Monga, Dipesh C.; Numan, Omar; Andraud, Martin; Halonen, Kari

A temperature and process compensation circuit for resistive-based in-memory computing arrays

Published in:
ISCAS 2023 - 56th IEEE International Symposium on Circuits and Systems, Proceedings

DOI:
10.1109/ISCAS46773.2023.10181619

Published: 01/01/2023

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Monga, D. C., Numan, O., Andraud, M., & Halonen, K. (2023). A temperature and process compensation circuit for resistive-based in-memory computing arrays. In *ISCAS 2023 - 56th IEEE International Symposium on Circuits and Systems, Proceedings* (IEEE International Symposium on Circuits and Systems proceedings; Vol. 2023-May). IEEE. https://doi.org/10.1109/ISCAS46773.2023.10181619

A temperature and process compensation circuit for resistive-based In-Memory Computing arrays

Dipesh C. Monga, Omar Numan, Martin Andraud and Kari Halonen
Department of Electronics and Nanoengineering
Aalto University School of Electrical Engineering, Espoo, Finland
Email: firstname.lastname@aalto.fi

***Abstract*—In-Memory Computing (IMC) architectures promise increased energy-efficiency for Artificial intelligence (AI) accelerators. IMC circuits are typically based on analog computation, which is more sensitive to process and temperature variations than digital. Thus, IMC circuit elements may require process and temperature compensation to maintain suitable computation accuracy and linearity. In this work, we propose an ultra-low power circuit to compensate for the temperature and process-based non-linearity of resistive computing elements (i.e., memristors) in IMC arrays. The proposed circuit has been implemented in a 65nm CMOS bulk technology. It can provide a temperature coefficient between 10 and 1938 ppm/°C for a wide temperature range (-40°C to 80°C) and output current range (few pA up to 600 nA) at 1.2 V operating voltage. Used in a resistive IMC array, the variation of output currents from each synapse (column accumulation) can be reduced by up to 84% to maintain computation accuracy across process and temperature variations.**

***Index Terms*—Thermal compensation, process compensation, ultra-low power, variable temperature coefficient, In-memory computing, Resistive random access memory.**

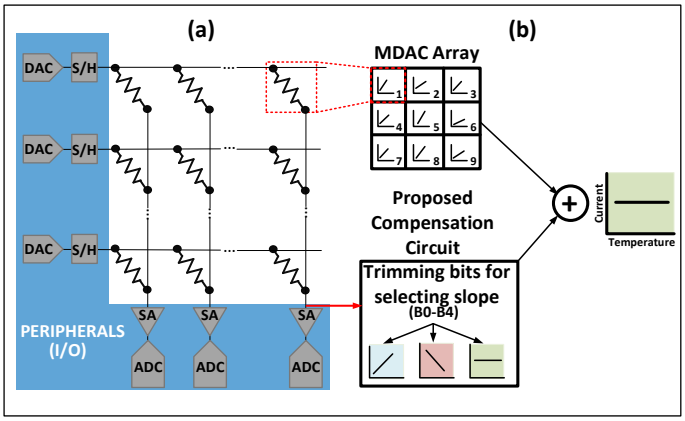

Fig. 1. (a) Illustration of a general resistive-based IMC core; (b) illustration of the proposed compensation cell.

I. Introduction

The recent push for embedded Artificial Intelligence (AI) computation has raised the need for dedicated analog and mixed-signal hardware AI processors [1]. The computation of AI models, specifically Neural Networks (NNs), mainly consists of simple and repetitive operations (products or sums). Using Von Neumann processors, memory transfers dominate the power consumption as memory and calculation units are separated. In-Memory Computing (IMC), which intend to eliminate this bottleneck by performing the computation directly inside the memory for increased efficiency [2], has recently gained significant interest [3]. IMC relies on analog computation, and a typical architecture using resistors is illustrated in Fig. 1(a). Inputs are analog voltages converted from a DAC. They are multiplied by their respective synaptic weights stored as conductance. Using Ohm's law, each multiplication is accumulated as a current column-wise, summed through a summing amplifier (SA) and converted to digital with an ADC. As analog computations are more sensitive to process and temperature variations, a calibration circuitry may be necessary for optimal performance. Yet, compensation techniques have not been widely investigated. For instance, [4] uses process variation calibration based on an operational amplifier, yet the exact architecture is not described. Here, we focus on resistive-based IMC arrays. These arrays rely, for instance, on emerging non-volatile memories (eNVM), such as Resistive or Magnetic Random Access Memories (ReRAM, MRAM) [5], [6]. eNVMs can store weights directly as analog values, promising more efficient computation and better integration. Alternatively, structures combining poly-silicon resistors and SRAM as R-2R Multiplying DACs (MDACs) have been proposed [7], providing excellent compatibility with standard CMOS processes. Yet several shortcomings limit the performance of resistive-based IMC arrays, such as the non-linearities of the resistive elements, read/write noises, and robustness [8]. As a result, eNVM-based arrays are still lacking behind SRAM-based arrays, in terms of area and efficiency [9]. Process and thermal stability represent severe limitations to the computing accuracy, causing significant resistance changes, whether an eNVM [10] or an MDAC [11]. This non-linearity accumulates over the computations reducing the inference accuracy [5], [10].

To solve this challenge, we propose a low-power and small-area process and temperature compensation technique for resistive-based IMC arrays. The proposed circuit compensates for the variations in the accumulated current of each IMC column to minimize computing accuracy losses. We demonstrate the effectiveness of the proposed approach using an R-2R MDAC weight synapse [7], [12]. As shown in Fig. 1(b) the compensation circuit, based on a beta multiplier current reference, produces a current whose thermal slopes can be varied

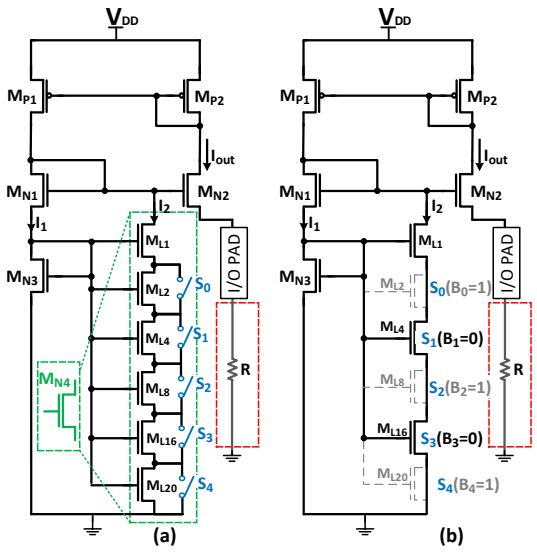

Fig. 2. (a) Implemented schematic of the proposed variable TC current generation with off-chip resistor R, (b) Circuit behaviour when the switches M_{S0}-M_{S4} are 10101.

by trimming bits. It results in a low-temperature coefficient (TC) output current, compensating the current degradation over a wide range of temperatures and process corners while preserving the computation accuracy. The efficiency of the circuit, fabricated and measured in a 65nm-CMOS process, is demonstrated on a MDAC cell.

II. Proposed Circuit Description

A. Variable TC current generator

Figure 2(a) depicts the proposed circuit, which consists of current generator that can generate complementary to absolute temperature (CTAT), proportional to absolute temperature (PTAT), or temperature invariant currents. The circuit is a modification of [13], where an additional trimming sub-circuit is added to control the slope of the output current w.r.t. to process and temperature variations. The aspect ratio of the PMOS M_{P1} and M_{P2} is 1:1, and the aspect ratio (or current mirroring ratio) of M_{N1} to M_{N2} is 1:8. The circuit uses a temperature dependence control sub-circuit consisting of NMOS transistors of equal widths, M_{N4}, M_{L1}, M_{L2}, M_{L4}, M_{L8}, M_{L16}, M_{L20} and switches S_0-S_4. Switches S_0-S_4 are NMOS with minimum ratios for low-resistance controlled by trimming bits B_0-B_4, respectively. Subscripts denote the ratio of length of transistors M_{L1}, M_{L2}, M_{L4}, M_{L8}, M_{L16}, M_{L20}. For example, if the length of NMOS M_{L1} is 1 µm, then the length M_{L16} is 16 µm. An off-chip biasing high-precision resistor R controls the range of biasing currents. The absence of on-chip resistor decreases the area overhead, reduces process variations and mismatches and eliminates the need for post-silicon trimming of resistors. It can be noted that the capacitance of I/O pads due to bonding has no effect on the output DC current.

Figure 2(b) shows the circuit operation for a trimming bit code B_0 to B_4 on S_0-S_4. Switches are closed for M_{L2}, M_{L8}, M_{L20}, indicating that then the drain and source of these transistor are connected, resulting in V_{DS}=0. Switches are open for M_{L4}, M_{L16}, which become part of right branch. Since the gates of M_{L2}-M_{L20} are connected together, they can be denoted by their equivalent series transistor M_{N4} as:

$$S_{N4} = \frac{W}{L_1 + \overline{B_0}.L_2 + \overline{B_1}.L_4 + \overline{B_2}.L_8 + \overline{B_3}.L_{16} + \overline{B_4}.L_{20}} \quad (1)$$

where S_{N4} is the equivalent aspect ratio of transistor M_{N4}, W is the width of transistor M_{L2}-M_{L20}, and L_2, L_4, L_8, L_{16}, L_{20} denote the length of respective transistors. Equation (1) shows that the aspect ratio of the equivalent transistor M_{N4} can effectively be controlled by the trimming bits. The output current is given as [13]:

$$I_{out} = \frac{V_{TH}}{R} + \frac{1}{R^2 K_n (S_{N3} + S_{N4})} - \sqrt{\frac{V_{TH}^2}{R^2} + \frac{2.V_{TH}}{R^3 K_n (S_{N3} + S_{N4})} + \frac{1}{(R K_n (S_{N3} + S_{N4}))^2} - V_{TH}^2} \quad (2)$$

where K_n is the transconductance parameter, V_{TH} is the threshold voltage, R is the off-chip resistor, S_{N3} and S_{N4} are the aspect ratios of M_{N3} and M_{N4}. In equation (2), V_{TH} and R are both temperature dependent, and V_{TH} and K_n are process-dependent. These dependencies can be compensated by changing the magnitude and behaviour of the output current using the trimming bits to adjust M_{N4} according to (1).

B. Weight synapse as a Multiplying DAC

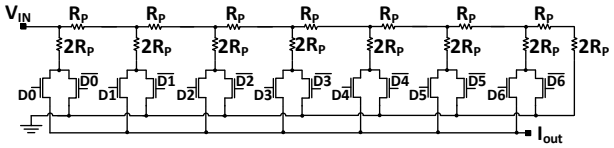

Fig. 3. Schematic of the MDAC used in synapse circuits for CNN.

Figure 3 depicts the schematic of the resistor-based R-2R MDAC used as a weight synapse, similar to [7], [12]. The circuit forms a 7-bit binary-weighted multiplier, based on a poly-silicon resistor-based R-2R MDAC with a unit resistance of R_P. This topology could, for instance be, used to build competitive weight synapses with high-density resistor technologies [14], [15], where MΩ resistor values can be obtained on areas close to foundry-based SRAM cells in modern CMOS processes. The output current I_{out} is written as:

$$I_{out} = \frac{V_{IN}}{R_P} \cdot \frac{\sum_{i=0}^{i=6} 2^i \cdot D_i}{2^7} \quad (3)$$

A 7-bit MDAC can produce (2^7=128) different I_{out} levels. As shown by equation (3), I_{out} is proportional to the product of V_{IN} and the binary-weighted conductance of R_P. As shown in Fig.1, in this work we built and analyzed an array of 3x3 MDAC synapses to test the performance of the current temperature and process compensation circuit. It represents a 3x3 convolution baseline, between a weight kernel and input data in a convolutional NN [4]. After multiplication operation of the weights kernel and input data, the outputs of 9 MDACs are summed together through a wire. This forms

the 3x3 convolution output which is then combined with a complimentary thermal slope current of constant magnitude (for all weights) to eliminate the temperature effect over the MDAC output current.

R-2R MDACs built with poly-silicon resistors are susceptible to output errors when performing in different temperatures [11]. The relationship between resistance, R_P, and the TC of resistance, TCR, can be described as:

$$R_p(T) = \alpha_2 \cdot T^2 + \alpha_1 \cdot T + \alpha_0 \tag{4}$$

and

$$TCR(T) = \frac{dR_P(T)}{dT} = \alpha_0(\frac{2\alpha_2}{\alpha_0}T + \frac{\alpha_1}{\alpha_0}) \tag{5}$$

where T is the temperature, and α_0, α_1, α_2 are constants [11].

As seen from equation (5), the thermal dependence of the poly-silicon resistor is PTAT; hence with the variation of temperature, the output current increases as R_p is increased with temperature. Variations in the unit resistance due to temperature and process variations can cause changes in the MDAC output currents resulting in degradation in the accuracy of the IMC multiplication operations.

The multiplication precision of the IMC multipliers must stay consistent over these variations to maintain the NN accuracy. In addition, it is a good practice to ensure that the output inaccuracy of the MDAC does not exceed ½ LSB, where 1 LSB defines the width of one step of an N-bit MDAC and is given by:

$$1\,LSB = \frac{FSR}{2^N} \tag{6}$$

Here, FSR is the MDAC full-scale range, in this case of current-mode MDAC FSR is maximum I_{out} achieved when all MDAC cells have full-scale weight, and largest V_{IN}.

III. Results and Discussion

The proposed process and temperature compensation current generator circuit has been fabricated in a 65nm CMOS technology. The MDAC array performance has been evaluated with post-layout simulations to demonstrate the application of the fabricated circuit.

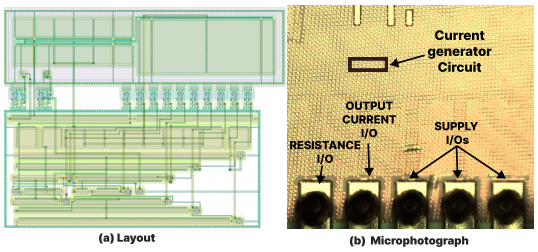

Fig. 4. (a) Layout (b) Die microphotograph of the TC and process compensation current generator.

Figures 4(a) and (b) show the layout and photograph of the implemented circuit. The circuit occupies an active area of 540 μm^2 (20 µm x 27 µm). The final area overhead depends on how many elements the IMC array has for each column and the range of the output current.

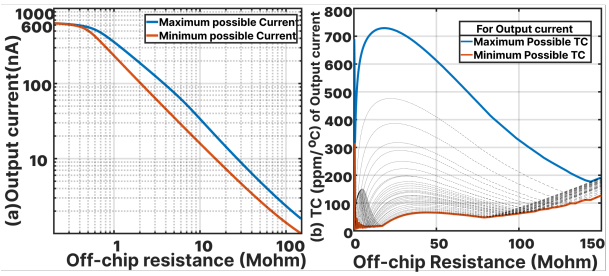

Fig. 5. Maximum and minimum achievable (a) Current (b) TC, by changing the off-chip resistance when trimming bits are varied.

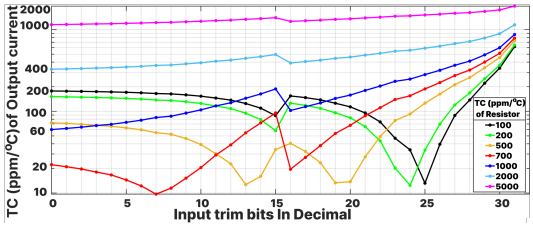

Fig. 6. Simulated thermal dependence of the output current with the variation of the trimming bits for various TC resistors.

A. Evaluation of Variable TC current generator

The temperature measurements are performed in a Weisstechnik Labevent temperature chamber for a temperature range of -40°C to 80°C where the device under test is soldered to a PCB along with the off-chip resistor and kept inside the chamber. The output current is measured using a Keithley 6514 electrometer. The trimming bits are controlled by programming on chip SPI using a Xilinx Zynq-7000 SoC ZC706 FPGA. Figure 5(a) depicts the change of the output current value, which can be varied from a few pA to 600 nA by changing the off-chip resistance. Moreover, changing the trimming bits can achieve the desired TC and process spread. Figure 5(b) plots an example with the maximum and minimum achievable TC of the output current. Figure 5(a) and Figure 5(b) show the envelop of TC (i.e., the maximum and minimum TC values). Intermediate values can also be achieved but are not discussed in this plot. To evaluate the thermal dependence of the output current, we performed post-layout simulation with different TC values of the $5M\Omega$ off-chip resistor. The circuit consumes a power of $\approx 80nW$ (1.2V supply, room temperature) for this setup. Figure 6 shows the TC of output current by varying the trimming bits B_0-B_4, where B_0 is the MSB and B_4 is the LSB. The circuit produces a current with TC as low as 10 ppm/°C for off-chip resistances with TC of up to 700 ppm/°C. For resistance with TC up to 5000 ppm/°C the circuit achieves an output current with a maximum TC of 1938 ppm/°C. Figure 7(a) shows measured values of output current when the trimming bits are changed at room temperature, validating the circuit's functionality. Fig.7(b) shows the distribution of the output current measured for 12 samples at room temperature and 1.2 V supply. The measured samples shows mean current of 42.17 nA with a process variation ($3\sigma/\mu$) of 4.2%. Figure 8(a) shows the measured values relative current (i.e., the output current subtracted from the average current value) to show the thermal sensitivity of the output current. As verified from measurements, the circuit can generate PTAT, CTAT and

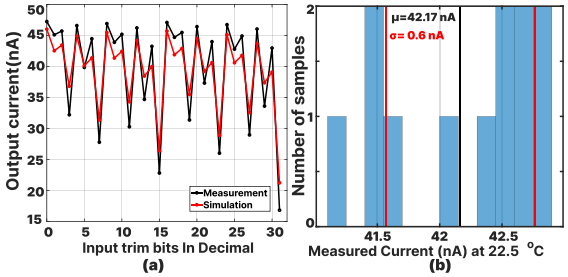

Fig. 7. (a) Measured and simulated output current when trimming bits are varied. (b) Histogram of the output current of measured samples for a fixed trimming bit at room temperature.

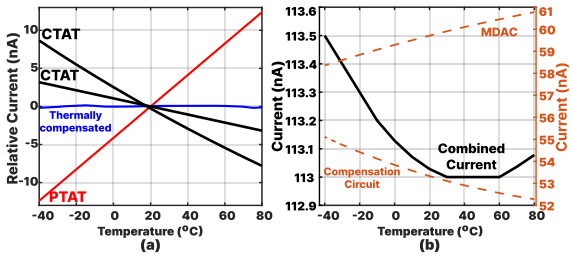

Fig. 8. (a) Measured thermal variation and a few possible output currents (b) Output current of MDACs array and compensation circuit (orange) and the whole system (black).

thermally compensated currents.

B. Performance of the system with an MDAC synapse

Post-layout simulations have been performed to evaluate the performance of the MDAC synapse with the compensation circuit. The system is assessed by adding the measured output current of the compensating circuit with post-layout simulation currents of the synapse.

Figure 8(b) shows the output current of the MDAC array, the variable TC generator, and the whole system. As discussed in section II, the output of the MDAC array is PTAT. Thus, the variable TC circuit is set to generate CTAT current so that the overall current is thermally compensated. The TC of MDAC current is 335 ppm/°C, and the overall TC of compensated current is 70 ppm/°C at 1.2V supply.

The discussion in Section III-B indicates that the reduction in thermal and process dependence of the current increases the accuracy of the synapse multiplication operation. In the simulated synapse network, the FSR of the output current in this configuration is 60 nA and the 1 LSB for the 7-bit MDAC is 0.47 nA (see equation(6)). Figure 8(b) shows that the variation of the synapse output current is 3 nA over a range of 120°C temperature range. Hence, the accuracy error in the output current is 6.4 LSB. On the other hand, the accuracy error in the combined current is 1 LSB which is an 84% improvement compared to the non-compensated version.

To test process compensation, Fig. 9(a) shows post-layout simulations for different process corners for a fixed trimming bit value of 11. Figure 9(b) shows the process spread when the trimming circuit is used with a focus on achieving constant output current across corners. Here the average current error at room temperate across the corners is reduced from 5.76 nA to 1.72 nA (70%).

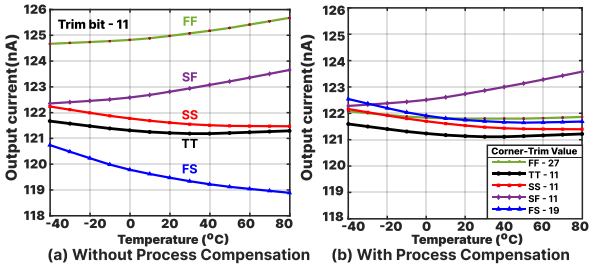

Fig. 9. Process spread of the MDACs synapse system for constant average output current (a) Without trimming bit compensation (b) With trimming bit compensation.

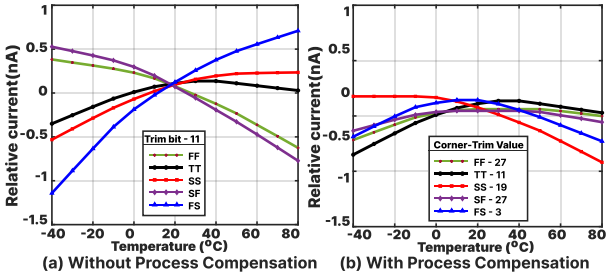

Fig. 10. Process variation of the MDACs synapse system (a) Without trimming bit compensation (b) With trimming bit compensation to improve the thermal sensitivity.

Another process compensation is discussed in Figure 10, where the trimming circuit is used to improve the thermal stability of the current across process corners. In Figure 10(a), the relative current (output current subtracted from its average value) at a fixed trimming bit is shown. Figure 10(b) shows the thermal sensitivity improvement the trimming circuit is used for TC compensation. Here the process compensation has improved the maximum TC variation from 125 ppm/°C to a maximum of 43 ppm/°C which is an improvement of 65%. Figure 11(a) shows the thermal behaviour of the input voltage

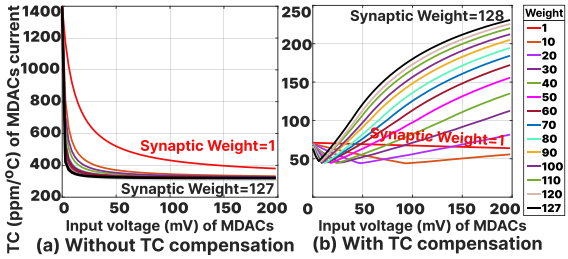

Fig. 11. Variation of input voltage of MDACs synapse (a) without TC compensation (b) with TC compensation.

V_{IN} of the synapse, where the TC of output current ranges from 1400 ppm/°C to 400 ppm/°C, depending on the input voltage and synaptic weights. The combined currents of the MDAC array and the compensation circuit reduce the TC to a minimum of 44 ppm/°C and a maximum of 230 ppm/°C, as shown in Figure 11(b).

IV. Conclusions

This work presents a temperature and process compensation circuit, compensating for resistor non-linearities in resistive-based IMC arrays. The circuit is implemented in 65nm bulk

CMOS. The design can generate current ranging 600 nA to 5 pA with TC varying from 1938 ppm/°C to 10 ppm/°C. The circuit is demonstrated on a MDAC weight synapse circuit, which maintains computation accuracy across process and temperature variations.

ACKNOWLEDGMENTS

This work was supported by Academy of Finland through the projects EHIR (grant XXXXX) WHISTLE project (grant 332218)

REFERENCES

[1] B. Murmann, "Mixed-Signal Computing for Deep Neural Network Inference," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 29, no. 1, pp. 3-13, Jan. 2021.

[2] N. Verma et al., "In-Memory Computing: Advances and Prospects," in *IEEE Solid-State Circuits Magazine*, vol. 11, no. 3, pp. 43-55, Summer 2019.

[3] J. -s. Seo et al., "Digital Versus Analog Artificial Intelligence Accelerators: Advances, trends, and emerging designs," in IEEE Solid-State Circuits Magazine, vol. 14, no. 3, pp. 65-79, Summer 2022.

[4] J. -O. Seo, M. Seok and S. Cho, "ARCHON: A 332.7TOPS/W 5b Variation-Tolerant Analog CNN Processor Featuring Analog Neuronal Computation Unit and Analog Memory," *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 258-260, 2022.

[5] J. M. Correll et al., "A Fully Integrated Reprogrammable CMOS-RRAM Compute-in-Memory Coprocessor for Neuromorphic Applications," *in IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, pp. 36-44, June 2020.

[6] L. Chang et al., "Trend of Emerging Non-Volatile Memory for AI Processor," *2021 18th International SoC Design Conference (ISOCC)*, pp. 223-224, 2021.

[7] R. J. Kier, R. R. Harrison and R. D. Beer, "An MDAC synapse for analog neural networks," *2004 IEEE International Symposium on Circuits and Systems*, pp. V-V, 2004.

[8] Daniele Ielmini., "Resistive switching memories based on metal oxides: mechanisms, reliability and scaling," *Semicond. Sci. Technol.*, 2016.

[9] N. R. Shanbhag and S. K. Roy, "Comprehending In-memory Computing Trends via Proper Benchmarking," in *2022 IEEE Custom Integrated Circuits Conference (CICC)*, pp. 01–07, 2022.

[10] Heng Xu et al., "Statistical temperature coefficient distribution in analog RRAM array: impact on neuromorphic system and mitigation method," *Journal of Physics D: Applied Physics*, 2021.

[11] Chuang, Hung-Ming & Thei, Kong-Beng & Tsai, Sheng-Fu & Lu, Chun-Tsen & Liao, Xin-Da & Lee, Kuan-Ming & Chen, Hon-Rung & Liu, Wen-Chau., "A comprehensive study of polysilicon resistors for CMOS ULSI applications", *Superlattices and Microstructures*. Vol 33. pp 193-208, 2003.

[12] R. Laajimi, B. Hamdi and N. Ayari, "A low power bulk-driven MDAC synapse," *2011 International Conference on Applied Electronics*pp. 1-5, 2011.

[13] Chouhan, S.S., Halonen, K., "Ultra low power beta multiplier-based current reference circuit". *Analog Integrated Circuits and Signal Processing*,pp 523–529, 2017.

[14] W. Choi, M. Kwak, S. Heo, K. Lee, S. Lee, and H. Hwang, "Hardware Neural Network using Hybrid Synapses via Transfer Learning: WOx Nano- Resistors and TiOx RRAM Synapse for Energy-Efficient Edge-AI Sensor," in *2021 IEEE International Electron Devices Meeting (IEDM)*. IEEE, pp. 23–1, 2021.

[15] D. Saito, et al. "Analog In-memory Computing in FeFET-based 1T1R Array for Edge AI Applications," in 2021 Symposium on VLSI Circuits, pp. 1–2, 2021.

[16] B. Razavi, Design of analog CMOS integrated circuits, Boston, MA: McGraw-Hill.

[17] Huang, Shanshi Sun, Xiaoyu & Peng, Xiaochen & Jiang, Hongwu & Yu, Shimeng,"Achieving High In Situ Training Accuracy and Energy Efficiency with Analog Non-Volatile Synaptic Devices". *ACM Transactions on Design Automation of Electronic Systems*.pp 1–19, 2022.

[18] Y. Luo and S. Yu, "Benchmark Non-volatile and Volatile Memory Based Hybrid Precision Synapses for In-situ Deep Neural Network Training," *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 422-427, 2020.