A CMOS Analog Neuron Circuit with A Multi-Level Memory

Melvin D. Edwards II, *Student Member, IEEE*, Nabil J. Sarhan, *Member, IEEE*, Mohammad Alhawari, *Member, IEEE*

2023 International Conference on Microelectronics (ICM) | 979-8-3503-8082-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICM60448.2023.10378912

***Abstract*—This paper presents a CMOS-based analog neuron circuit that utilizes a multi-level analog memory that is useful for mixed signal neural networks. The implementation of neural networks in the analog/mixed signal domain is crucial for edge applications of Artificial Intelligence (AI). The proposed circuit is able to generate both positive and negative output Multiply and Accumulate (MAC) currents by utilizing a bipolar supply. The multi-level analog memory is able to generate eight distinct analog voltages to drive the analog neuron circuit. Simulation results show that the analog neuron circuit is able to perform linear addition over the entire MAC current range. The MAC current per input-weight pair is between $\pm 25\mu$A to $\pm 56\mu$A. The analog neuron circuit is able to generate 0A for zero-weight or zero-input. Each input-weight pairs consumes 170μW. The circuit is designed and simulated in the 65 nm technology node.**

***Index Terms*—Analog Neuron, Two-Quadrant Analog Neuron, Positive and Negative MAC (Multiply and Accumulate) Currents, Bipolar Supply, Analog Memory Circuit, 3-Bit Memory, Neuron Circuit, Edge Computing, 65 nm technology**

I. Introduction

The integration of Artificial Intelligence (AI) into various devices and systems, particularly through the use of Deep Neural Networks (DNNs), has led to significant advancements in computer vision [1]–[3] and natural language processing [4]–[6]. However, as the Internet of Things (IoT) continues to advance, the need for devices that integrate AI models while optimizing energy usage is growing. Consequently, alternative computing approaches such as analog computation are being investigated to address this demand [7], [8].

Analog computation is particularly well-suited for low-energy and low-latency applications [9], [10], as it allows for energy-efficient Vector Matrix Multiplication (VMM) through Kirchoff's Current Law (KCL) and Ohm's Law. Additionally, In-Memory Computing (IMC) systems utilizing non-conventional memory technologies, such as analog memory or RRAM, can potentially store entire models on-chip [11], [12], eliminating the need for off-chip memory accesses. However, the use of these non-conventional memory technologies introduce new issues that must be taken into account such as CMOS compatibility, durability, sneak path [13], and variation [14] for successful application to the field of DNNs.

The authors are with the Electrical and Computer Engineering Department, Wayne State University, Detroit, USA (email: ev7854@wayne.edu, nabil.sarhan@wayne.edu, alhawari@wayne.edu).

Corresponding author is Melvin D. Edwards II.

This work was supported by the National Science Foundation (NSF) under grant number 2221753.

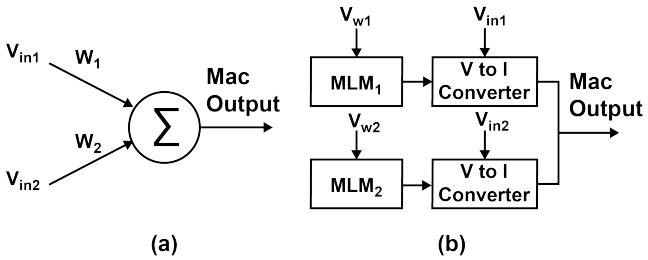

Fig. 1. Perceptron architecture. (a) High-level architecture. (b) Architecture of the proposed perceptron.

The signed MAC operation in the crossbar architecture is generally accomplished with one of two methods. The first method involves performing the signed operation digitally and assigning a negative sign in the digital domain to certain columns. This approach is referred to as the dual array approach to signed MAC operations. The second method, known as the dual row approach, which is implemented in this work, relies on a bipolar supply to perform signed operations in the analog domain. Although the dual array approach can reduce the latency of performing more operations in the analog domain, it is more sensitive to variation in the memory technology.

This paper presents a new design for an analog neuron circuit that utilizes a two-quadrant operation. The proposed circuit uses a recently introduced analog memory technology to store weights [15]–[17]. This memory has been shown to have low-power consumption and robustness to process variations (PVT).

The contribution of this paper is as follows:

- A low-power, analog perceptron with an all-CMOS implementation.
- Dual row implementation of the signed MAC operation.
- Usage of the same analog memory for both the positive and negative weights.
- Simulation results showcasing the operation of the analog neuron.

The remainder of this paper is structured as follows. Section II discusses the perceptron circuit design and the I/O signals from the perceptron. Section III presents the analytic results of the proposed neuron. Section IV discusses the simulation results of the proposed neuron. Section V concludes the paper.

II. Proposed Analog Neuron Circuit Design

The focus of this work is on inference operations. The overall perceptron architecture is shown in Figure 1. This work utilizes a $\pm 1.2V$ supply. The architecture of this circuit is modular and extra input-weight pairs can be added as needed for any particular layer in the overall DNN by adding another Multi-Level Memory (MLM) and V-to-I converter as shown in Figure 1(b). Each input-weight pair consists of an Analog Multi-Level Memory (MLM) and a current summing architecture (V to I Converter). The current summer is connected to the shared MAC node which could be fed into a trans-impedance amplifier in system-level implementations of this neuron.

A. Analog Multi-Level Memory

Figure 2 shows the analog memory in this work, Figure 2(a), takes an analog input and quantizes that input to one of eight distinct memory levels which produces a stable voltage that drives the current summer. The analog MLM has two input signals and one output signal as shown in Figure 2(b). This memory is described in detail in [15]–[17].

B. Neuron Architecture

As shown in Figure 3, the neuron uses a bipolar supply $+V_{in}$ and $-V_{in}$ to generate the signed MAC current. The MAC node is grounded. This circuit has two pairs of transmission gates: TG1/TG2 which control zero versus non-zero weights and TG3/TG4 which control the sign of the MAC operation. There is a read transistor, M_{read}, which controls the V-to-I operation. There is a current mirroring structure that allows the current to be steered into the MAC node or out of the MAC node depending on the state of the V_{sel} pin.

The circuit in Figure 3 has two inputs $\pm V_{in}$ and $V_{w,q}$. $\pm V_{in}$ is the input activations in 1-bit bipolar binary form and $V_{w,q}$ is the weight which is coming from the Analog Multi-Level Memory. TG1 and TG2 control a zero weight versus non-zero weight. If V_{zs} is HIGH, TG1 is on and TG2 is off, voltage $V_{w,q}$ is allowed to through to the gate of M_{read} for a non-zero weight. If V_{zs} is LOW, TG1 is off, TG2 is on, the gate of M_{read} is connected to V_{in}, effectively making the gate-source voltage of M_{read} zero resulting in a zero weight.

The transistor M_{read} acts as a current source for the current steering architecture. M_{mirror} acts as a mirror for both M_1 and M_2. M_1 is the current source for the negative MAC current path while M_2 is the current source for the positive MAC current path. The positive MAC current path includes M_2, TG4, M_3, and M_4. The negative MAC current path includes M_1 and TG3. The positive MAC current path is enabled when TG3 is on and TG4 is off, this occurs when V_{sel} is HIGH. The negative MAC current path is enabled when TG3 is off and TG4 is on, this occurs when V_{sel} is LOW. M_3 and M_4 are added to allow for current to be sourced by M_4 for the positive MAC while M_1 acts as a current sink for the negative MAC.

Matching between positive and negative currents is a key design parameter. The unsuitability of NMOS and PMOS transistors for the roles of current sinking and sourcing, respectively, arises from their opposite response to gate voltage modulation. Specifically, the overdrive of a PMOS transistor exhibits a decreasing trend as the gate voltage is increased. In contrast, an NMOS transistor demonstrates an increasing trend as the gate voltage is increased. Ideally, the current will be equal in magnitude and have opposite signs for all voltages of $V_{w,q}$. There is some mismatch in the currents from the positive and negative MAC operation. This non-ideality comes from the difference in current gain from the positive and negative current sources in the architecture. More details will be provided in the next section.

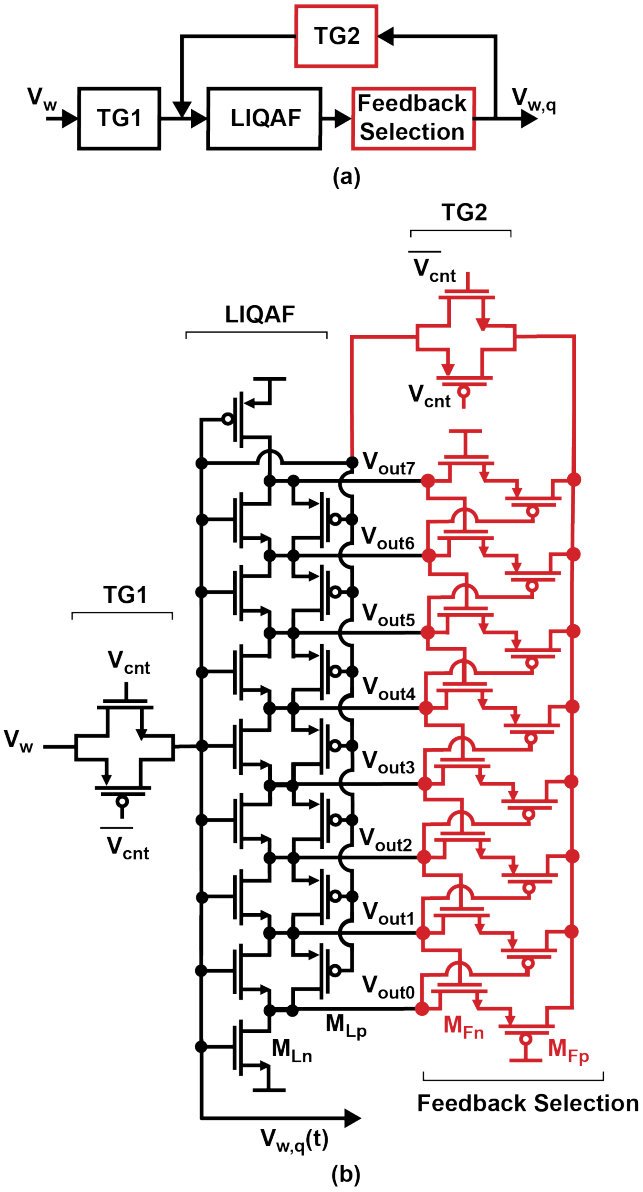

Fig. 2. (a). High-level block diagram of the analog memory. (b). transistor-level schematic of the analog memory.

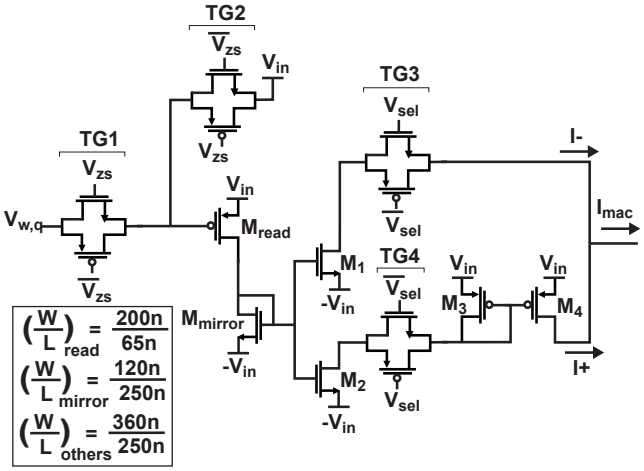

Fig. 3. Current Summing Architecture.

III. Analytic Results

The read transistor, M_{read}, converts the weight voltage into a current. The read transistor is biased to operate in the saturation region over the range of voltages that the analog multi-level memory outputs. The transconductance of the read transistor roughly follows that of the following equation.

$$g_m = \mu_n C_{ox} \frac{W}{L}(V_{SG} - |V_{th}|) \tag{1}$$

The read transistor has a different gain for different weight values due to the over-drive voltage being modulated by the output of the analog memory's output. This is not an issue for two reasons. First, the architecture does not change weight values in run-time, weights are configured at the beginning of operation and remain until new weights are loaded into the inference machine. Second, this architecture is designed for 1-bit input activation which results in a discrete output of eight different levels or an output of zero in the case of zero input or zero weight.

The read transistor forms a current source for the central biasing branch of the current steering architecture shown in Figure 3. The read transistor expresses differences in weight voltages as a change in the over-drive voltage of the read transistor. Due to velocity saturation, there is a linear relationship between gate-source voltage and drain current in the saturation region. This results in a variable current source that is dependent on the weight voltage while the input voltage is binary. The current in the read transistor roughly follows the following equation.

$$I_D = \frac{1}{2}\mu_n C_{ox} \frac{W}{L}(V_{SG} - |V_{th}|)^2 \tag{2}$$

The read transistor, M_{read}, is a current source to the current mirroring structure. The current mirroring structure takes this current and replicates it in the two signed branches connected to the drains of M_1 and M_2. Equation 2 explains the downward trend in the MAC current as the weight voltage is increased. As the weight voltage decreases, the voltage V_{SG} increases since the source is fixed at V_{DD}. As the gate voltage decreases the over-drive increases which results in a larger current at the central biasing branch.

The current at the signed branches are roughly the following equations for transistors M_1 and M_2.

$$I_- = \frac{W_1}{W_{mirror}} I_{read} \tag{3}$$

$$I_{M2} = \frac{W_2}{W_{mirror}} I_{read} \tag{4}$$

In equation 4, I_{read} is being generated from M_{read}. I_- is being fed directly to the MAC node since it is sinking current from the node. The positive MAC node must undergo another conversion to convert from sinking current to sourcing current into the node. This equation is defined by the equation shown below.

$$I_+ = \frac{W_2 W_4}{W_{mirror} W_3} I_{read} \tag{5}$$

Based on equations 3 and 5 sizing of transistors W_1 and W_2 must be equal and transistors W_3 and W_4 must be equal to ensure matching between the negative and positive MAC operation. Another consideration is channel length modulation which needs to be taken into account in design with cascoding techniques or increasing the transistor channel length. The length for this design was chosen to be 250 nm to minimize the effect of channel modulation on the mirroring operations.

IV. Simulation Results

This section goes over the simulation results of the dual row MAC operation. Figure 4 shows the MAC current when $V_{in,1}$ is positive and $V_{in,2}$ is zero. The maximum current is 56 μA and the minimum current is 25 μA. The stairwell characteristic comes from the analog memory which is driving the current steering architecture. The analog memory outputs a voltage in the 0 to V_{DD} range with the analog outputs centered on $0.5V_{DD}$.

Figure 5 shows the output characteristic stairwell when both $V_{in,1}$ and $V_{in,2}$ are positive. The current summing action of the MAC architecture can be seen in this figure, where the maximum MAC current is 112 μA and the minimum current is 50 μA. This is due to both the inputs being positive and having the same weight applied to them. In actual operation, the two inputs do not need to have the same weight as each input/weight pair has a different analog memory to drive the current steering architecture.

Figure 6 depicts the MAC current when $V_{in,1}$ is negative and $V_{in,2}$ is zero. The maximum current is -25 μA and the minimum current is -58 μA. It should be noted that Figures 4 and 6 should have equal magnitude and opposite sign if perfect matching between the positive and negative path is accomplished based on section III. The difference between Figures 4 and 6 leads to Figure 7. Figure 7 quantifies the mismatch between the positive and negative paths. The maximum mismatch is at the two endpoint weight values which corresponds to 1.5 μA in each case. The mismatch of 1.5 μA is very small compared to the minimum non-zero

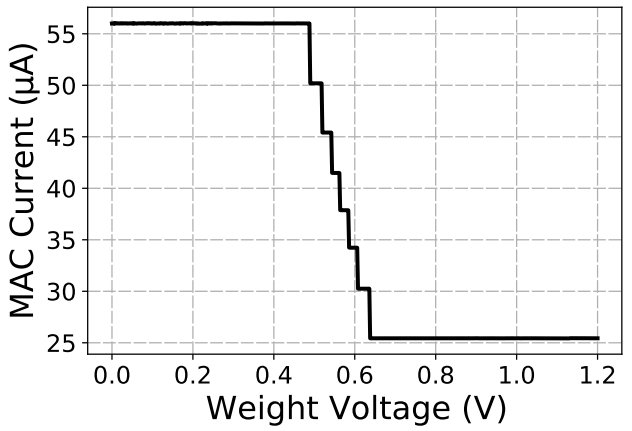

Fig. 4. Output MAC Currents when $V_{in,1}$ is positive and $V_{in,2}$ is zero.

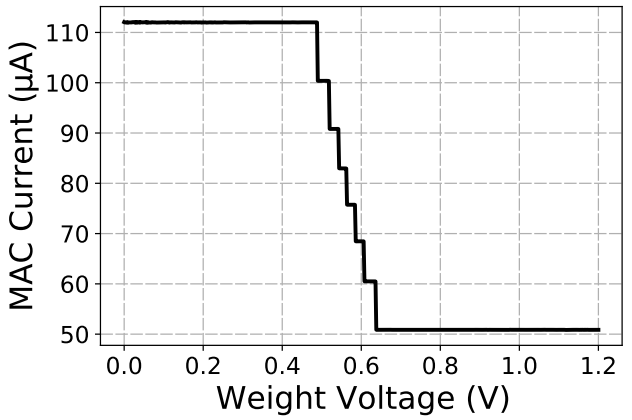

Fig. 5. Output MAC Currents when $V_{in,1}$ is positive and $V_{in,2}$ is positive.

current of 25 μA. This can effectively be considered zero in downstream operations where a trans-impedance amplifier (TIA) would amplify the current signal generated by this stage.

Figure 8 represents the actual MAC current in the time domain as a weight is programmed into the memory and held. In Figure 8 the programming of the memory happens from 0 s to 50 μs, this is outside the scope of this paper so the programming stage will not be discussed as the focus is on the inference stage of the MAC operation. Once the weight is chosen and the input voltage is input to the current summing circuit, the current is held indefinitely. The input to this architecture is one-bit, so it is either zero or a non-zero value determined by the analog memory.

Each input-weight pair consumes 170μW. The circuit shown in this work consumes a maximum of 340μW when both input-weight pairs are active.

V. Conclusion

The modular nature of the proposed analog neuron circuit enables the creation of neural networks with layers of varying

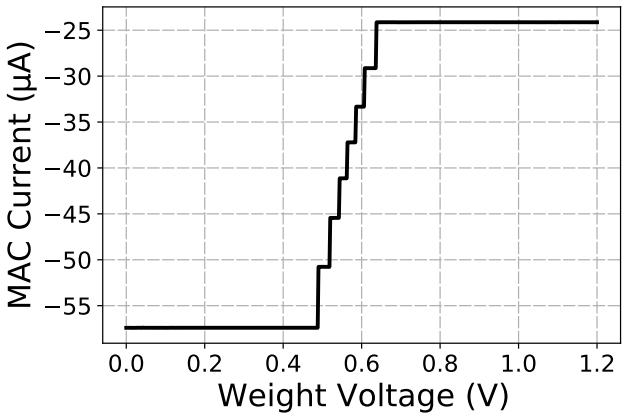

Fig. 6. Output MAC Currents when $V_{in,1}$ is negative and $V_{in,2}$ is zero.

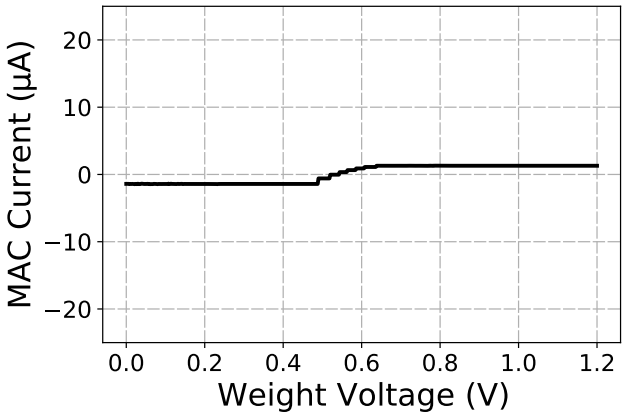

Fig. 7. Output MAC Currents when $V_{in,1}$ is positive and $V_{in,2}$ is negative with equal magnitude weights on both inputs.

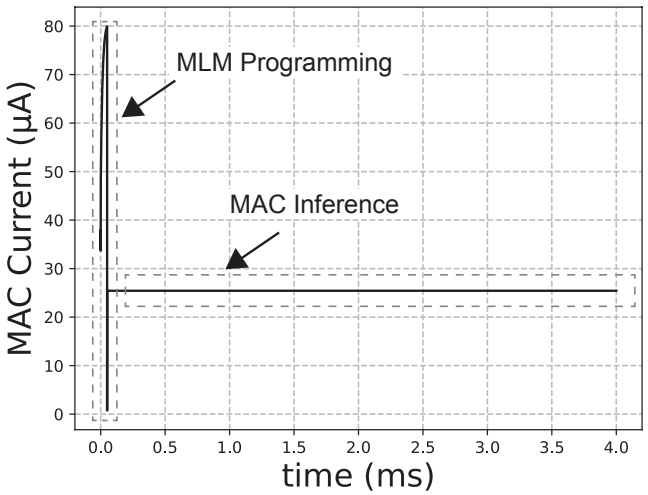

Fig. 8. Output MAC Current when $V_{in,1}$ is positive and $V_{in,2}$ is zero. A 1.2V weight is applied.

sizes. If a larger or smaller neural network is needed, input-weight pairs can be added or removed from the MAC node

in the same manner as the first two input-weight pairs shown in this work. This level of flexibility enables this architecture to be used in neural networks that require varying numbers of connections for neurons, depending on the specific layer of the network. The modular and flexible nature of this block is a key feature that makes it an extremely versatile and valuable tool for building neural networks of all types and sizes.

REFERENCES

[1] Y. Liu, J. He, J. Gu, X. Kong, Y. Qiao, and C. Dong, "Degae: A new pretraining paradigm for low-level vision," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23292–23303, 2023.

[2] F. Sammani, T. Mukherjee, and N. Deligiannis, "Nlx-gpt: A model for natural language explanations in vision and vision-language tasks," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8312–8322, 2022.

[3] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12176–12185, 2022.

[4] A. Soni, B. Amrhein, M. Baucum, E. J. Paek, and A. Khojandi, "Using verb fluency, natural language processing, and machine learning to detect alzheimer's disease," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pp. 2282–2285, 2021.

[5] M. Guo, Y. Chen, J. Xu, and Y. Zhang, "Dynamic knowledge integration for natural language inference," in *2022 4th International Conference on Natural Language Processing (ICNLP)*, pp. 360–364, 2022.

[6] S. Das, M. Ashrafuzzaman, F. T. Sheldon, and S. Shiva, "Network intrusion detection using natural language processing and ensemble machine learning," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 829–835, 2020.

[7] H. Tsai, P. Narayanan, S. Jain, S. Ambrogio, K. Hosokawa, M. Ishii, C. Mackin, C.-T. Chen, A. Okazaki, A. Nomura, I. Boybat, R. Muralidhar, M. M. Frank, T. Yasuda, A. Friz, Y. Kohda, A. Chen, A. Fasoli, M. J. Rasch, S. Woźniak, J. Luquin, V. Narayanan, and G. W. Burr, "Architectures and circuits for analog-memory-based hardware accelerators for deep neural networks (invited)," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2023.

[8] N. Udayanga, S. I. Hariharan, S. Mandal, L. Belostotski, L. T. Bruton, and A. Madanayake, "Continuous-time algorithms for solving maxwell's equations using analog circuits," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–1, 2020.

[9] M. de Prado, M. Rusci, R. Donze, A. Capotondi, S. Monnerat, L. Benini, and N. Pazos, "Robustifying the deployment of tinyml models for autonomous mini-vehicles," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2021.

[10] K. Matsubara, L. Hanno, M. Kimura, A. Nakamura, M. Koike, K. Terashima, S. Morikawa, Y. Hotta, T. Irita, S. Mochizuki, H. Hamasaki, and T. Kamei, "4.2 a 12nm autonomous-driving processor with 60.4tops, 13.8tops/w cnn executed by task-separated asil d control," in *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, vol. 64, pp. 56–58, 2021.

[11] Y. He, Y. Huang, J. Yue, W. Sun, L. Zhang, and Y. Liu, "C-rram: A fully input parallel charge-domain rram-based computing-in-memory design with high tolerance for rram variations," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 3279–3283, 2022.

[12] J. Read, W. Li, and S. Yu, "Enabling long-term robustness in rram-based compute-in-memory edge devices," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2023.

[13] S. K. Kingra, V. Parmar, S. Negi, S. Khan, B. Hudec, T.-H. Hou, and M. Suri, "Methodology for realizing vmm with binary rram arrays: Experimental demonstration of binarized-adaline using oxram crossbar," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2020.

[14] J. Yang, X. Xue, X. Xu, Q. Wang, H. Jiang, J. Yu, D. Dong, F. Zhang, H. Lv, and M. Liu, "24.2 a 14nm-finfet 1mb embedded 1t1r rram with a 0.022μm2 cell size using self-adaptive delayed termination and multi-cell reference," in *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, vol. 64, pp. 336–338, 2021.

[15] M. Alhawari, N. Albelooshi, and M. H. Perrott, "A 0.5v ¡4μw cmos photoplethysmographic heart-rate sensor ic based on a non-uniform quantizer," in *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, pp. 384–385, 2013.

[16] M. Alhawari, N. A. Albelooshi, and M. H. Perrott, "A 0.5 v $<$ 4 μw cmos light-to-digital converter based on a nonuniform quantizer for a photoplethysmographic heart-rate sensor," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 271–288, 2014.

[17] M. D. Edwards, H. Al Maharmeh, N. J. Sarhan, M. Ismail, and M. Alhawari, "A low-power, digitally-controlled, multi-stable, cmos analog memory circuit," in *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 872–875, 2020.