Analysis of Dual-Row and Dual-Array Crossbars in Mixed Signal Deep Neural Networks

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

***Abstract*—This paper makes a comparative analysis between dual row and dual array mixed-signal neural network architectures through the use of a case study. Dual row and dual array architectures are used to implement signed operations in mixed signal neural network circuits. The dual row arrangement uses two ideal data paths for signed operations in the analog domain. The dual array arrangement uses two Analog to Digital Converters (ADCs) to perform the signed operation in the digital domain wheras the dual row arrangement uses only one. The trade-offs studied for each topology are: speed/latency, which has similar performance between both architectures. Power consumption, which if including the Multiply and Accumulate (MAC) circuit alone is much lower than the dual row approach. Area, which is lower for the dual array approach than it is for the dual row approach. Variability, which is similar for both approaches, but has limitations when considering the memory technology used.**

***Index Terms*—mixed signal, neural networks, analog memory, dual row, dual array**

I. Introduction

Neural networks have found widespread application in numerous fields including computer vision [1]–[5], natural language processing [6], [7], fraud detection [8], [9], autonomous technologies [10], [11], healthcare [12], [13], and behavioral prediction [14], [15]. As neural networks become larger and more complex there is a growing concern about power consumption, especially for edge devices that have limited energy sources. This has led to an increase in research for smaller and more efficient neural networks which include techniques to reduce the networks size such as pruning and weight dropout [16], [17], and thus have the potential to operate on the edge since they can be scaled down in size and complexity.

Neural networks can be designed using digital or mixed-signal architectures [18], [19]. Digital neural networks perform all calculations related to the vector matrix multiplication (VMM) in the digital domain [20]. In contrast, mixed-signal neural networks perform the VMM in the analog domain and convert the result to the digital domain for further processing [21]. In particular, mixed-signal neural networks outperform the digital ones in energy efficiency [22] due to the usage of memory technologies that enables in-memory computations and thus substantially reduce the energy of data transfer.

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.

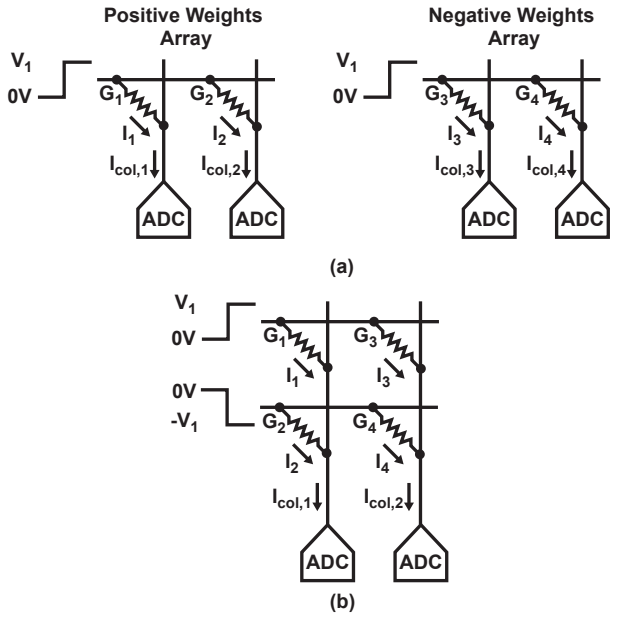

Fig. 1. A high-level schematic of (a) Generic dual array and (b) generic dual row.

A key component in mixed-signal neural networks is the memory to store the values of the weights. Analog memory technologies that provide in-memory computation include embedded flash memory, Resistive RAM (RRAM), Phase-Change RAM (PCRAM), and Magnetic RAM (MRAM). Other memory technologies can be used to implement near-memory computation, such as SRAM that stores the weights digitally and then convert them to analog using pulse width modulation. The interest in multi-level circuits has grown to realize a CMOS-based, multi-bit, analog memory circuits that can perform near memory computation, where the weights are stored in analog domain [23].

VMMs require signed operation since the weights can be positive or negative values. There are two ways to accomplish signed operations: dual array and dual row architectures as shown in Figures 1 (a) and (b), respectively [24]–[27]. The dual array implementation shown in Figure 1(a) uses two identical sub-arrays within a crossbar to subtract the partial products in the digital domain. In contrast, the dual row

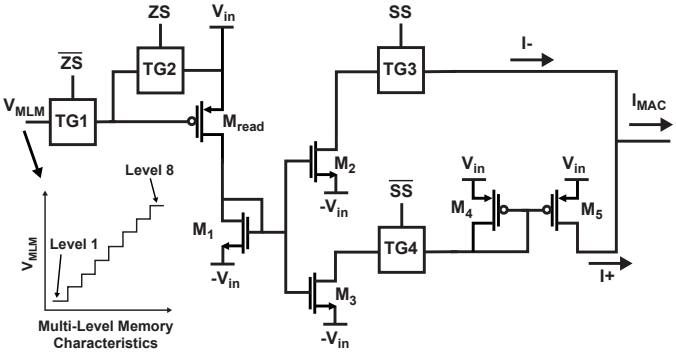

Fig. 2. Dual row architecture.

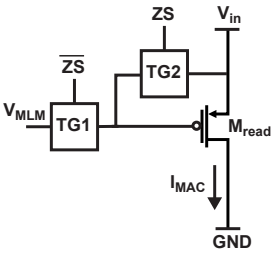

Fig. 3. Dual array architecture.

implementation displayed in Figure 1(b) uses two identical memories which have an equal magnitude and opposite sign applied to each memory for subtraction in the analog domain. Each approach has advantages and disadvantages over the other which are discussed throughout this paper.

This paper investigates the design decisions and constraints that come with the dual row and dual array approaches to implementing signed operations in mixed-signal neural networks. This investigation will be carried out through the use of a case study topology which uses an analog multi-level memory, developed by the authors. This paper is organized as follows. Section II will present and explain the two architectures, dual array and dual row. Section III will provide simulation results which form the basis of the design decisions. Section IV will conclude the paper.

II. Mixed Signal Neural Network Architectures

Mixed Signal accelerators employ two methods to implement signed weights. The dual row approach in Figure 2 performs the subtraction in the analog domain by subtracting currents, while the dual array approach in Figure 3 performs the subtraction in the digital domain after each current is passed through an ADC. This section will explain each architecture of the case study in detail.

A. Multi-level Analog Memory

The memory utilized in this work uses a recently developed CMOS-based, multi-level memory (MLM) topology, published by the authors in [28], [29]. The memory takes an analog input and outputs one of eight distinct analog voltages. The memory uses a inverter based structure to generate the eight distinct memory voltages which are then fed into a feedback structure to produce the distinct memory voltage (V_{MLM}) that is closest to the analog input value. The characteristic of the MLM is shown in the graph in Figure 2, where 8-level MLM is utilized.

B. Dual Row Mixed Signal Neural Network

The dual row architecture is shown in Figure 2, where the transmission gates, TG1 and TG2, are used to control if the output of that neuron is zero or non-zero. The control input to this block is called zero select (ZS). If ZS is logic HIGH (1V) then the M_{read} transistor is OFF and the current in M_{read} will be zero. If ZS is logic LOW (-1V), then the overdrive of M_{read} will be controlled by the output of the analog memory, V_{MLM}. M_{read} works as a current source over the range of V_{MLM}. The transmission gates, TG3 and TG4, are used to control if the result is positive or negative. The control input to this block is called sign select (SS), if SS is logic HIGH (1V) then M_2 will act as a current sink, sinking current from the multiply and accumulate (MAC) node thus removing current from the MAC node. If SS is logic LOW (-1V), then the current mirror M_4, M_5 will act as a current source for the MAC node and supply current current to the node.

The input to this block must be bipolar to allow for the sourcing and sinking of the MAC node. This requires usage of Deep N-Well transistors, causing area overhead. For dual row signed implementations, the input voltage needs to be bipolar around the bias point of the MAC node. In this implementation that was chosen to be ground or 0V. Another possibility can be to set the MAC node to be biased to 0.5V to allow for bipolar operation with a uni-polar supply. The drawbacks of having the MAC node biased at 0.5V are reduced range of operation if multi-bit inputs are needed. Also, if a trans-impedance amplifier (TIA) is used to interface with the neuron, then the TIA will need to be adjusted based on a 0.5V input to one terminal which requires extra biasing circuitry. Another drawback of the dual row implementation is the inherent sensitivity to mismatch between the positive and negative MAC current paths. The negative MAC current path only has M_2 whereas the positive MAC current path has M_3, M_4, and M_5. Given the same memory input, if we set SS to logic HIGH or LOW, then the magnitude of current should be the same, but the sign should be opposite.

In total, the dual row implementation has four transmission gates, one read transistor, and five transistors to implement the current mirroring. The total area of the design in 65nm CMOS is $2087.36\mu m^2$. This larger area minimizes the scalability of this particular implementation of the dual row architecture. There are two control pins zero select (ZS) and sign select (SS) which require registers if the architecture is scaled up to avoid a large usage of pins on the IC package.

C. Dual Array Mixed Signal Neural Network

The dual row architecture is shown in Figure 3, which is simpler than the dual row approach. As depicted in the figure, M_{read} is a voltage controlled current source that converts

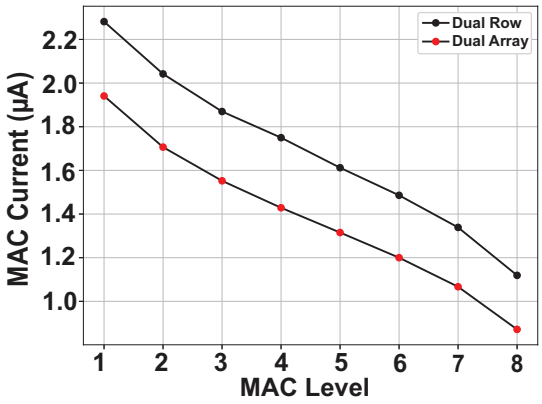

Fig. 4. MAC current of the dual array vs. dual row architectures across all eight analog multi-level memory output levels.

the analog memory voltage V_{MLM} to a MAC current. The transmission gates TG1 and TG2 control zero or non-zero weight similar to the implementation in the dual row mixed signal neural network. The input voltage V_{in} for this block can be uni-polar. A drawback of this topology is the large number of memories required. Each input-weight pair requires two memories rather than one as shown in the dual row section, thus matching between the two memories is important.

This architecture has the advantage of only one device operating in the analog domain to perform the non-signed part of the VMM. In practice this approach is implemented by assigning a crossbar to be a positive crossbar and another crossbar to be a negative crossbar, such as in Figure 1(a). These crossbar outputs are then combined in the digital domain to produce the full signed sum.

This structure also requires two ADCs per neuron which impacts the peripheral power consumption and area. This limitation is not present in the dual row architecture. For the MAC structure itself the area is very small which is dictated by the size of the read transistor. The transconductance of the read transistor by itself is a function of process, voltage, and temperature variations which limits the robustness of this structure thus a limiting the scalability of this implementation.

This structure has a lower MAC current across all input memory voltages, given the same sizing for M_{read}, which results in lower power consumption as shown in the next section. Figure 4 shows the MAC current of the dual array and the dual row architectures.

III. Simulation Results

A few metrics of comparison between the dual row and dual array architectures are considered to compare the two implementations from a system level. This includes speed, power consumption, and variation.

The latency is determined as the time it takes for an input to produce a stable output. The first type of input is a non-zero input, which is applied from zero initial condition and the second type of input is when zero is applied from non-zero initial condition. These two latencies differ from each other.

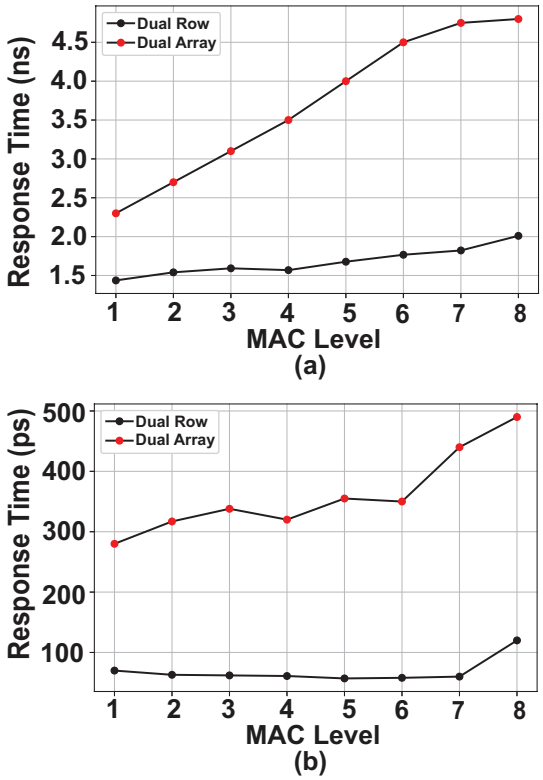

Fig. 5. (a). Latency for dual row and dual array when a non-zero excitation is applied. (b). Latency for dual row and dual array when a zero excitation is applied.

In order to determine the maximum frequency that the dual row/array architecture can operate at, the largest latency must be chosen. The latency is the settling time which is defined as the time it takes to reach the final value for the current. The latency can be decreased by clocking at a period smaller than the settling time, but results in sampling a current that is not the final value. This is a trade-off between accuracy and speed. The decision on the amount of over-clocking is based on ADC resolution and the minimum delta between current levels before an adjacent level is chosen.

Figure 5 shows the latency at different weight storage values. To calculate the maximum usable frequency without performance degradation the worst case for each architecture must be chosen. In Figure 5(a), the maximum latency for the dual array architecture is 4.8 ns at MAC level 8. The maximum operating frequency of the dual array architecture is 208 MHz without degradation in performance. In Figure 5(a) the maximum latency for the dual row architecture is 2.02 ns at MAC level 8. The maximum operating frequency of the dual row architecture is 495 MHz without degradation in performance. The latency and frequency is calculated under the worst case scenario of a square waveform input.

The power consumption has two components: static and dynamic. The static power consumption is dominated by the transconductance of the read transistor. This relationship implies the static power consumption is directly proportional

TABLE I
SUMMARY OF THE POWER CONSUMPTION AND FREQUENCY PERFORMANCE OF EACH ARCHITECTURE.

	Dual Row		Dual Array	
MAC Level	Frequency (MHz)	Power Consumption (μW)	Frequency (MHz)	Power Consumption (μW)
1	696	10.7	434	1.95
2	649	9.50	370	1.71
3	628	8.70	322	1.56
4	637	8.05	285	1.43
5	596	7.45	250	1.32
6	566	6.84	222	1.20
7	548	6.14	210	1.07
8	497	5.09	208	0.88

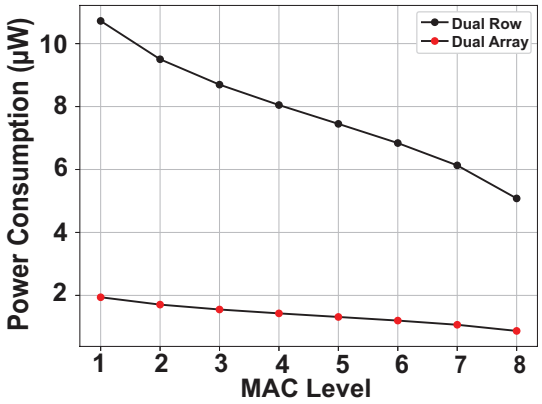

Fig. 6. Power Consumption for the Dual Row and Dual Array architectures.

to transistor width and over-drive and is inversely proportional to transistor length. The over-drive of the read transistor is controlled by the output voltage of the analog memory, leaving the aspect ratio as the design variable. The dynamic power consumption is directly proportional to the frequency and the gate-source capacitance: ωC_{gs}. The dynamic power consumption also depends on $C_{tot}fV^2$. This implies another system level trade-off between power consumption and speed. There are two methods to reduce dynamic power consumption: slope control and transistor sizing. Slope control will reduce the amount of high frequency content at the falling and rising edges which sees the gate-source capacitance as low impedance. Reducing the width of the transistor will reduce the gate-source capacitance.

The power consumption of the two architectures is shown in Figure 6. The power consumption of the dual array implementation is less than that of the dual row implementation with smaller size and little sacrifice in operating frequency. The lower power consumption of the dual array approach is due to the MAC current being lower given the same sizing for the M_{read} transistor. Also, the dual array structure does not have the current mirroring topology to occupy more area. The dual row structure has a power consumption of 10.5μW where the dual array structure has a power consumption of 2μW per neuron. This corresponds to MAC level 1 which is the highest over-drive on the M_{read} transistor. In the zero state, the dual row structure consumes 6nW while the dual array structure consumes 68pW, which is the leakage of each topology. Table I shows a summary of the dual row and dual array architectures explored in this paper.

The trade-off of limiting power consumption from the MAC circuit has implications on the ADC resolution. the trade-off between read transistor transconductance and power consumption means that the ADC must have a higher resolution since the transconductance of the read transistor is lower.

The MAC circuits of the dual array or dual row must be scaled up for a large neural network. Therefore, the impact of variation and any possible trade-offs within the MAC circuit itself or impacts on ADCs or TIAs must be investigated. It is important to understand this impact on a large scale because variation on the device current will impact the overall MAC current and degrade overall system accuracy. The drawback of using MOSFETs (NMOS or PMOS) for VMMs is the variation in transconductance. In both the dual array and dual row architectures presented in this work the transconductance varies with the weight voltage. This gives a non-linear response to the variation, meaning that the distribution will be dependent on the weight value. The variation performance points to another trade-off between robustness and power consumption. If the transconducatance is minimized to minimize power consumption then the spacing of the distributions will be small and robustness will be low. The robustness is directly proportional to the system level accuracy, on a large scale.

IV. CONCLUSION

This paper made a comparative analysis between dual row and dual array architectures through the use of two case study architectures. The trade-offs between speed/latency, power consumption, accuracy, and robustness were explored. The trade-off between speed and accuracy exists because more settling time allows for higher accuracy, but more settling settling time also means slower operating frequency. The trade-off between speed and power consumption exists because at a higher frequency, the dynamic power consumption becomes dominant over static power consumption. The trade-off between power consumption and accuracy exists because to increase the spacing between MAC current distributions and to allow for more selectivity, the current must be increased. Downstream trade-offs were also explored which is why the crossbar architecture must be designed with the ADC in mind.

REFERENCES

[1] M. Alnemari and N. Bagherzadeh, "Efficient deep neural networks for edge computing," in *2019 IEEE International Conference on Edge Computing (EDGE)*, pp. 1–7, 2019.

[2] E. Nishani and B. Çiço, "Computer vision approaches based on deep learning and neural networks: Deep neural networks for video analysis of human pose estimation," in *2017 6th Mediterranean Conference on Embedded Computing (MECO)*, pp. 1–4, 2017.

[3] F. Utaminingrum, M. A. Fauzi, R. C. Wihandika, S. Adinugroho, T. A. Kurniawan, D. Syauqy, Y. A. Sari, and P. P. Adikara, "Development of computer vision based obstacle detection and human tracking on smart wheelchair for disabled patient," in *2017 5th International Symposium on Computational and Business Intelligence (ISCBI)*, pp. 1–5, 2017.

[4] Y. Yang, S. Kim, and J. Joo, "Explaining deep convolutional neural networks via latent visual-semantic filter attention," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8323–8333, 2022.

[5] L. Sharara, A. Politis, H. Syed, E. Kronell, D. Dunsmore, T. Thierfelder, J. Wolf, J. Süß, L. Mansour, K. Thelen, L. Alazzawi, and M. Ismail, "A real-time automotive safety system based on advanced ai facial detection algorithms," *IEEE Transactions on Intelligent Vehicles*, pp. 1–22, 2023.

[6] M. Qin, "Machine translation technology based on natural language processing," in *2022 European Conference on Natural Language Processing and Information Retrieval (ECNLPIR)*, pp. 10–13, 2022.

[7] 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.

[8] M. Li, M. Sun, Q. Liu, and Y. Zhang, "Fraud detection based on graph neural networks with self-attention," in *2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pp. 349–353, 2021.

[9] M. L. Gambo, A. Zainal, and M. N. Kassim, "A convolutional neural network model for credit card fraud detection," in *2022 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 198–202, 2022.

[10] S. Sindhu and M. Saravanan, "Part-based convolutional neural network and dual interactive wasserstein generative adversarial networks for land mark detection and localization of autonomous robots in outdoor environment," in *2022 1st International Conference on Computational Science and Technology (ICCST)*, pp. 1062–1066, 2022.

[11] J. Liu, "Survey of the image recognition based on deep learning network for autonomous driving car," in *2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*, pp. 1–6, 2020.

[12] S. M. Varnosfaderani, R. Rahman, N. J. Sarhan, L. Kuhlmann, E. Asano, A. Luat, and M. Alhawari, "A two-layer lstm deep learning model for epileptic seizure prediction," in *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pp. 1–4, 2021.

[13] R. Rahman, S. M. Varnosfaderani, O. Makke, N. J. Sarhan, E. Asano, A. Luat, and M. Alhawari, "Comprehensive analysis of eeg datasets for epileptic seizure prediction," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2021.

[14] N. Fatehi, A. Politis, L. Lin, M. Stobby, and M. H. Nazari, "Machine learning based occupant behavior prediction in smart building to improve energy efficiency," in *2023 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pp. 1–5, 2023.

[15] T. Ishitaki, R. Obukata, T. Oda, and L. Barolli, "Application of deep recurrent neural networks for prediction of user behavior in tor networks," in *2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pp. 238–243, 2017.

[16] D. Sinha and M. El-Sharkawy, "Thin mobilenet: An enhanced mobilenet architecture," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pp. 0280–0285, 2019.

[17] S. Bouguezzi, H. Faiedh, and C. Souani, "Slim mobilenet: An enhanced deep convolutional neural network," in *2021 18th International Multi-Conference on Systems, Signals Devices (SSD)*, pp. 12–16, 2021.

[18] B. Murmann, "Mixed-signal computing for deep neural network inference," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 29, no. 1, pp. 3–13, 2021.

[19] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, "An always-on 3.8j/86% cifar-10 mixed-signal binary cnn processor with all memory on chip in 28nm cmos," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, pp. 222–224, 2018.

[20] Y.-H. Chen, T. Krishna, J. Emer, and V. Sze, "14.5 eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 262–263, 2016.

[21] Z. Jiang, S. Yin, M. Seok, and J.-s. Seo, "Xnor-sram: In-memory computing sram macro for binary/ternary deep neural networks," in *2018 IEEE Symposium on VLSI Technology*, pp. 173–174, 2018.

[22] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[23] U. Cilingiroglu and Y. Ozelci, "Multiple-valued static cmos memory cell," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 48, no. 3, pp. 282–290, 2001.

[24] Q. Wang, Y. Park, and W. Lu D., "Device variation effects on neural network inference accuracy in analog in-memory computing systems," in *2022 Advanced Intelligent Systems*, vol. 4, 2022.

[25] D. Kim, C. Yu, S. Xie, Y. Chen, J.-Y. Kim, B. Kim, J. P. Kulkarni, and T. T.-H. Kim, "An overview of processing-in-memory circuits for artificial intelligence and machine learning," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 12, no. 2, pp. 338–353, 2022.

[26] V. Joshi, M. Le Gallo, S. Haefeli, I. Boybat, S. R. Nandakumar, C. Piveteau, M. Dazzi, B. Rajendran, A. Sebastian, and E. Eleftheriou, "Accurate deep neural network inference using computational phase-change memory," *Nature Communications*, vol. 11, no. 2473, 2020.

[27] X. Wang, Q. Wang, F.-H. Meng, S. H. Lee, and W. D. Lu, "Deep neural network mapping and performance analysis on tiled rram architecture," in *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pp. 141–144, 2020.

[28] M. Alhawari and M. H. Perrott, "A clockless, multi-stable, cmos analog circuit," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1764–1767, 2014.

[29] 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.