2024 IEEE 24th International Conference on Nanotechnology (NANO)
July 8-11, 2024. Gijón, Spain

5-bit Signed SRAM-Based In-Memory Computing Cell

F. Karimpour, F. Pardo, D. García-Lesta
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
Santiago de Compostela, Spain
email: faranak.karimpour@usc.es

***Abstract*— Hardware accelerators are critical in providing real-time processing for edge computing applications, particularly in the context of convolutional neural networks. A crucial challenge in this context is achieving low power consumption while maintaining an appropriate performance in terms of accuracy. This work delves into a thorough analysis of prospective architectures for the core cell of the multiply-and-accumulate function, monitoring each structure's crucial benefits and drawbacks. It includes electrical simulations comparing their performance in a 180 nm process node for 1.8 V and 3.3 V. Moreover, a process corner simulation is proposed to identify on-chip process variations in the voltage error of the proposed design under different input voltages. Notably, the minimum corner errors observed at +15 and -7 sign bits are 0.45% and 0.63%, respectively. The significant outcome highlights that the single-switch implementation achieves optimal performance, displaying the lowest error value of 0.14%, specifically at the +15 sign bit and operating at 1.8 V.**

***Index Terms*— In-memory computing, convnets, neural network hardware accelerator, mixed-mode, CNN.**

I. Introduction

Computer vision has made great strides in recent years due to the increased adoption of deep learning algorithms. These methods have significantly advanced the state-of-the-art in numerous tasks, including speech analysis, object recognition, and face recognition [1]. A deep neural network (DNN) comprises successive layers with different operations, where convolutions are the primary focus and critical components in DNN accelerator designs [2]. Convolutional Neural Networks (CNNs) use convolutional layers for feature extraction. However, they struggle with real-time performance and energy efficiency, as convolutional layers involve multiplication and accumulation (MAC) operations [3]. While MAC DNN accelerators have been developed to meet computational demands, they often lead to increased power consumption, posing challenges for deploying Edge Artificial Intelligence (AI) [4].

Unlike traditional Von Neumann architectures, DNN hardware accelerators employ memory-efficient designs such as CMOS or emerging memory technologies like FeRAM for In-Memory-Computing (IMC), as depicted in Fig. 1. These innovative designs aim to minimize data transfers between memory and processing units, ultimately reducing communication and power requirements [5]. Therefore, leveraging IMC to accelerate MAC operations is strongly advocated to achieve superior performance and energy efficiency in on-chip DNNs [6].

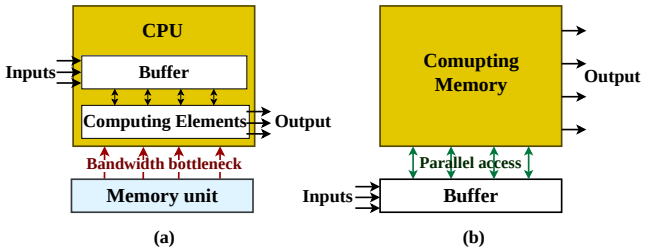


Fig. 1: (a) Von Neumann architecture, in which CPU and memory are separated and connected through a high-bandwidth bus, (b) SRAM-based IMC, in which the computation is performed directly in the SRAM memory array.

This design choice is significant for edge-based AI, especially in the Internet of Things (IoT). The paper evaluates diverse 5-bit IMC designs in a 180 nm process node operating at 1.8 V and 3.3 V. Our investigation includes a comprehensive study of architecture and voltage's impact on system performance and a thorough analysis of errors in single and 3×3 multipliers. The multipliers used in this work are based on the concept of time-current multiplication [7], [8]. These cells rely on a controllable current source that can be selectively turned on and off in diverse ways. The paper addresses the impact of various strategies to control the current source. It delves into two approaches to positioning switches behind (Pre-Swtich Structure - PrSS) or after (Post-Switch Structure - PSS) a single transistor current source. These switches utilize a single transistor (S-SW) controlled by some logic or three transistors in series (M-SW). These analyses are repeated for the two voltage domains presented in the technology node and evaluate whether increasing the power consumption of the 3.3 V domain works in the case of the 1.8 V voltage domain.

The paper is structured as follows: Section II introduces the baseline architecture and the design of the 5-bit multiplier. Section III explores various implementation

This work has received funding from projects PID2021-128009OB-C32 and TED2021-132372B-I00 from the MCIN/AEI/10.13039/501100011033 and FEDER and European Union NextGenerationEU/PRTR and the European Regional Development Fund: Xunta de Galicia-Consellería de Cultura, Educación e Ordenación Universitaria Accreditation 2019–2022 ED431G-2019/04 and Reference Competitive Group Accreditation 2021–2024, GRC2021/48, and from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101016734.

2024 IEEE 24th International Conference on Nanotechnology (NANO) | 979-8-3503-8624-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/NANO61778.2024.10628891

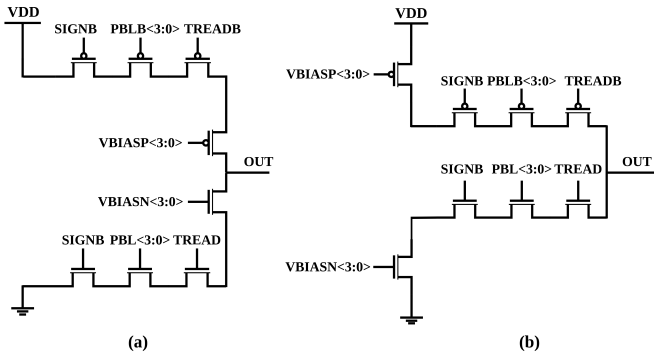


Fig. 2: (a) In-memory multiplier concept in pre-switch structure (PrSS) and (b) post-switch structure (PSS).

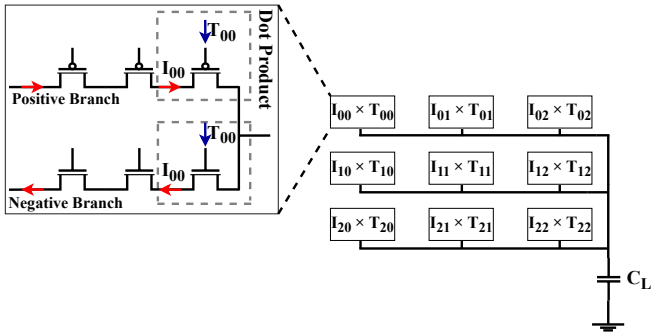


Fig. 3: MAC operation with different weights and inputs for CNN.

approaches, while in Section IV, we present a quantitative analysis of the multiplier's performance. Finally, Section V summarizes our findings and gives an outlook and conclusions.

II. 5-bit Multiplier Design

In this work, we employ current-time multiplication to carry out the MAC operations in the analog domain. This method involves switching a controlled current, proportional to a filter weight, across the entire period of the input feature map. Subsequently, this current is integrated into a capacitor, yielding the desired result [9]. It's worth noting that image filters and algorithms often have the potential to reuse the same weights across different sections or layers. This characteristic suits them particularly for IMC architectures, where the same weights can be stored and utilized without frequent updates. In our baseline architecture, the primary operation is convolution, which efficiently computes inner products, generating output feature map values ($Out^n_{x,y}$). Here, $I^n_{i,j,k}$ represents the weights associated with a set of N 2-D filters (where n ranges from 1 to N), and $T_{x,y,k}$ corresponds to the input data. This collective process creates a 3-D feature map [10]. The mathematical representation of this process is as follows:

$$Out^n_{x,y} = \sum_{i,j,k} \mathrm{I}^n_{i,j,k} \times T_{x+i,y+j,k} \tag{1}$$

We introduce the customized IMC multipliers for a 5-bit signed system and delve into a detailed analysis of a single multiplier, providing insights into its operation and performance. As previously mentioned, two different 5-bit signed multipliers, PrSS and PSS, illustrated in Fig. 2(a), Fig. 2(b) respectively. A 5-bit signed multiplier optimized for computational efficiency in DNN operations, with a distinct focus on the matrix dot product, is presented. The configuration depicted in Fig. 3 comprises a grid of MAC units in a 3×3 layout. Each unit multiplies a ratio element, such as I_{00}, with a corresponding input value, such as T_{00}, as depicted in the magnified section of Fig 3. This representation corresponds to the initial multiplier within the 3×3 weight matrix. The contributions of these units are accumulated in a shared capacitor (C_L), yielding the final voltage result.

Both designs are configured as signed 5-bit multipliers with weighting bits ($I_{i,j,k}$) obtained from nearby digital memory cells like SRAM memory cells. In the case of 0 stored in the bit-cell, NMOS remains OFF, while PMOS transistors are activated, resulting in zero current in the bottom branch and achieving the desired current level in the upper branch. On the other hand, a 1 (PBL = 1, PBLB = 0) reverses this behavior. Pulse Width Modulation (PWM) signals pass input feature maps through transistors denoted as $Tread_{i,j,k}$. The products of all multipliers, represented as ($I_{i,j,k} \times Tread_{i,j,k}$), are accumulated as charges and converted to voltages using a load capacitor (C_L), initially charged to a predetermined voltage (V_{PRE}). The voltage evolution (V_{Out}) at the capacitor C_L over time increases when current is drawn into the capacitor and decreases when the opposite occurs. This behavior is mathematically expressed by Eq. (2).

$$V_{out,x,y,n} = \frac{1}{C_L} \sum_{i,j,k} \mathrm{I}^n_{i,j,k} \times Tread_{x+i,y+j,n+k} \tag{2}$$

The flexibility in adjusting key parameters such as C_L, $Tread_{i,j,k}$, and the least significant bit current offer advantages in hardware reuse across different layers.

Capacitance values were precisely chosen to ensure the transistor saturation based on weight current and Tread LSBs, optimized at 128 fF, 288 fF at 1.8 V, and 3.3 V, respectively, in a single 5-bit multiplier. The PrSS structure shows an uninterrupted path for the input current with the summing capacitor, effectively eliminating switch-related interference caused by parasitic capacitance. Conversely, the PSS structure, with switches in the input current path, could reduce the total current reaching the capacitor due to the introduction of undesired interference.

Regarding switching errors, the PrSS keeps providing current for a specific time once the source is switched off. This phenomenon occurs because the drain to source voltage of the transistor that implements the current source remains high while the parasitic capacitance in the supply path is not fully discharged. On the other hand, in the

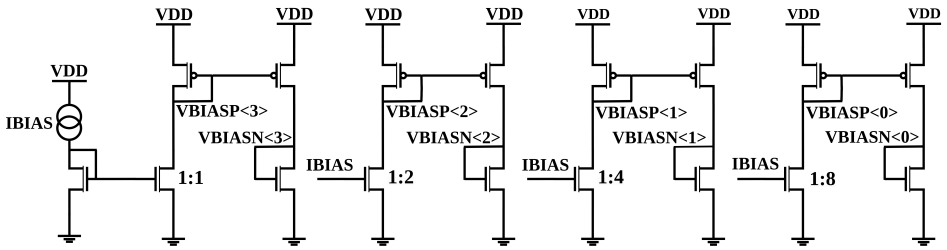


Fig. 4: Current mirror biasing circuit for computing MAC operation with 4-bit weight precision. The transistors are mirrored in the ratio 1: 2: 4: 8 for the four columns.

PSS structure, dynamic and switching errors emerge due to charge injection from the on/off switch. The PSS structure has higher parasitic capacitance but can be mitigated using a larger C_L capacitance [7].

Fig. 4 shows the biasing circuitry responsible for generating the required biasing voltages (VBIASP and VBIASN) for the current sources. It uses a reference current to develop all four currents needed for the 5-bit multipliers, positive and negative. Transistors that implement the current sources share identical dimensions but differ in bias voltages based on position and polarity.

III. Implementation Variations and Comparative Analysis

We initially implemented a three-switch configuration before transitioning to a simplified single-switch design. Our evaluations revealed approximately 283 nA and -289 nA for the PrSS structure and 260 nA and -262 nA for the PSS configuration with multiple switches. These currents nearly reached the 300 nA target for all significant bits. Fig. 5 shows our transition from a three-switch design to a single-transistor solution, enhancing performance and reducing error voltage. Table I compares errors between single and multiple switch configurations.

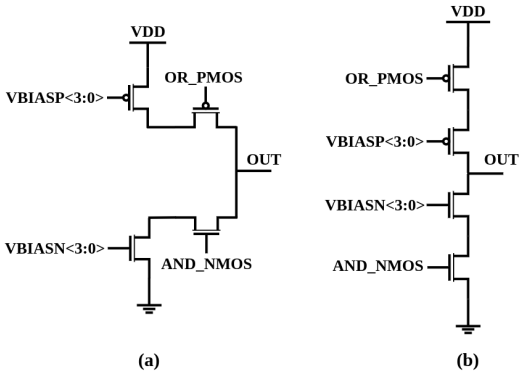


Fig. 5: Simulation of single switch performance in (a) PSS and (b) PrSS structures at 1.8 V.

Subsequently, we observed the system's performance in the 1.8 V and 3.3 V voltage domains using a single switch and three switch configurations, each utilizing a single signed 5-bit multiplier. Fig. 6 depicts voltage changes over time in

TABLE I: Maximum current output error in Single-Switch (S-SW) and Multiple-Switch (M-SW) for a single multiplier at 1.8 V and 3.3 V.

Configuration	1.8 V Error (%)	3.3 V Error (%)
S-SW PSS	6.50	1.33
M-SW PSS	12.60	4.25
S-SW PrSS	4.16	0.66
M-SW PrSS	4.56	1.83

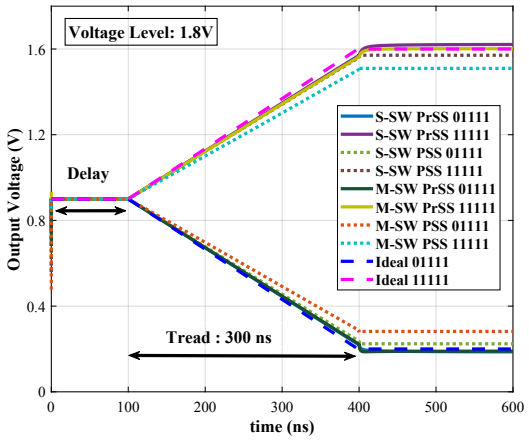


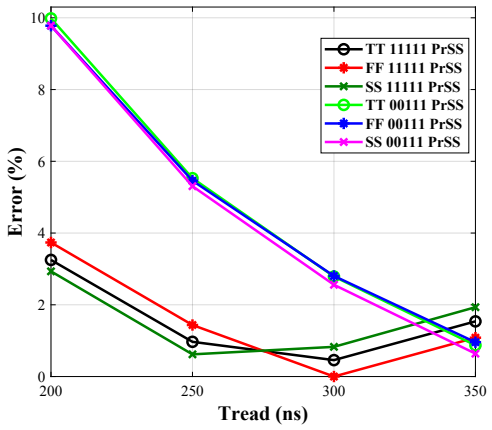
Fig. 6: Simulation results of voltage versus time for single and multiple switches in the one multiplier (1MLP) configuration accompanied by the ideal value at 1.8 V.

the load capacitor (C_L), considering 5-bit signed weights. These simulations evaluate the system performance under specific conditions, including consequences of -15 and +15 and maximum input values. In these simulations, the parameters include the least significant bit current (I_{LSB}), which is 20 nA, a load capacitor value of 128 fF, and processing time 0.3μs at 1.8 V. In the 3.3 V scenario, these parameters change to 40 nA, 288 fF, and 0.6μs. The highest output voltage (V_{outmax}) can be calculated using the following equation:

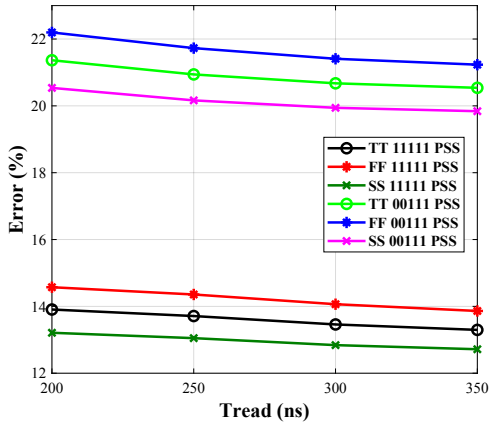
$$V_{outmax} = 15 \times 31 \times \frac{I_{LSB} T_{LSB}}{C_{sum}} + V_{Pre} \tag{3}$$

Analyzing the error in output voltages across various input voltages is the focus of this study, and the error percentage, represented as η and calculated through:

$$\eta = \frac{V_{out_{ideal}} - V_{out_{sim}}}{V_{out_{ideal}}} \tag{4}$$



(a)



(b)

Fig. 7: Examining typical (TT), fast (FF), and slow (SS) corners involves evaluating voltage error at various input voltages, with a specific emphasis on the -7 and +15 sign bits in both (a) PrSS and (b) PSS structures.

Eq. (4) quantifies the deviation between the ideal value ($V_{out_{ideal}}$), which is calculated using Eq. (1), introduced earlier and the simulated output ($V_{out_{sim}}$). For the 3×3 multiplier, we adjusted the load capacitance (C_L) to be nine times the value of a single multiplier, totaling 1152 fF. Notably, the single switch configuration consistently outperforms the multiple switch configuration regarding total current and error performance across both voltage domains.

Various statistical simulation methods are widely used for circuit analysis. These models and simulation approaches are employed to explore the effects of local variations and identify the most critical performance values across different corners. The objective is to understand how changes in input characteristics influence the accuracy of the system's output. Ensuring that the designed circuit meets all design objectives and constraints in various corners enhances its robustness and yield [11]. The PrSS and PSS structures are evaluated at three corners: TT (typical NMOS, typical PMOS), SS (slow NMOS and slow PMOS case), and FF (fast NMOS and fast PMOS case). This evaluation helps assess the circuit's performance under different process variations.

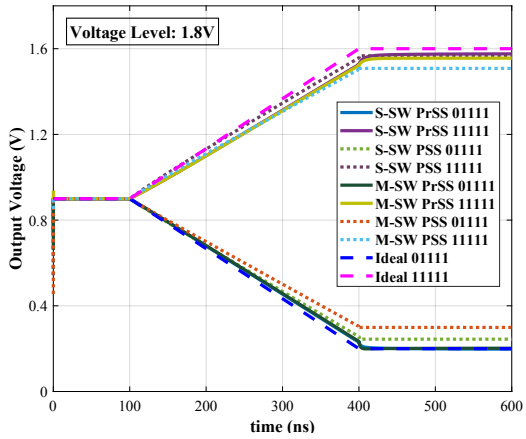


Fig. 8: Simulation results of 3×3 5-bit signed multiplier (9MLP) with 3×3 input with single and multiple switches accompanied by the ideal value at 1.8V

The simulations for process variations are depicted in Fig. 7.

The evaluation analyzes the error voltage under various input voltages (Treads) at intervals of 200ns, 250ns, 300ns, and 350ns. Examining two different sign bits, +15 and -7, within the PSS structure revealed error percentages ranging from a minimum of 12.72% to a maximum of 14.57% for the +15 sign bit and from a minimum of 19.84% to a maximum of 22.20% for the -7 sign bit. Similarly, the PrSS structure demonstrated error percentages spanning from a minimum of 0.45% to a maximum of 3.73% for the +15 sign bit and from a minimum of 0.63% to a maximum of 9.78% for the -7 sign bit.

These findings highlight the circuit's robust performance across various critical scenarios. As can be seen, the PrSS structure supports minor variations, so it must be optimized to guarantee better performance [12].

IV. Quantitative Analysis of the In-Memory Multiplier Design

It is worth noting that the number of multipliers that can be joined together is not limited, and this feature makes the design adaptable for any kernel size. This adaptability significantly enhances the versatility of the 5-bit multipliers in accommodating different kernel sizes. As a use case in this study, we have analyzed a 3×3 kernel size. Fig. 8 illustrates the voltage evolution in the summing capacitor for inputs and maximum weight values, with positive and negative sign bits. In analyzing the errors quantitatively, Table II underscores the superior performance of the S-SW PrSS configuration, achieving a remarkable 0.14% error for positive weights at 1.8 V. However, the M-SW PrSS configuration exhibits notably higher errors, reaching 14.23% and 13.16% for positive and negative weights at 1.8 V. These outcomes are particularly noteworthy due to the current flow deviating into parasitic capacitors rather than the load capacitor. Adjusting the load capacitor and period could address this issue, but we maintained consistent current values across structures to ensure a uniform performance evaluation.

TABLE II: Error Evaluation in Single-Switch (S-SW) and Multiple-Switch (M-SW) Configurations for 3×3 MLP at 1.8 V and 3.3 V for Positive and Negative Weights.

Configuration	1.8 V Error (%)	3.3 V Error (%)
S-SW PSS	6.23, 4.74	0.34, 2.14
M-SW PSS	14.23, 13.16	3.93, 2.24
S-SW PrSS	0.14, 1.55	1.06, 4.80
M-SW PrSS	0.79, 7.52	1.46, 7.11

A quantitative analysis of those errors is depicted in Table II.

In validating the system's overall performance within the PrSS structure, an initial simulation employed a 3×3 vertical edge detector kernel on the input image, i.e.:

$$Kernel = \begin{pmatrix} -15 & 0 & 15 \\ -15 & 0 & 15 \\ -15 & 0 & 15 \end{pmatrix} \tag{5}$$

The normalized outputs resulting from the convolution of the 16x16 pixel image of Fig. 9(a) with the kernel (Eq. (5)) using a stride of 1 are presented in Fig. 9(b) and Fig. 9(c) for the ideal and electrical simulation, respectively. Subsequently, the difference between the ideal and simulated output is illustrated in Fig. 9(d). Overall, these figures highlight the effective operation of the edge detection process. The root mean square error (RMSE) between the ideal and simulated output is 0.173, which quantitatively measures the dissimilarity between the expected and simulated results.

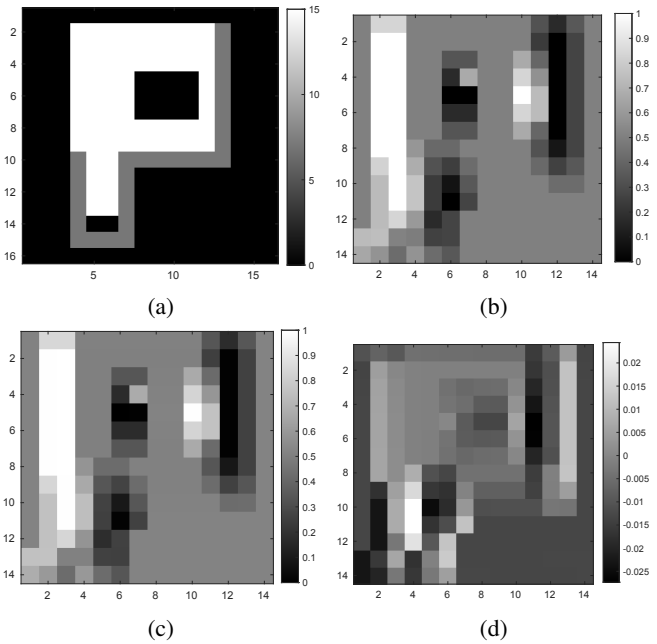


Fig. 9: Example of convolution using the PrSS structure: (a) a 16×16×1 input image, (b) ideal edge detection, (c) simulated edge detection, (d) difference between (b) and (c) matrices.

V. Outlook and Conclusions

This paper compares CNN hardware accelerators based on current-time multiplication architectures. Adjusting the switching scheme and parasitic capacitance of the core cell significantly affects performance.

The results favor a single-switch implementation for optimal performance despite the increase in silicon area. Considering switch configuration, the PSS architecture achieves notable accuracy; however, combining multiple 3×3 multiplier groups for kernel patches decreases accuracy. This effect is mainly due to parasitic capacitance influencing the summing dependence of weights in each kernel patch.

Corner simulations indicate varying errors across different structures, with the PrSS circuit demonstrating promising performance alignment with nominal cases.

This work is the initial step in developing a complex system with many kernel patches connected to compute 3D filters. Therefore, future work will include studying the effect on system performance. Moreover, additional figures of merit, including power consumption and processing speed, will be analyzed to enhance the comprehensive evaluation of the optimal solution.

References

[1] A. Ankit, I. Chakraborty, A. Agrawal, M. Ali, and K. Roy, "Circuits and architectures for in-memory computing-based machine learning accelerators," *IEEE Micro*, vol. 40, no. 6, pp. 8–22, 2020.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] S.-S. Park and K.-S. Chung, "Cenna: cost-effective neural network accelerator," *Electronics*, vol. 9, no. 1, p. 134, 2020.

[4] N. Verma, H. Jia, H. Valavi, Y. Tang, M. Ozatay, L.-Y. Chen, B. Zhang, and P. Deaville, "In-memory computing: Advances and prospects," *IEEE Solid-State Circuits Magazine*, vol. 11, no. 3, pp. 43–55, 2019.

[5] Z. Chen, Z. Yu, Q. Jin, Y. He, J. Wang, S. Lin, D. Li, Y. Wang, and K. Yang, "Cap-ram: A charge-domain in-memory computing 6t-sram for accurate and precision-programmable cnn inference," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 6, pp. 1924–1935, 2021.

[6] X. Wang, G. Li, J. Sun, H. Fan, Y. Chen, and H. Jiao, "Ternary in-memory mac accelerator with dual-6t sram cell for deep neural networks," in *2022 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*. IEEE, 2022, pp. 246–250.

[7] O. Pereira-Rial, D. García-Lesta, V. Brea, P. López, and D. Cabello, "Design of a 5-bit signed sram-based in-memory computing cell for deep learning models," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2022, pp. 702–706.

[8] D. G. Lesta, Ó. P. Rial, V. Brea, P. López, and D. Cabello, "A general-purpose cmos vision sensor with in-pixel 5-bit convolutional layer computation," in *2022 29th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. IEEE, 2022, pp. 1–4.

[9] L. Baischer, M. Wess, and N. TaheriNejad, "Learning on hardware: A tutorial on neural network accelerators and co-processors," *arXiv preprint arXiv:2104.09252*, 2021.

[10] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A 64-tile 2.4-mb in-memory-computing cnn accelerator employing charge-domain compute," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 6, pp. 1789–1799, 2019.

[11] L. Yang, D. Li, X. Yang, X. Feng, L. Tan, C. Shen, and H. Cai, "A temperature-insensitive process corner detection circuit based on self-timing ring oscillator," *Microelectronics Reliability*, vol. 138, p. 114779, 2022.

[12] V. H. Carbajal-Gomez, E. Tlelo-Cuautle, J. M. Muñoz-Pacheco, L. G. de la Fraga, C. Sanchez-Lopez, and F. V. Fernandez-Fernandez, "Optimization and cmos design of chaotic oscillators robust to pvt variations," *Integration*, vol. 65, pp. 32–42, 2019.