories [30]. The first, "digital shift-add", directs the partial MAC value (pMACV) of the selected column through a Multiplexer (MUX) to an analog-to-digital converter (ADC), which converts pMACV to digital form in a single cycle. Over n sequential cycles, these n digital values are summed on bit significance via a digital shift-add circuit. This process is slow due to time-multiplexing, and consumes additional area overhead for components like Multipliers [31]. The alternative process, "analog shift-add", relocates the shift-add process for weight before analog-to-digital conversion. This allows parallel generation of pMACV results for each column, followed by their weighted summation in the analog domain. Several analog shift-add circuit modules have been developed to support various analog domain based IMC implementations [6, 7, 9]. However, these solutions still involve significant energy and area overheads. Furthermore, both of the above shift-add circuit modules are separate from the array implementing the partial MAC operations, indicating a substantial opportunity for joint optimization of these two processes.

In this paper, we propose novel dual designs for FeFET-based high precision analog IMC with inherent shift-add capability. Unlike previous approaches, we leverage FeFET's analog storage property to establish a novel FeFET-based IMC array paradigm that not only has partial MAC ability for each column, but also inherently integrates the shift-add process for 4-bit weights. This effectively eliminates the need for extra dedicated shift-add circuits in multi-bit weight processing. This FeFET based IMC paradigm supports both 2's/non-2's complement mode (2CM/N2CM) MAC for 4-/8-bit weight data in 2's complement format. Building upon this paradigm, we propose two designs for prevalent analog domain IMC architectures: the current mode (CurFe) and the charge mode (ChgFe) architectures. The CurFe design uses 1nFeFET1R cells with binary-weighted conductance values, and a trans-impedance amplifier (TIA) for both the MAC and shift-add processes. ChgFe utilizes single-level cell (SLC) 1pFeFET and MLC 1nFeFET to store the sign bit and remaining bits, respectively, enabling inherent shift-add operations through

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DAC '24, June 23–27, 2024, San Francisco, CA, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0601-1/24/06...\$15.00

<https://doi.org/10.1145/3649329.3655990>

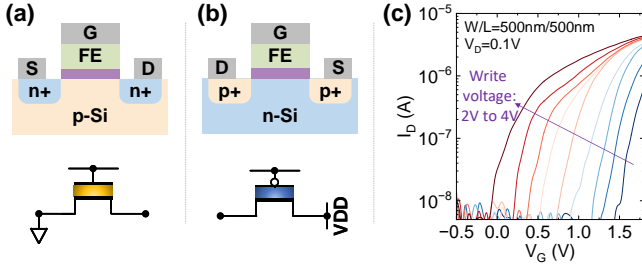


Figure 1: (a)/(b) Structure of nFeFET/pFeFET. (c) Measured I_D - V_G characteristics with MLC V_{th} states of a fabricated nFeFET.

charge sharing among capacitors. SPICE and system-level simulation results suggest that our proposed FeFET-based analog IMC is $1.56\times/1.37\times$ more energy efficient at the circuit/system-level compared to the state-of-the-art analog IMC designs.

The rest is organized as follows: Section 2 reviews the basics and relevant prior works. Section 3 introduces our proposed IMC dual designs, i.e., CurFe and ChgFe. Section 4 presents evaluation results. Section 5 concludes the paper.

2 BACKGROUND

Here we review the FeFET device, analog IMC and related works.

2.1 FeFET Basics

FeFET, a three-terminal NVM device, is widely used in IMC designs due to its CMOS compatibility, high ON/OFF ratio, and compact structure. Most research focuses on nFeFET, similar to a standard nMOS transistor but featuring a thick doped HfO₂ ferroelectric (FE) layer on its gate, as shown in Fig. 1(a). Measurements on fabricated nFeFET devices as shown in Fig. 1(c) have indicated that nFeFETs can be programmed to exhibit MLC threshold voltage V_{TH} states by applying different write pulses at gate. This capability enables non-volatile storage of information within an 1nFeFET cell. In parallel, the pFeFET device, based on pMOS transistor technology (Fig. 1(b)), has been experimentally demonstrated to exhibit similar switching behavior to nFeFET [32]. Furthermore, the 1pFeFET cell has been validated as a programmable synapse [33], demonstrating its usefulness in IMC applications.

2.2 Analog IMC preliminaries

Numerous SRAM-based analog IMC designs have been proposed [5–10]. On the one hand, these designs explore a variety of cell structures, including 8T-SRAM [6, 7, 9], 10T-SRAM [5], and Block-wise 6T SRAM [8, 10], among others. In terms of implementation methods in analog domain, there are primarily two modes: current mode and charge mode. Current mode aggregates currents from multiple computing units to obtain pMACV [6, 8, 9], while charge mode represents multiplication results as charges [5, 7, 10]. Additionally, analog IMC architectures based on various NVMs have emerged. ReRAM-based designs are studied for 8-bit precision inference implementation [14–16]. FeFET has recently gained attention as a potential candidate for analog IMC, thanks to its three-terminal structure, high ON/OFF ratio, and compact structure. However, most existing FeFET-based IMC designs are limited to storing only binary states [17, 19], not fully utilizing the MLC states of FeFET. Recently, Soliman et al. [25] proposed an IMC design using a MLC FeFET for 2-bit multiplication operations in an 1nFeFET1R cell. Nevertheless, current limitations in the precision that can be stored in a single FeFET make extending this design to high-precision IMC, such as 8-bit, a notable challenge.

2.3 Related Works and Motivation

Storing an n -bit weight in a single IMC cell is typically challenging, requiring the use of multiple cells in adjacent columns. As a result, this necessitates combining the pMACVs from various columns based on their weight bit significance through subsequent peripheral circuitry, a process known as the shift-add process for weight. The conventional "digital shift-add" process is executed after ADC conversion with a digital circuit module. However, due to the considerable ADC overhead, multiple columns have to share an ADC through time-multiplexing, facilitated by a MUX, thus demanding multiple clock cycles to complete a MAC operation. To optimize this process, several works adopt the "analog shift-add" approach [6, 7, 9], where the shift-add operation occurs before the ADC conversion in the analog domain. This method allows the ADC to directly generate the MAC result with embedded weight significance, leading to a substantial throughput improvement. Si et al. [6] proposed a 2's complement weight mapping scheme with a processing unit to achieve "analog shift-add" for 5-bit weights, it requires additional proportional capacitors, resulting in extra area overhead. Dong et al. [7] employed binary-weighted computation capacitors in 4-bit Flash ADCs for charge sharing and column-wise pMACVs combining for 4-bit precision, but the scalability to higher weight bit precision is challenging due to the significant capacitance difference between the least significant bit (LSB) capacitor and the most significant bit (MSB) capacitor. Yue et al. [9] addressed the above issues and designed an "analog shift-add" based ADC that flexibly supports both 2CM/N2CM for signed/unsigned 4-bit weight, thus allowing for 8-bit signed weight MAC operations using 2CM for the high 4-bit and N2CM for the low 4-bit. For example, an m -bit unsigned input X (where $m=2$ in [9]) and an 8-bit signed weight Y in 2's complement format. The multiplication is segmented into two parts corresponding to the 2CM/N2CM ADCs:

$$X = \sum_{i=0}^{m-1} x_i 2^i, Y = (-y_7 2^7 + \sum_{j=4}^6 y_j 2^j) + (\sum_{j=0}^3 y_j 2^j) \quad (1)$$

$$XY = (-y_7 \sum_{i=0}^{m-1} x_i 2^{i+7} + \sum_{i=0}^{m-1} \sum_{j=4}^6 x_i y_j 2^{i+j}) + (\sum_{i=0}^{m-1} \sum_{j=0}^3 x_i y_j 2^{i+j}) \quad (2)$$

When combining the accumulation module, such MAC operation support can be extended to 2-/4-/6-/8-bit inputs and 4-/8-bit weights. However, like the approach in [6], extra binary-weighted capacitors are also required. Besides, the separation of both types of the shift-add process for weight from the multi-bit MAC operation suggests potential for further integration of these processes. In this paper, we propose energy efficient dual designs for FeFET-based analog IMC with inherent shift-add capability, aiming to integrate the shift-add process for weight with the partial MAC operation in the IMC array through FeFET's analog storage characteristics.

3 PROPOSED ANALOG IMC DUAL DESIGNS

Here we introduce the dual designs of analog IMC paradigm that utilizes the adjustable analog characteristics of FeFET cells for storing multi-bit weights. Our current mode design uses 1nFeFET1R cells with varying resistance to conduct weighted currents corresponding to weighted bits. Meanwhile, our charge mode design conducts weighted currents by programming different V_{TH} to 1nFeFET cells.

3.1 Current Mode FeFET-Based IMC

This section introduces the design of the current mode FeFET-based IMC, referred to as CurFe. Fig. 2(a) illustrates the overall architecture, comprising a wordline (input) driver, a BL/SL switch

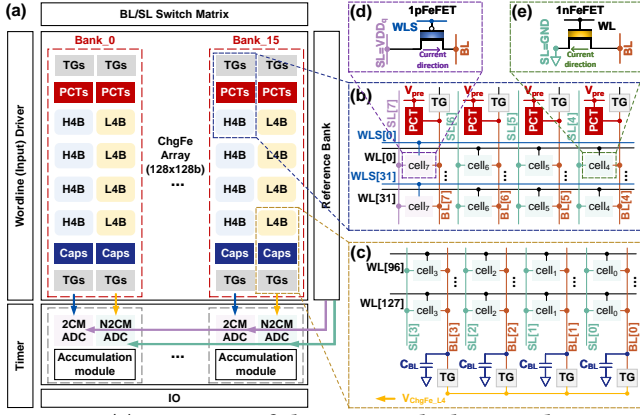


Figure 4: (a) Structure of the proposed ChgFe architecture. (b) Structure of H4B with PCTs and TGs. (c) Structure of L4B with capacitors and TGs. (d) 1pFeFET structure for $cell_7$. (e) 1nFeFET structure for $cell_0$ - $cell_6$.

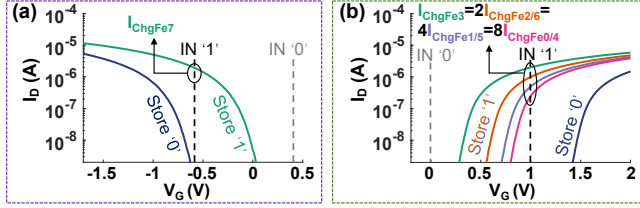


Figure 5: (a) I_D - V_g curves of $cell_7$ in ChgFe. (b) I_D - V_g curves of $cell_0$ - $cell_6$ in ChgFe.

with a pre-charged transistor (PCT) and a capacitor, replacing the TIA used in CurFe. Notably, the V_{TH} state positioned on the right side in Fig. 5(a) is designated as the high V_{TH} state of pFeFET, indicating that the sign bit for the weight has a value of '1'. To signify weight significance, as shown in Fig. 5(b), the low V_{TH} states (representing '1') of the 1nFeFET cells for distinct significant weight bits vary, and the ON current magnitude of the high V_{TH} state of the 1pFeFET in $cell_7$ matches that of $cell_3$, creating a binary-weighted pattern for the ON state current of $cell_0$ - $cell_3$ (denoted as $I_{ChgFe0}-I_{ChgFe3}$)/ $cell_4$ - $cell_7$ (denoted as $I_{ChgFe4}-I_{ChgFe7}$).

The proposed ChgFe performs 32 accumulations for a 1-bit input and an 8-bit weight in the charge domain. Within each bank, SL[7] is set to V_{DDq} , while other SLs are grounded, and all TGs are OFF. Initially, each BL capacitor is pre-charged to V_{pre} (1.5V) within 1ns via PCT. 32 1-bit input data are then applied to the corresponding WL and WLS via wordline drivers within 0.5ns. In this stage, the MAC operation for 1-bit input and 1-bit weight is executed on each BL. Specifically, the capacitor on BL[7] is charged through activated $cell_7$ s, while the capacitors on BL[0]-BL[6] are discharged through the respective activated $cell_0$ s- $cell_6$ s. Since all FeFETs operate in the saturation region, the ON state currents vary less with time, causing the voltage changes in BL voltages (denoted as $\Delta V_{ChgFe0}-\Delta V_{ChgFe3}$ for $cell_0$ - $cell_3$ and $\Delta V_{ChgFe4}-\Delta V_{ChgFe7}$ for $cell_4$ - $cell_7$, respectively) following a binary-weighted pattern, with ΔV_{ChgFe7} being positive and others negative. Subsequently, controlled by multiple TGs, charge sharing operations between BL capacitors in H4B/L4B are performed to produce the output voltages $V_{ChgFe-H4}/V_{ChgFe-L4}$:

$$V_{ChgFe-H4} = V_{pre} + \left(\sum \Delta V_{ChgFe7} + \sum \Delta V_{ChgFe6} + \sum \Delta V_{ChgFe5} + \sum \Delta V_{ChgFe4} \right) / 4 \quad (5)$$

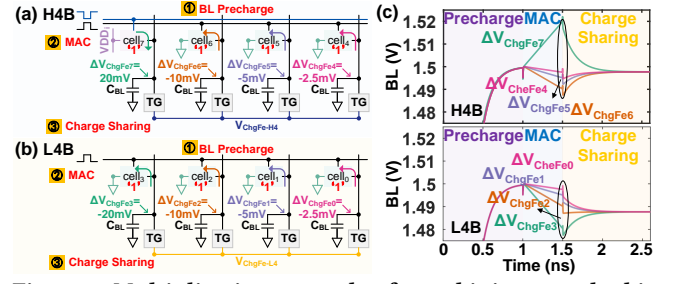


Figure 6: Multiplication example of an 1-bit input and 8-bit signed weight in ChgFe. The 8-bit weight is divided into (a) high 4-bit and (b) low 4-bit parts in H4B and L4B, respectively. (c) Transient simulation waveforms of this operation.

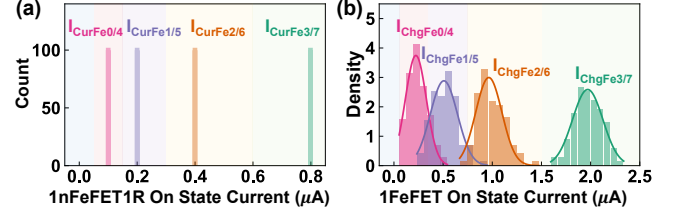


Figure 7: (a) Current histogram of $I_{CurFe0}-I_{CurFe7}$ in CurFe. (b) Current histogram of $I_{ChgFe0}-I_{ChgFe7}$ in ChgFe.

$$V_{ChgFe-L4} = V_{pre} + \left(\sum \Delta V_{ChgFe3} + \sum \Delta V_{ChgFe2} + \sum \Delta V_{ChgFe1} + \sum \Delta V_{ChgFe0} \right) / 4 \quad (6)$$

Hence, both the 1-bit partial MAC operation and the shift-add process for 4-bit weight in 2CM/N2CM depend on the same BL capacitors, eliminating the need for extra binary-weighted computation capacitors as seen in [7]. Fig. 6(a) and (b) illustrate a multiplication scenario of an 1-bit input '1' and an 8-bit weight "11111111" in ChgFe, with no other rows in this H4B/L4B activated. As shown in Fig. 6(c), each BL voltage experiences a slight drop at the beginning of the charge-sharing operation, but this does not affect linearity.

Similar to CurFe, ChgFe can accommodate 1-8 bit unsigned inputs and 4-/8-bit signed weights in 2's complement format, aided by ADCs, accumulation modules, and external control. The details of this operation are similar to those in CurFe and are not repeated.

4 VALIDATION AND EVALUATION

In this section, we validate and evaluate our proposed dual designs of FeFET-based analog IMC at both circuit level and system level, which are then compared with state-of-the-art analog IMC designs.

4.1 Circuit-Level Validation and Evaluation

We conducted SPICE simulations on our proposed dual designs of FeFET-based analog IMC using the Cadence Spectre Simulator. The simulations are based on the experimentally calibrated Preisach FeFET model [34] and a commercial CMOS 40nm processing development kit. The write method described in [35] was adopted. The BL capacitor in ChgFe was set to 50fF. FeFET devices are assumed to have V_{TH} variability with $\sigma=40mV$ for each state, as per [25].

We first performed a Monte Carlo simulation to assess the impact of FeFET variation on different ON state currents. In the CurFe architecture, as shown in Fig. 7(a), the drain resistance of the 1nFeFET1R cell significantly mitigates fluctuations in the ON state current. For ChgFe, to maintain adequate linearity in the charge domain, we employed the 1nFeFET and 1pFeFET cell structure. The simulation results are presented in Fig. 7(b). We then exhaustively examined

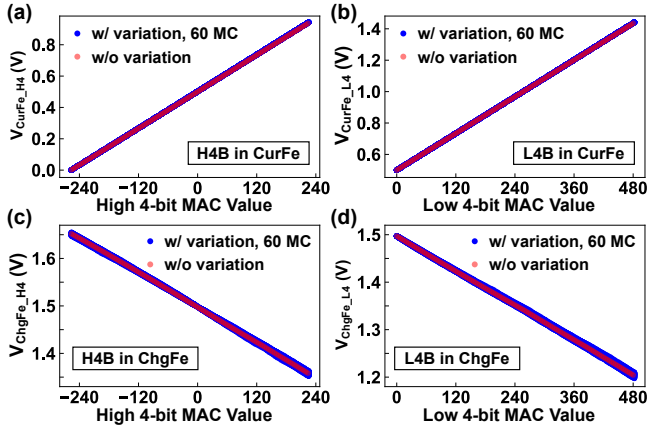


Figure 8: The MAC outputs for 32 accumulations of 1-bit input and 4-bit weight in (a)/(c) H4B and (b)/(d) L4B under CurFe/ChgFe.

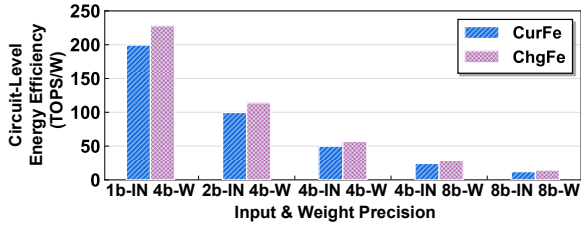


Figure 9: Average energy efficiency for 32 accumulations with different input and weight precision in CurFe/ChgFe.

all possible input and weight combinations to generate complete MAC outputs for 32 accumulations of 1-bit input and 4-bit weight in H4B and L4B under both CurFe and ChgFe. As illustrated in Fig. 8, the results exhibit good linearity for both CurFe and ChgFe. By conducting 60 Monte Carlo simulations for each case, we observe the impact of FeFET variation on the output voltages, consistent with the effect on the currents in Figure 7. Furthermore, we accounted for different output fluctuations in CurFe and ChgFe to assess the DNN inference accuracy in Section 4.2.

In terms of energy efficiency, as depicted in Fig. 9, we evaluated the average circuit-level energy efficiency for 32 accumulations involving different input and weight precision in both CurFe and ChgFe. It is essential to highlight that x b-IN/ y b-W represents the x -bit input/ y -bit weight precision in Fig. 9 and subsequent figures. This evaluation was conducted with a 5-bit ADC precision setting. A detailed analysis of ADC precision will be presented in Section 4.2. As expected, energy efficiency decreases with input/weight precision. Notably, the energy efficiency in CurFe is lower than that in ChgFe at the same precision level. This difference is attributed to the higher energy requirement for the TIA in CurFe compared to the energy needed for precharging in ChgFe.

4.2 Benchmark Results and Discussion

We further benchmark the system performance of our proposed analog IMC dual designs using NeuroSim V1.4 [36], an integrated framework for DNN inference on IMC-based hardware accelerator with the support for various device technologies. We selected two representative networks, i.e., VGG8 and ResNet18, and two datasets, CIFAR10 and ImageNet, for our analysis. We assume an H-tree structure for routing among modules in each hierarchy. The sub-array size is set to 128x128, and the partial parallel model for 32 input

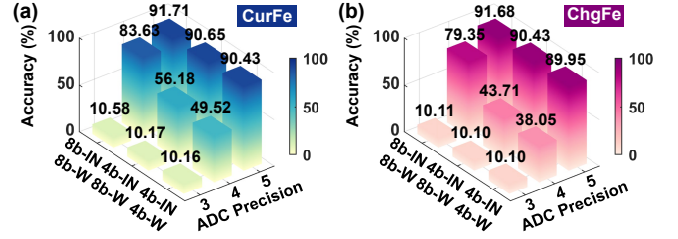


Figure 10: Impact of ADC resolution and input/weight precision on accuracy for CIFAR10 dataset in (a) CurFe and (b) ChgFe architectures.

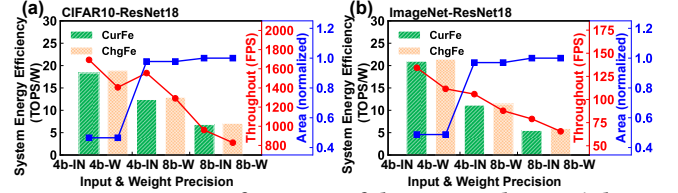


Figure 11: System performance of the proposed CurFe/ChgFe architectures with different input and weight precision on ResNet18 for (a) CIFAR10 and (b) ImageNet datasets.

parallelism is enabled as illustrated in Fig. 2 and 4. Moreover, the ON/OFF ratio of FeFET is set to 10^5 [17]. Modifications have been made to NeuroSim to accommodate our proposed architectures.

In evaluating DNN inference accuracy, we consider the impact of ADC resolution and input/weight precision, as well as device variations under the CurFe and ChgFe architectures, as analyzed in Section 4.1. Using the VGG8 network on the CIFAR10 dataset as an example, with a baseline accuracy of 92%, the experimental results in Fig. 10 show that 5-bit ADC is necessary to avoid significant accuracy loss. This finding aligns with the analysis reported in [36]. Consequently, the ADC precision is set to 5-bit in the subsequent performance analysis. The accuracy under ChgFe is slightly lower than that under CurFe, consistent with Fig. 8, but this discrepancy is acceptable. Even at 4-bit input/weight precision, accuracy with 5-bit ADC under ChgFe is less than 0.5% compared to CurFe.

Next, we evaluate the performance of the ResNet18 network on CIFAR10 and ImageNet. The system-level performance of our proposed dual designs, considering different input/weight precision, is shown in Fig. 11. Consistent with circuit-level results, ChgFe exhibits higher system energy efficiency than CurFe at the same precision level. However, the throughput in ChgFe is lower than that in CurFe due to longer time required for MAC operation in ChgFe. The system area costs are similar for both architectures. Furthermore, for the case of 4-bit input/weight precision, Fig. 12 provides a detailed breakdown of dynamic energy consumption and latency for each layer of ImageNet-ResNet18 in CurFe and ChgFe.

4.3 Comparison With State-of-the-Art Designs

Table 1 compares our proposed designs with the state-of-the-art analog IMC designs that enable 8-bit precision DNN inference. For a relatively fair comparison of energy efficiency, all designs are scaled to 40nm with MAC operations of 8-bit input and 8-bit weights by multiplying λ^2 , where λ is the ratio of the realistic technology node to 40nm. It should be noted that [9] includes additional sparse optimizations. Our analysis shows that without considering sparse optimization, our FeFET-based analog IMC designs achieves the highest energy efficiency at the circuit level for 8-bit input and weight precision, which is 1.56 \times and 2.22 \times higher than the latest

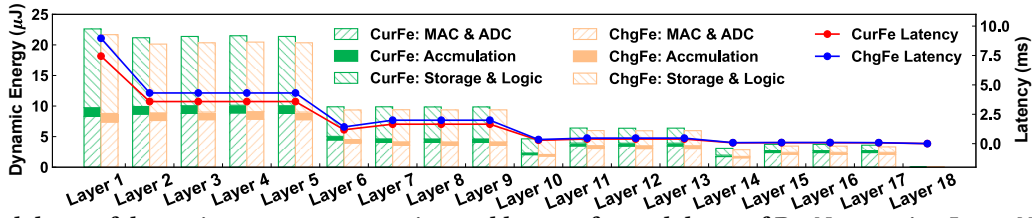


Figure 12: Breakdown of dynamic energy consumption and latency for each layer of ResNet18 using ImageNet dataset with 4-bit input and 4-bit weight precision in CurFe/ChgFe architectures.

Table 1: Comparison with the state-of-the-art analog IMC designs

Reference	[8]	[9]	[10]	[14]	[15]	[16]	CurFe	ChgFe
Technology	CMOS	CMOS	CMOS	ReRAM	ReRAM	ReRAM	FeFET	FeFET
Cell Type	6T-SRAM+LLC	8T-SRAM	6T-SRAM+LMC	1T1R	1T1R	1T1R	1nFeFET1R	1nFeFET/1pFeFET
Node	28nm	65nm	28nm	22nm	22nm	22nm	40nm	40nm
Input Precision (bit)	4/8	2/4/6/8	4/8	1/4/8	1/2/4/8	1-8	1-8	1-8
Weight Precision (bit)	4/8	4/8	4/8	2/4/8	2/4/8	1-8	4/8	4/8
Computing Mode	current domain	current domain	charge domain	current domain	current domain	charge domain	current domain	charge domain
Multi-Bit Weight Processing	digital shift-add	analog shift-add	digital shift-add	digital shift-add	digital shift-add	digital shift-add	inherent shift-add	inherent shift-add
Average Circuit/Macro-Level Energy Efficiency (TOPS/W) [†]	6.90@(8b,8b)*	41.67@(4b,8b) (with sparse optimization)	9.26@(8b,8b)	3.60@(8b,8b)	4.72@(8b,8b)	6.53@(8b,8b)	12.18@(8b,8b)	14.47@(8b,8b)
Average System-Level Energy Efficiency (TOPS/W) ^{†‡}	N/A	9.40@(4b,8b)	N/A	N/A	N/A	N/A	12.41@(4b,8b)	12.92@(4b,8b)

[†]: Scaled to 40nm, assume energy \propto (Node)². [‡]: Based on CIFAR10-ResNet18. *: α :(x,b,yb): x-bit input, y-bit weight.

SRAM [10] and ReRAM [16] based analog IMC designs, respectively. At the system level, using the same CIFAR10-ResNet18 configuration, the system energy efficiency of our FeFET-based analog IMC as derived from the Neurosim framework, is 1.37 \times higher than the latest analog IMC based system realization [9]. This enhanced efficiency is primarily due to the inherent shift-add capability in our CurFe/ChgFe designs, which eliminates the need for additional hardware for the shift-add operation in multi-bit weight processing. While the energy efficiency of the CurFe architecture is slightly lower than that of the ChgFe architecture, CurFe exhibits better robustness against device variations, as detailed in Fig. 10.

5 CONCLUSION

In this paper, we present novel dual designs for FeFET-based high precision analog IMC with inherent shift-add capabilities. These designs capitalize on the analog storage characteristics of FeFETs to establish a unique FeFET-based IMC array paradigm. This approach not only offers partial MAC capabilities for each column, but also inherently integrates the shift-add process for 4-bit weights directly within the array. As a result, we eliminate the need for additional digital or analog shift-add circuits typically required for multi-bit weight processing. Our proposed FeFET based IMC paradigm is designed to support both 2CM/N2CM MAC, thus providing flexible support for 4-/8-bit weight data in 2's complement format. We then develop the CurFe and ChgFe designs to accommodate the analog domain IMC architecture in both the current mode and charge mode, respectively. Evaluation results at circuit and system levels indicate that our novel dual designs offer superior energy efficiency compared to existing state-of-the-art analog IMC approaches.

ACKNOWLEDGMENTS

This work was supported in part by NSFC (92164203, 62104213) and SGC Cooperation Project (Grant No. M-0612).

REFERENCES

- [1] Y. LeCun et al., "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [2] Z. Yan et al., "Computing-in-memory neural network accelerators for safety-critical systems: Can small device variations be disastrous?," in *IEEE/ACM ICCAD*, pp. 1–9, 2022.
- [3] N. Verma et al., "In-memory computing: Advances and prospects," *IEEE Solid-State Circuits Magazine*, vol. 11, no. 3, pp. 43–55, 2019.
- [4] Z. Yan et al., "Swim: Selective write-verify for computing-in-memory neural accelerators," in *ACM/IEEE DAC*, pp. 277–282, 2022.
- [5] A. Biswas et al., "Conv-ram: An energy-efficient sram with embedded convolution computation for low-power cnn-based machine learning applications," in *2018 IEEE ISSCC*, pp. 488–490, 2018.
- [6] X. Si et al., "24.5 a twin-8t sram computation-in-memory macro for multiple-bit cnn-based machine learning," in *2019 IEEE ISSCC*, pp. 396–398, 2019.
- [7] Q. Dong et al., "15.3 a 351tops/w and 372.4 gops compute-in-memory sram macro in 7nm finfet cmos for machine-learning applications," in *2020 IEEE ISSCC*, pp. 242–244, 2020.
- [8] X. Si et al., "15.5 a 28nm 64kb 6t sram computing-in-memory macro with 8b mac operation for ai edge chips," in *2020 IEEE ISSCC*, pp. 246–248, 2020.
- [9] J. Yue et al., "14.3 a 65nm computing-in-memory-based cnn processor with 2.9-to-35.8 tops/w system energy efficiency using dynamic-sparsity performance-scaling architecture and energy-efficient inter/intra-macro data reuse," in *2020 IEEE ISSCC*, pp. 234–236, 2020.
- [10] J.-W. Su et al., "16.3 a 28nm 384kb 6t-sram computation-in-memory macro with 8b precision for ai edge chips," in *2021 IEEE ISSCC*, pp. 64, pp. 250–252, 2021.
- [11] C. Zhuo et al., "Design of ultra-compact content addressable memory exploiting 1t1mtj cell," *IEEE TCAD*, 2022.
- [12] Q. Huang et al., "Fefet based in-memory hyperdimensional encoding design," *IEEE TCAD*, 2023.
- [13] A. Sebastian et al., "Memory devices and applications for in-memory computing," *Nature Nanotechnology*, vol. 15, no. 7, pp. 529–544, 2020.
- [14] C.-X. Xue et al., "16.1 a 22nm 4mb 8b-precision rram computing-in-memory macro with 11.91 to 195.7 tops/w for tiny ai edge devices," in *2021 IEEE ISSCC*, pp. 245–247, 2021.
- [15] J.-M. Hung et al., "A four-megabit compute-in-memory macro with eight-bit precision based on cmos and resistive random-access memory for ai edge devices," *Nature Electronics*, vol. 4, no. 12, pp. 921–930, 2021.
- [16] J.-M. Hung et al., "8-b precision 8-mb rram compute-in-memory macro using direct-current-free time-domain readout scheme for ai edge devices," *IEEE JSSC*, vol. 58, no. 1, pp. 303–315, 2022.
- [17] T. Soliman et al., "Ultra-low power flexible precision fefet based analog in-memory computing," in *2020 IEEE IEDM*, pp. 29–2, 2020.
- [18] X. Yin et al., "Ferroelectric ternary content addressable memories for energy-efficient associative search," *IEEE TCAD*, vol. 42, no. 4, pp. 1099–1112, 2022.
- [19] D. Saito et al., "Analog in-memory computing in fefet-based 1t1r array for edge ai applications," in *2021 IEEE Symposium on VLSI Technology*, pp. 1–2, 2021.
- [20] Q. Huang et al., "Computing-in-memory using ferroelectrics: From single-to multi-input logic," *IEEE Design & Test*, vol. 39, no. 2, pp. 56–64, 2021.
- [21] X. Yin et al., "Ferroelectric compute-in-memory annealer for combinatorial optimization problems," *Nature Communications*, vol. 15, no. 1, p. 2419, 2024.
- [22] L. Liu et al., "A reconfigurable fefet content addressable memory for multi-state hamming distance," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2023.
- [23] H. Xu et al., "On the challenges and design mitigations of single transistor ferroelectric content addressable memory," *IEEE Electron Device Letters*, 2023.
- [24] C. Li et al., "A scalable design of multi-bit ferroelectric content addressable memory for data-centric computing," in *2020 IEEE IEDM*, pp. 29–3, 2020.
- [25] T. Soliman et al., "First demonstration of in-memory computing crossbar using multi-level cell fefet," *Nature Electronics*, vol. 14, no. 1, p. 6348, 2023.
- [26] X. S. Hu et al., "In-memory computing with associative memories: A cross-layer perspective," in *2021 IEEE IEDM*, pp. 25–2, 2021.
- [27] X. Yin et al., "Fecam: A universal compact digital and analog content addressable memory using ferroelectric," *IEEE Transactions on Electron Devices*, vol. 67, no. 7, pp. 2785–2792, 2020.
- [28] S. Shou et al., "See-mcam: Scalable multi-bit fefet content addressable memories for energy efficient associative search," in *2023 IEEE/ACM ICCAD*, pp. 1–9, 2023.
- [29] X. Yin et al., "An ultracompact single-ferroelectric field-effect transistor binary and multibit associative search engine," *Advanced Intelligent Systems*, vol. 5, no. 7, p. 2200428, 2023.
- [30] H. Jiang et al., "Analog-to-digital converter design exploration for compute-in-memory accelerators," *IEEE Design & Test*, vol. 39, no. 2, pp. 48–55, 2021.
- [31] J. Hur et al., "Nonvolatile capacitive crossbar array for in-memory computing," *Advanced Intelligent Systems*, vol. 4, no. 8, p. 2100258, 2022.
- [32] D. Kleimaier et al., "Demonstration of a p-type ferroelectric fet with immediate read-after-write capability," *IEEE Electron Device Letters*, vol. 42, no. 12, pp. 1774–1777, 2021.
- [33] S. Thomann et al., "All-in-memory brain-inspired computing using fefet synapses," *Frontiers in Electronics*, vol. 3, p. 833260, 2022.
- [34] K. Ni et al., "A circuit compatible accurate compact model for ferroelectric-fets," in *2018 IEEE symposium on VLSI technology*, pp. 131–132, 2018.
- [35] D. Reis et al., "Design and analysis of an ultra-dense, low-leakage, and fast fefet-based random access memory array," *IEEE Jx CDC*, vol. 5, no. 2, pp. 103–112, 2019.
- [36] X. Peng et al., "Dnn+ neurosim: An end-to-end benchmarking framework for compute-in-memory accelerators with versatile device technologies," in *2019 IEEE IEDM*, pp. 32–5, 2019.