# Low Power and Temperature-Resilient Compute-In-Memory Based on Subthreshold-FeFET

Yifei Zhou[1], Xuchu Huang[1], Jianyi Yang[1,2,*], Kai Ni[3], Hussam Amrouch[4], Cheng Zhuo[1,5,*], and Xunzhao Yin[1,5,*]

[1]Zhejiang University, Hangzhou, China; [2]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou, China;
[3]Department of Electrical Engineering, University of Notre Dame, Notre Dame, USA
[4]Chair of AI Processor Design, Technical University of Munich; TUM School of Computation, Information and Technology; Munich Institute of Robotics and Machine Intelligence, Munich, Germany
[5]Key Laboratory of Collaborative Sensing and Autonomous Unmanned Systems of Zhejiang Province, Hangzhou, China
*Corresponding authors, email: {czhuo, yangjy, xzyin1}@zju.edu.cn

***Abstract*—Compute-in-memory (CiM) is a promising solution for addressing the challenges of artificial intelligence (AI) and the Internet of Things (IoT) hardware such as "memory wall" issue. Specifically, CiM employing nonvolatile memory (NVM) devices in a crossbar structure can efficiently accelerate multiply-accumulation (MAC) computation, a crucial operator in neural networks among various AI models. Low power CiM designs are thus highly desired for further energy efficiency optimization on AI models. Ferroelectric FET (FeFET), an emerging device, is attractive for building ultra-low power CiM array due to CMOS compatibility, high $I_{ON}/I_{OFF}$ ratio, etc. Recent studies have explored FeFET based CiM designs that achieve low power consumption. Nevertheless, subthreshold-operated FeFETs, where the operating voltages are scaled down to the subthreshold region to reduce array power consumption, are particularly vulnerable to temperature drift, leading to accuracy degradation. To address this challenge, we propose a temperature-resilient 2T-1FeFET CiM design that performs MAC operations reliably at subthreahold region from 0°C to 85°C, while consuming ultra-low power. Benchmarked against the VGG neural network architecture running the CIFAR-10 dataset, the proposed 2T-1FeFET CiM design achieves $89.45\%$ CIFAR-10 test accuracy. Compared to previous FeFET based CiM designs, it exhibits immunity to temperature drift at an 8-bit wordlength scale, and achieves better energy efficiency with 2866 TOPS/W.**

## I. Introduction

With the rapid development of artificial intelligence (AI) and Internet of things (IoT), there is an increasing demand for processing massive amount of data, thus calling for computing devices with low power consumption and read/write latencies [1]. Unfortunately, data-intensive applications are challenging to address with von Neumann architectures because of the "memory wall" problem [2], where frequent data transfers lead to high power consumption. To overcome such challenges, computing-in-memory (CiM) has been proposed as a potential solution for energy-efficient AI and IoT hardware designs [3–7], as it allows highly parallel arithmetic or logical calculations within the memory, which greatly increases the computing efficiency by eliminating the data transfers [8–13].

A number of non-volatile memory (NVM) devices have been considered for CiM designs, such as resistive random access memeory (ReRAM) [14], phase-change memory (PCM) [15], ferroelectric FET (FeFET) [16–22], etc. Among these NVMs, FeFET offers advantages such as CMOS compatibility, ultra-low leakage current, high $I_{ON}/I_{OFF}$ ratio and low write/read power. Prior works have reported FeFET based CiM designs for multiply-accumulation (MAC) operation, the core operator in AI models. Various FeFET based cells have been proposed to optimize the power consumption, speed, area and robustness [17, 19, 23], most of which operate in the saturation region of FeFETs. That said, continuous optimizations on the energy efficiency of CiM solutions are still desirable, given the billion level number of MAC operations required for neural network processing. This is where subthreshold computing comes into play. By scaling the operating voltage of FeFET to the subthreshold region (subthreshold-FeFET), FeFET based CiM array can achieve further reductions in power consumption compared to those operating in the saturation region.

On the other hand, the increased computation density in a compact area leads to higher power density and temperature elevation [24]. Also, the operating temperature will be affected by environmental conditions. The varying temperatures that may change the circuit operation states are normally neglected by FeFET based CiM designers, assuming a thermostatic condition. Unfortunately, FeFET is a highly temperature-sensitive device, and even more sensitive in the subthreshold region [25]. As temperature changes, the output of FeFET will change significantly, resulting in the error of circuits. Therefore, merely scaling the operating voltage to the subthreshold region is not sufficient, necessitating temperature-resilient FeFET based design.

In this paper, to further scale down the power consumption of FeFET based CiM design, and to solve the temperature drift problem associated with the FeFET operating conditions, for the first time, we propose a 2T-1FeFET cell design that almost eliminates the impact of temperature ranging from 0°C to 85°C on the subthreshold-FeFET based CiM array, while achieving ultra-low power consumption. The operation and the temperature resilience of the proposed design have been illustrated and validated at both cell and array levels. Evaluation results of the proposed design on VGG nerual network for Cifar-10 datasets demonstrate evident power consumption improvements with great resilience to temperature drift compared to other FeFET based CiM designs.

The paper is organized as follows: Sec. II provides a review

979-8-3503-4859-0/DATE24/© 2024 EDAA

of FeFET and subthreshold computation, and points out the design challenges. Sec. III analyzes impacts of temperature drift and proposes a temperature-resilient 2T-1FeFET cell. Sec. IV evaluates the performance of our design. Sec. V concludes.

## II. Background

In this section, we review the FeFET device and the model capturing the temperature dependency of FeFET, introduce the subthreshold CiM design, and analyze the design challenges of temperature-resilient subthreshold-FeFET based CiM.

### A. FeFET Basics

FeFETs, which utilize $HfO_2$ as the ferroelectric dielectric, have emerged as a competitive option for embedded NVM due to their ultra-low leakage current, high $I_{ON}/I_{OFF}$ ratio, voltage driven mechanisms and CMOS compatibility [26]. FeFETs allow for the switching of ferroelectric polarization within the gate layer by applying positive or negative gate pulses, which sets FeFET to the low-$V_{TH}$/high-$V_{TH}$ state [27], as shown in Fig. 1. The stored data (i.e., '1' and '0') can then be read through the drain current (i.e., $I_{ON}$ and $I_{OFF}$) by applying a gate voltage between the low-$V_{TH}$ and high-$V_{TH}$. Compared with other NVM devices, i.e., ReRAM [14], PCM [15], which require high power during the write due to large current, FeFETs provide superior energy efficiency due to the electric field driven write scheme [28].

Various models have been proposed to simulate FeFET devices in circuit designs, such as the negative capacitance FET (NCFET) based model [29], the multi-domain Preisach model [30], and the comprehensive Monte Carlo model [31]. However, these models lack consideration for temperature effects on devices, hindering the validation of functions and performance evaluation of FeFET based circuits in practical edge devices. While some studies have explored the impact of temperature on FeFET devices [25, 32] and vulnerability of CiM architectures to device variations [7], this work goes further by investigating and addressing the temperature effects on subthreshold-FeFET based CiM structures. To analyze the collective temperature effects on CiM at the circuit level, we utilize the experimentally calibrated Preisach FeFET compact model [30] in conjunction with the Intel FinFET model.

### B. Subthreshold Computing Design

Scaling the operating voltage to subthreshold region rather than saturation region is a promising way to optimize the energy efficiency, as the energy consumption is typically proportional to the conducting current and the voltage. That said, subthreshold computation has only been studied in mature devices such as CMOS [33]. These designs consume large area overhead and high power consumption, sacrificing the advantages of voltage scaling, and yet CiM designs based on subthreshold devices have not been studied so far.

To address the challenges faced by CMOS based subthreshold circuit design, in this work, for the first time, we propose a subthreshold-FeFET based CiM design for MAC operations. By leveraging the FeFET characteristics that are similar to MOSFET devices as shown in Fig. 1, subthreshold-FeFETs are acquired by applying scaled read voltage $V_{read}$ between the programmed memory window as shown in Fig. 1. As a result, the current $I_D$ of FeFETs is lowered down, achieving significant power reduction. However, similar to the CMOS based subthreshold computing, the subthreshold-FeFET is highly influenced by the temperature, as illustrated in Fig. 1. It can be seen that temperature changes have a stronger impact on the high-$V_{TH}$ state compared to the low-$V_{TH}$ state. The detailed impact of the temperature drift on the subthreshold-FeFET based CiM design is analyzed in Sec. III.

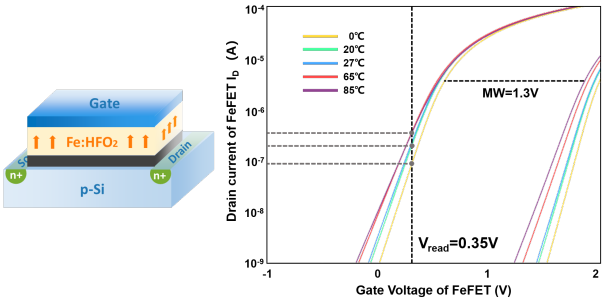


Fig. 1. FeFET structure and characteristic line at different temperatures and at two states (low-$V_{TH}$/high-$V_{TH}$). FeFETs working under the chosen $V_{read} = 0.35V$ in our design are fully lying in the subthreshold region.

### C. Subthreshold Operation vs. Temperature Resilience

Subthreshold-FeFET offers the advantage of reduced power consumption, while scaling voltages also makes the devices more susceptible to temperature drift, which deteriorates their performance or even functionality due to the exponential I-V relationship of FeFETs in the subthreshold region. Previous studies on CiM designs for MAC have already highlighted the challenges of achieving temperature-resilient design and subthreshold MAC operations. Combining both aspects further exacerbates the difficulty. Therefore, the challenge lies in finding an appropriate balance point between lower operating voltage and better resilience to temperature drift.

## III. Proposed Temperature-Resilient 2T-1FeFET CiM Design

In this section, we study the impacts of temperature drift on existing FeFET CiM designs operating in the subthreshold region, and propose a novel cell structure that aims at improving temperature resilience and energy efficiency based on subthreshold-FeFET.

### A. Temperature Drift on Existing Designs

To assess the significance of our proposed design, we first analyze traditional FeFET based CiM designs under different temperature conditions. As an example, we consider the basic 1FeFET-1R structure proposed in [17]. Fig. 2 illustrates the basic 1FeFET-1R array structure. For our analysis, we maintain all given parameters, e.g., write voltage and latency, and adjust the read voltage to investigate both the saturation region ($V_{read} = 1.3V$) and the subthreshold region ($V_{read} = 0.35V$).

First we analyze the output current of a single 1FeFET-1R cell. Fig. 3(a) and (b) illustrate the output current and the

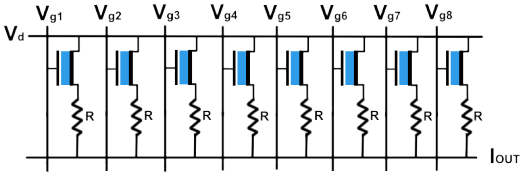


Fig. 2. The traditional 8-bit wordlength 1FeFET-1R structure [17].

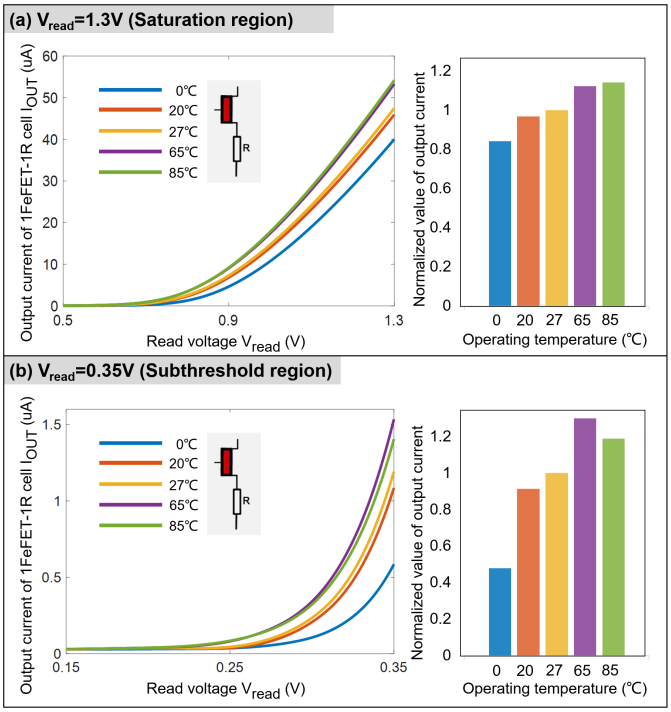


Fig. 3. Output current of 1FeFET-1R cell at temperatures from 0°C to 85°C under the same input voltages and normalized output current with reference temperature at 27°C. (a) $V_{read} = 1.3V$ (saturation region, which is the operating voltage of [17]). (b) $V_{read} = 0.35V$ (subthreshold region). Current fluctuation is measured by the difference of nominal output current to 1.

normalized output current to show current fluctuations (over the reference current at 27°C) at different temperatures with the same input voltages upon read when cells are operating in the saturation region and subthreshold region, respectively. When temperature rises from 0°C to 85°C, the fluctuation of the output currents reaches up to 20.6% (saturation region) and 52.1% (subthreshold region), respectively. These large current fluctuations in output current pose challenges in constructing larger-scale CiM arrays, where different MAC outputs may have overlapped output currents due to temperature drift, thus causing computation inaccuracies. Fig. 4 displays the range of output results obtained from an 1FeFET-1R CiM array with 8 cells per row, with a read voltage $V_{read}$ of 0.35V, under varying temperatures. It can be seen that problems arise when two distinct MAC outputs overlap with each other due to the impact of temperature drift.

We extend our analysis to several other FeFET based structures, specifically those operating in the subthreshold region, such as the one proposed in [23]. Upon examining these works, we observe that FeFET devices operating in or near the subthreshold region demonstrate expected performance within specific temperature conditions. However, these devices exhibit erroneous results when temperature changes, highlighting

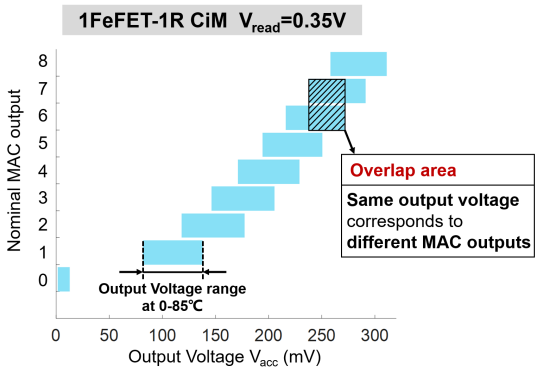


Fig. 4. Output voltage ranges at temperatures from 0°C to 85°C of an 1FeFET-1R CiM array with 8 cells per row. MAC outputs are from 0 to 8 and overlap with each other, causing computation errors.

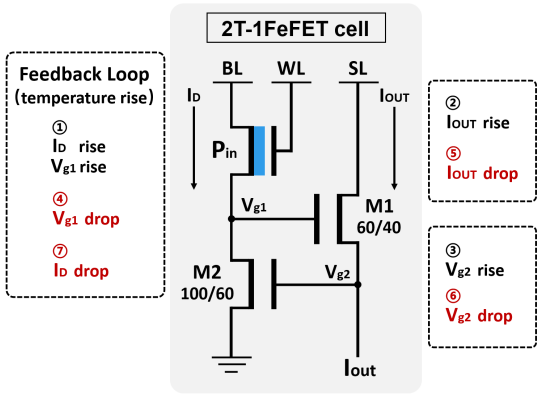


Fig. 5. Schematic of the proposed 2T-1FeFET cell design for CiM array, with the feedback loop for temperature drift compensation.

the need for robust temperature compensation techniques.

### B. *Proposed 2T-1FeFET Design*

To achieve temperature resilience, we propose a FeFET based cell as shown in Fig. 5. In this structure, two n-MOSFETs (M1 and M2) are connected in a ring configuration. Both M1 and M2 operate in the subthreshold region. The cell parameters, such as the W/L (width/length) ratio, read latencies, and write latencies, are tuned to improve the temperature resilience of the cell. As the temperature increases and the FeFET drain current $I_D$ rises, the gate voltage of M1 also increases due to the presence of M2. By keeping voltages on BL and SL lines constant, the drain current of M1 increases as a result of the increasing gate voltage. This increased drain current of M1, i.e., the output current, subsequently raises the voltage at the gate of M2. This leads to a decrease in $V_{gs1}$ (gate-source voltage of M1), limiting and reducing the drain current of M1, which is the output current of cell, and then leads to the voltage drop on $V_{gs2}$ (gate-source voltage of M2). Ultimately, the drain current of M2 and the FeFET is reduced. Similarly, the current drop of FeFET due to decreasing temperature will also rise back per the feedback loop. By employing this feedback mechanism, the drain currents of both M1 and M2 experience a drop/rise as temperature increases/decreases, effectively mitigating the impact of temperature drift on the cell output.

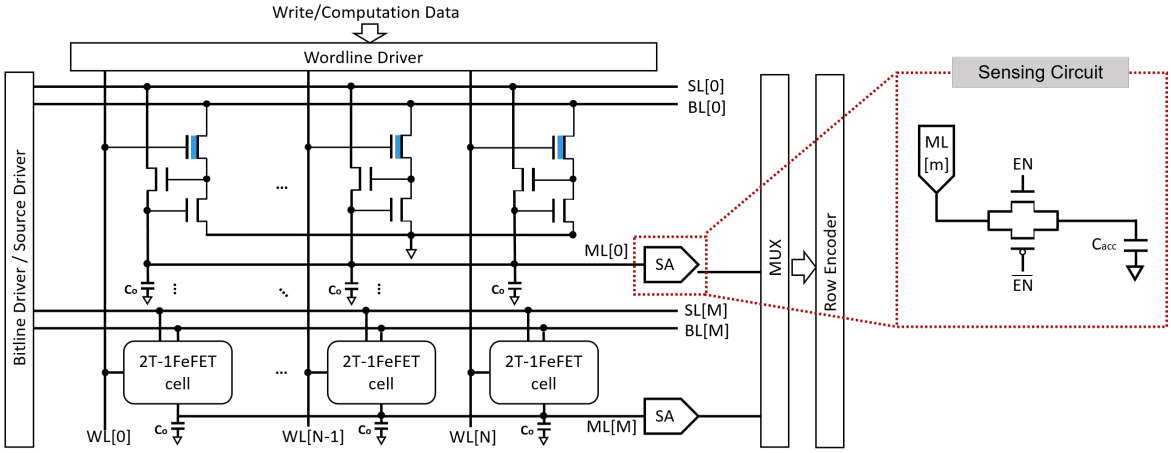


Fig. 6. Overall schematic of the proposed subthreshold 2T-1FeFET based CiM array and the associated sensing circuit.

The overall structure of the CiM array is built as shown in Fig. 6. Each row performs one MAC operation. In this design, we build a CiM array with 8 2T-1FeFET cells per row, with each cell associating with a small capacitor that converts current signals to stable voltage signals. The sensing circuit of array illustrated in Fig. 6 consists of a switch controlled by signal $EN$, and an accumulation capacitor $C_{acc}$ to obtain the output of MAC operations.

During the write operation of the 2T-1FeFET CiM array, we apply a -4V pulse for 200ns to WLs to program FeFETs to high-$V_{TH}$ state (logic '0'). On the other hand, to set the FeFETs to the low-$V_{TH}$ state (logic '1'), a +4V pulse for 115ns is applied. During the MAC operation, BL is set to 1.2V and SL is set to 0.2V. Within a row, each cell is controlled by an input WL line. When the input is '1', WL is set to 0.35V to activate subthreshold-FeFETs with low-$V_{TH}$ state, i.e., storing '1'. Conversely, when the input is '0', WL disables FeFETs, conducting no currents. If a FeFET stores high-$V_{TH}$ state, the multiplication result of the cell remains 0 regardless of input on WL. Each cell performs the multiplication operation based on the input voltage and the stored state, and multiplication currents are drawn from the SL lines, charging all the cell capacitance $C_o$s. After the charging, the $EN$ signal is set to a high level, and switches are opened simultaneously to charge the accumulation capacitance $C_{acc}$. The stored voltage of $C_{acc}$ can be calculated using equation (1):

$$V_{acc} = \frac{C_o}{nC_o + C_{acc}} \sum_{i=1}^{n} V_{Oi} \tag{1}$$

where $n$ is the number of cells connected to $C_{acc}$, and $V_{Oi}$ is the voltage level of each $C_o$. The array demonstrates resilience to temperature drift due to the temperature-resilient nature of its cells. Furthermore, the array exhibits enhanced performance as the effects of temperature drift on both the cells and the sensing circuit align in the same direction.

## IV. Validation and Evaluation

In this section, we first evaluate the performance of our proposed 2T-1FeFET CiM structure, including the resilience to temperature drift from 0°C to 85°C and process variation, and apply our design to VGG network with Cifar-10 dataset. In the end, we compare the the proposed 2T-1FeFET CiM structure with other emerging designs.The evaluations are performed on Cadence Virtuoso Spectre simulator.

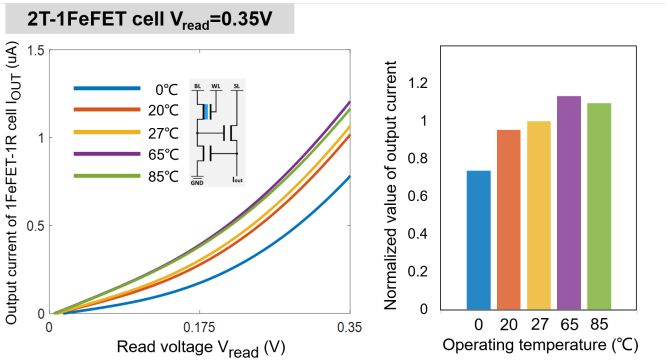


Fig. 7. Normalized output current of proposed 2T-1FeFET cell under the same input voltages with varying temperature, the reference temperature is 27°C. The 2T-1FeFET cell maintains similar temperature resilience to 1FeFET-1R structures that operate in the saturation region, and performs even better at temperatures above 20°C.

### A. Temperature Resilience Validation

Fig. 7 shows the temperature impact on our proposed cell design. The results are tested with the same input voltages and the reference temperature is 27°C (RT). The largest current fluctuation over the reference temperature of 2T-1FeFET cell is 26.6% at 0°C, which is close to the 20.6% fluctuation of 1FeFET-1R cell that operates in the saturation region (as revealed in Sec. III), and much better than subthreshold 1FeFET-1R cell, whose current fluctuation reaches 52.1%. Moreover, when temperature exceeds 20°C, our design outperforms the saturated 1FeFET-1R design, with the largest current fluctuation reduced to 12.4%.

We further perform simulation on 2T-1FeFET CiM array with 8 cells per row to measure the different outputs of MAC operation under different temperatures. Fig. 8 illustrates different MAC outputs and respective energy consumption of our proposed 2T-1FeFET CiM array with 8 cells per row. The different MAC outputs under different temperatures do not

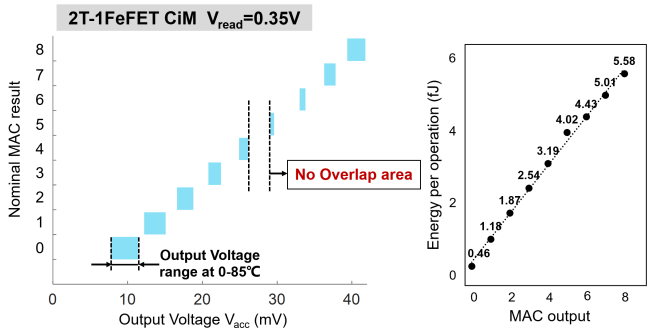


Fig. 8. (a) MAC output ranges of the proposed 2T-1FeFET array containing 8 cells per row with varying temperature. MAC outputs vary from 0 to 8. (b) Energy consumption per operation at different MAC outputs.

overlap, indicating stable CiM computation. Fig. 8(b) shows the energy consumption corresponding to different MAC outputs. With each MAC operation consisting of 8 multiplications and 1 accumulation, the average energy per operation is about 3.14fJ, and the energy efficiency reaches 2866 TOPS/W.

To better evaluate our 2T-1FeFET array, we define Noise Margin Rate ($NMR$) whose numerical expression is (2):

$$NMR_i = \frac{LV_{i+1} - HV_i}{HV_i - LV_i} \tag{2}$$

In this expression, $HV_i$ and $LV_i$ are the highest and lowest output voltages for $MAC = i$ within 0°C to 85°C. $NMR_i$ values the ratio of the gap distance between two different MAC outputs to the width of the $MAC = i$ output range. $NMR$ also signifies the probability of two different outcomes overlapping with each other. We further define $NMR_{min}$ expressed by equation (3):

$$NMR_{min} = min\{NMR_i\}, i = 0, 1, \ldots, 8 \tag{3}$$

to show the worst performance of the array from $MAC = 0$ to $MAC = 8$. A higher and positive value of $NMR_{min}$ indicates better performance of the CiM array, while a negative value suggests that the array may output overlaps and induce errors. As far as we have studied, all existing FeFET based CiM designs with 8 cells per row, whether operating in the saturation region or subthreshold region, exhibit $NMR_{min}$ values below 0 within 0°C to 85°C. Our array exhibits an $NMR_{min}$ value of $NMR_{min} = NMR_0 = 0.22$. When only considering the temperature range of 20°C to 85°C, the NMR value increases to $NMR_{min} = NMR_7 = 2.3$, indicating a significant improvement in sense margin. It is evident that our design has better optimization for the range of 20°C to 85°C than the range of 0°C to 20°C. Despite this, our design outperforms other related designs upon temperature drift, maintaining superior overall performance.

To further assess the impact of process variations on FeFET based CiM design, we run 100 Monte Carlo simulations of the 2T-1FeFET CiM array with an experimental FeFET Gaussian variability of $\sigma_{V_T} = 54mV$. The simulation results, shown in a histogram (Fig. 9), reveal that the highest error caused by process variation is approximately 25%, which is not significantly higher than other emerging CiM designs, such

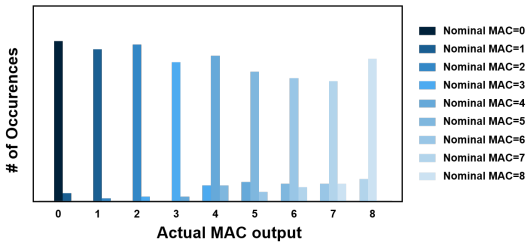


Fig. 9. The impact of process variation with $\sigma_{V_T} = 54mV$ at 27°C on CiM output via 100 Monte Carlo simulations.

TABLE I
THE STRUCTURE OF VGG EXECUTED ON CIFAR-10 DATASET

| Layer | Input Map | Output Map | Non Linearity |
|---|---|---|---|
| 64 3× 3 Conv1 | 32×32×3 | 32×32×64 | ReLU,dropout(0.3) |
| 64 3× 3 Conv2 | 32×32×64 | 32×32×64 | ReLU |
| [2, 2] MaxPool1 | 32×32×64 | 16×16×64 | – |
| 128 3× 3 Conv3 | 16×16×64 | 16×16×128 | ReLU,dropout(0.4) |
| 128 3× 3 Conv4 | 16×16×128 | 16×16×128 | ReLU |
| [2, 2] MaxPool2 | 16×16×128 | 8×8×128 | – |
| 256 3× 3 Conv5 | 8×8×128 | 8×8×256 | ReLU,dropout(0.4) |
| 256 3× 3 Conv6 | 8×8×256 | 8×8×256 | ReLU,dropout(0.4) |
| 256 3× 3 Conv7 | 8×8×256 | 8×8×256 | ReLU |
| [2, 2] MaxPool3 | 8×8×256 | 4×4×256 | – |
| 4096×4096 FC1 | 1×1×4096 | 1×1×4096 | ReLU,dropout(0.5) |
| 4096×4096 FC2 | 1×1×4096 | 1×1×4096 | ReLU,dropout(0.5) |
| 4096×10 FC3 | 1×1×4096 | 1×1×10 | – |

as the 6T SRAM CiM [34] that has a maximum error of 50% caused by process variation. Our proposed CiM array exhibits an error below 10% when reduced to 4 cells per row, which is comparable to the 1FeFET-1R design [17].

### B. Performance Evaluation of CiM Array

In this section, we evaluate the performance of the proposed design in the context of convolutional neural network (CNN) layers. Here we use the same CNNs and datasets as in [34], i.e., VGG and Cifar-10 dataset. The executed VGG network is summarized in Table I, and the performance of the 2T-1FeFET cell applied in the VGG layers is illustrated in Table II. Monte Carlo simulations of executing VGG on Cifar-10 using the proposed hardware suggest an average classification accuracy of 89.45%.

we compare our subthreshold-FeFET based CiM design with other existing CiM designs, which are SRAM CiM designs [34, 35], ReRAM CiM designs [14] and MTJ CiM designs [36]. As shown in Table II, we list several performance metrics of our design and other types of existing CiM designs. The 2T-1FeFET design demonstrates significantly lower power consumption compared to other energy-efficient designs, thanks to its high $I_{ON}/I_{OFF}$ ratio, and subthreshold computing mode. Other designs such as ReRAM and MTJ consume 64.6× and 445.9× more operation energy than 2T-1FeFET CiM array. This highlights the advantage on power consumption of subthreshold-FeFET based CiM designs over other designs. The proposed design has a latency of 6.9 ns for each MAC operation. While it may not be as fast as some other devices like the 1FeFET-1R, it remains competitively efficient. The lower operating voltage and accumulative capacitors contribute to the slightly higher latency.

From all these comparisons, we conclude that our 2T-1FeFET design has significant advantages on power consump-

TABLE II
PERFORMANCE SUMMARY

| Related Work | Device | Process | Cell Structure | Dataset | Network Architecture | Accuracy | Energy | Energy Efficiency |
|---|---|---|---|---|---|---|---|---|
| [34] | CMOS | 65nm | 6T SRAM | Cifar-10<br>MNIST | VGG<br>LeNet-5 | 88.83%<br>99.05% | NA<br>158.203nJ (/inference) | NA |
| [35] | CMOS | 65nm | 12T SRAM | Cifar-10 | BNN | 85.7% | 2.48-7.19fJ (/operation) | 403 TOPS/W |
| [17] | FeFET | 28nm | 1FeFET-1R | / | / | / | NA | 13714 TOPS/W |
| [19] | FeFET | 28nm | 1FeFET-1T | MNIST | MLP | 97.6% | 17.6uJ (/inference) | NA |
| [14] | ReRAM | 22nm | 1T-1R | Cifar-10 | VGG | 91.72% | ≈5.5uJ (/inference) | 26.66 TOPS/W |
| [36] | MTJ | 28nm | 1T-1MTJ | / | / | / | 1.4pJ (/operation) | 32 TOPS/W |
| This Work | FeFET | 14nm | 2T-1FeFET | Cifar-10 | VGG | 89.45% | 85.08nJ (/inference)<br>3.14fJ (/operation) | 2866 TOPS/W |

tion while being resilient to temperature variations, and the accuracy of our design applied in VGG network is relatively high. Notably, We are the first to realize the temperature-resilient subthreshold-FeFET based CiM array with 8 cells per row within 0°C to 85°C range, especially at temperature above 20°C, whose property other existing designs do not possess.

## V. CONCLUSION

In this paper, to build an ultra-low power CiM design for practical edge scenarios, we investigate the computation failure caused by thermal variations of FeFET based CiM structures, and propose a novel subthreshold 2T-1FeFET cell design and an ultra-low power CiM array with 8 cells per row that are immune to output overlap induced operation failure at temperatures ranging from 0°C to 85°C. The developed subthreshold-FeFET based CiM design has a remarkable reduction on power consumption compared to other emerging designs, and a remarkable energy efficiency of averaging 2866 TOPS/W for the CiM array with 8 cells per row. Evaluation also shows $89.45\%$ accuracy of the proposed design on the VGG neural network architectures running the Cifar-10 dataset. Overall, our proposed CiM structure offers a promising solution to mitigate the impact of temperature variations, reduce power consumption, and deliver reliable MAC operations for neural networks deployed in edges.

## ACKNOWLEDGEMENTS

This work was supported in part by National Key R&D Program of China (2020YFB1313501), Zhejiang Provincial Natural Science Foundation (LD21F040003, LQ21F040006), NSFC (62104213, 92164203).

## REFERENCES

[1] J. Deng *et al.*, "Energy-efficient real-time uav object detection on embedded platforms," *IEEE TCAD*, vol. 39, no. 10, pp. 3123–3127, 2019.
[2] W. A. Wulf *et al.*, "Hitting the memory wall: Implications of the obvious," *SIGARCH Comput. Archit. News*, vol. 23, no. 1, p. 20–24, 1995.
[3] X. Yin *et al.*, "An ultracompact single-ferroelectric field-effect transistor binary and multibit associative search engine," *Adv. Intelli. Syst.*, p. 2200428, 2023.
[4] Y. Wei *et al.*, "Imga: Efficient in-memory graph convolution network aggregation with data flow optimizations," *IEEE TCAD*, 2023.
[5] A. Eldebiky *et al.*, "Correctnet: Robustness enhancement of analog in-memory computing for neural networks by error suppression and compensation," in *DATE*, pp. 1–6, IEEE, 2023.
[6] S. Shou *et al.*, "See-mcam: Scalable multi-bit fefet content addressable memories for energy efficient associative search," in *IEEE/ACM ICCAD*, pp. 1–9, IEEE, 2023.
[7] Z. Yan *et al.*, "Computing-in-memory neural network accelerators for safety-critical systems: Can small device variations be disastrous?," in *ACM ICCAD*, pp. 1–9, 2022.
[8] X. Chen *et al.*, "Accelerating graph-connected component computation with emerging processing-in-memory architecture," *IEEE TCAD*, vol. 41, no. 12, pp. 5333–5342, 2022.
[9] Z. Yan *et al.*, "Swim: Selective write-verify for computing-in-memory neural accelerators," in *ACM/IEEE DAC*, pp. 277–282, 2022.
[10] X. Yin *et al.*, "Ferroelectric ternary content addressable memories for energy-efficient associative search," *IEEE TCAD*, vol. 42, no. 4, pp. 1099–1112, 2022.
[11] X. Wang *et al.*, "Triangle counting accelerations: From algorithm to in-memory computing architecture," *IEEE TC*, vol. 71, no. 10, pp. 2462–2472, 2021.
[12] Q. Huang *et al.*, "Fefet based in-memory hyperdimensional encoding design," *IEEE TCAD*, 2023.
[13] Z. Yan *et al.*, "Improving realistic worst-case performance of nvcim dnn accelerators through training with right-censored gaussian noise," in *IEEE/ACM ICCAD*, pp. 1–9, IEEE, 2023.
[14] S. Yu *et al.*, "Rram for compute-in-memory: From inference to training," *IEEE TCAS-I*, vol. 68, no. 7, pp. 2753–2765, 2021.
[15] R. Khaddam-Aljameh *et al.*, "Hermes core – a 14nm cmos and pcm-based in-memory compute core using an array of 300ps/lsb linearized cco-based adcs and local digital processing," in *2021 Symposium on VLSI Technology*, pp. 1–2, 2021.
[16] K. Ni *et al.*, "Ferroelectric ternary content-addressable memory for one-shot learning," *Nature Electronics*, vol. 2, no. 11, pp. 521–529, 2019.
[17] T. Soliman *et al.*, "Ultra-low power flexible precision fefet based analog in-memory computing," in *2020 IEEE IEDM*, pp. 29.2.1–29.2.4, 2020.
[18] J. Cai *et al.*, "Energy efficient data search design and optimization based on a compact ferroelectric fet content addressable memory," in *ACM/IEEE DAC*, pp. 751–756, 2022.
[19] M. R. Sk *et al.*, "1f-1t array: Current limiting transistor cascoded fefet memory array for variation tolerant vector-matrix multiplication operation," *IEEE Transactions on Nanotechnology*, vol. 22, pp. 424–429, 2023.
[20] C.-K. Liu *et al.*, "Cosime: Fefet based associative memory for in-memory cosine similarity search," in *IEEE/ACM ICCAD*, pp. 1–9, 2022.
[21] L. Liu *et al.*, "A reconfigurable fefet content addressable memory for multi-state hamming distance," *IEEE TCAS-I*, 2023.
[22] H. Xu *et al.*, "On the challenges and design mitigations of single transistor ferroelectric content addressable memory," *IEEE Electron Device Letters*, 2023.
[23] X. Lu *et al.*, "In-memory multi-bit multiplication and accumulation (mac) using fefet for energy efficient iot," in *ICITES*, pp. 27–33, 2022.
[24] J. Meng *et al.*, "Temperature-resilient rram-based in-memory computing for dnn inference," *IEEE Micro*, vol. 42, no. 1, pp. 89–98, 2022.
[25] A. Gupta *et al.*, "Temperature dependence and temperature-aware sensing in ferroelectric fet," in *IEEE IRPS*, pp. 1–5, 2020.
[26] X. S. Hu *et al.*, "In-memory computing with associative memories: A cross-layer perspective," in *IEDM*, pp. 25–2, IEEE, 2021.
[27] A. I. Khan *et al.*, "The future of ferroelectric field-effect transistor technology," *Nature Electronics*, vol. 3, no. 10, pp. 588–597, 2020.
[28] S. Salahuddin *et al.*, "The era of hyper-scaling in electronics," *Nature Electronics*, vol. 1, no. 8, pp. 442–450, 2018.
[29] A. Aziz *et al.*, "Physics-based circuit-compatible spice model for ferroelectric transistors," *EDL*, vol. 37, no. 6, pp. 805–808, 2016.
[30] K. Ni *et al.*, "A circuit compatible accurate compact model for ferroelectric-fets," in *VLSI*, pp. 131–132, IEEE, 2018.
[31] S. Deng *et al.*, "A comprehensive model for ferroelectric fet capturing the key behaviors: Scalability, variation, stochasticity, and accumulation," in *IEEE VLSI*, 2020.
[32] S. Thomann *et al.*, "On the reliability of in-memory computing: Impact of temperature on ferroelectric tcam," in *VTS*, pp. 1–6, IEEE, 2021.
[33] C. Loyez *et al.*, "Subthreshold neuromorphic devices for spiking neural networks applied to embedded a.i," in *IEEE NEWCAS*, pp. 1–4, 2021.
[34] M. Ali *et al.*, "Imac: In-memory multi-bit multiplication and accumulation in 6t sram array," *IEEE TCAS-I*, vol. 67, no. 8, pp. 2521–2531, 2020.
[35] S. Yin *et al.*, "Xnor-sram: In-memory computing sram macro for binary/ternary deep neural networks," *IEEE JSSC*, vol. 55, no. 6, pp. 1733–1743, 2020.
[36] V. K. Jacob *et al.*, "A nonvolatile compute-in-memory macro using voltage-controlled mram and in-situ magnetic-to-digital converter," *IEEE JxCDC*, 2023.