Singh, Gaurav; Numan, Omar; Monga, Dipesh; Andraud, Martin; Halonen, Kari

On-chip Built-In Self-Calibration of Thermal Variations for Mixed-Signal In-Memory Computing

Published in:
Proceedings - 2024 29th IEEE European Test Symposium, ETS 2024

DOI:
10.1109/ETS61313.2024.10567960

Published: 01/01/2024

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

On-chip Built-In Self-Calibration of Thermal Variations for Mixed-Signal In-Memory Computing

Gaurav Singh*, Omar Numan*, Dipesh Monga*, Martin Andraud*†, and Kari Halonen*
*Aalto University, Department of Electronics and Nanoengineering, Finland {firstname.lastname}@aalto.fi
†UC Louvain, ICTEAM {firstname.lastname}@uclouvain.be

***Abstract*—In-memory computing (IMC) accelerators have become a pivotal architecture for enhancing AI algorithm computations, particularly critical for embedding deep neural networks (DNNs) in edge devices. The efficiency of these systems is paramount, yet IMC cores are prone to fluctuations due to process, temperature, and voltage variations, which can detrimentally impact DNN accuracy. This research introduces an innovative Built-In Self-Calibration (BISC) methodology, specifically designed to compensate for temperature-induced variations in mixed-signal IMC cores. The methodology enables real-time, on-chip adjustment of DNN weights during computation within the IMC core without modifying the computation path. The proposed approach, implemented on a silicon prototype, not only maintained DNN computation accuracy under substantial temperature variations but also fully compensated for almost 90% of the offset caused by these variations, without introducing any non-idealities.**

***Index Terms*—In-memory computing (IMC), multiply-and-accumulate (MAC), thermal compensation, temperature sensor.**

I. Introduction

Computing artificial intelligence workloads with increased energy efficiency is crucial to enable the next generation of automotive, medical, wireless communication or ubiquitous sensing systems. Hence, the *acceleration* of machine-learning models, particularly Deep Neural networks (DNNs), has gained a huge interest over the past years. In this context, In-Memory Computing (IMC) accelerators represent a significant advancement in overcoming the limitations of traditional computing paradigms, by eliminating the need for frequent data transfer between storage and processing elements [1]. Specifically, IMC architecture can use efficient analog computations for multiply-and-accumulate (MAC) operations, which consume most of DNN's computing energy [2]. In analog computations, a MAC is computed using basic Kirchoff's laws, i.e., multiplication using resistances or capacitances, and native addition in current or charge domains [3]. This motivates a large body of research on emerging architectures for IMC, building the IMC core with classical charge-based memories [3], or emerging memory technologies, for instance resistive memories such as resistive RAM (ReRAM) or phase change memories (PCM) [4]. These emerging technologies pave the way towards significantly more integrated and efficient architectures.

However, large-scale integration of IMC computing cores using emerging memory technologies, such as ReRAM, faces challenges related to reliability and robustness due to inherent variability and endurance issues [5]. These challenges can significantly impact the accuracy of DNN computations. Current research efforts are directed towards developing innovative circuit designs and system architectures to improve the reliability of IMC cores, addressing these variability and endurance concerns. In particular, there has been significant research work documenting the effect of temperature variations on SRAM-based IMC cores [6], resistive IMC architectures with high-density resistors [7] or emerging memories such as ReRAM [8], [9] or PCMs [10].

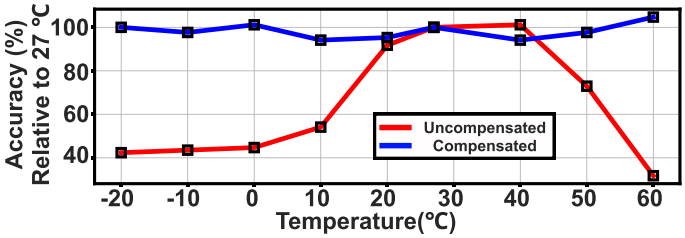

Fig. 1: Impact of temperature on mixed-signal IMC cores w.r.t. relative inference accuracy.

This is illustrated in Fig. 1, taking as a case study a mixed-signal IMC core (further detailed in this work) on a typical DNN workload (MNIST). In this particular case, the worst-case accuracy loss, compared to the baseline at room temperature, can be as high as 68%. Hence, compensation methods for thermal variations are highly required to maintain a reliable DNN inference across wide temperature ranges. To tackle this challenge, this work proposes a thermal aware Built-in Self-Calibration (BISC) technique, which performs on-chip real-time compensation on mixed-signal IMC cores. With this compensation method, the effects of temperature fluctuations on MAC units' performance are reduced up to 6%, indicating a 91% improvement in inference accuracy across various temperatures.

The structure of this paper is as follows: The next section provides an overview of related work and highlights its limitations. Section III details the proposed compensation architecture, including its building blocks. Section IV elaborates on the proposed thermal compensation methodology. Experimental and measurement results are presented in Section V. Finally, Section VI concludes the paper.

II. Related work and limitations

A. Existing thermal compensation techniques

Thermal compensation methods can be broadly classified into two categories: offline compensation, which occurs during or after training, and online compensation, which takes place before or during inference. Typically, a system can implement both categories simultaneously.

1) Offline: Offline compensation targets limiting DNN accuracy degradation from temperature effects by preemptively reordering data, weights, or tasks at the IMC core or system level. Example approaches include weight decomposition or assigning computations based on core temperature to reduce thermal variations [8]. While offline methods effectively manage workloads, they need to be paired with local adaptation strategies to comprehensively tackle temperature variations [8].

2) Online: Online compensation aims to compensate for local temperature variations during DNN execution on the accelerator. There are two subcategories: online system adaptation and BISC.

a) Online system adaptation: These methods have a similar system-level approach to offline compensation but enable a finer granularity as they are performed during the runtime. A first approach pairs IMC cores with different temperatures, splits the workload between them, and combines their results to be thermally compensated. A finer-grained approach would modify the DNN model's parameters (weights, bias, etc.) according to different temperatures, using an offline model trained for various temperatures. As updating all weights before each inference would be too costly, the batch normalization parameters of the DNN can be adapted instead (one set of parameters per chosen temperature range) [11]. In this case, these parameters are determined using a progressive knowledge distillation algorithm, injecting noise/variations in the trained model for various temperatures before deployment [11]. Although effective, online system adaptation methods involve complex algorithms and face challenges in addressing local temperature changes.

b) Built-in Self-Calibration: BISC differs from system adaptation as it intervenes *locally* at the hardware level (typically the MAC units or the IMC column). The objective is to directly calibrate the computed results to account for local temperature changes. BISC can can be used in complement to system-level approaches. One example of BISC is to adjust the output current of the IMC column, using current mirrors of variable ratios according to the temperature [8]. The column current compensation can also be done using an on-chip temperature dependent, variable current generator, at the same time compensating for process variations [7]. However, these methods require precise current ratios for each column [8] or off-chip components [7], complicating the final integration. An alternative approach is to use one additional column of the IMC core for temperature compensation [6], [10]. Taking the example of [10], the dummy column weights are learned offline from thermal model simulations, stored in a look-up table, and set according to different temperatures online. A more fine-grained approach would adjust the weight of each resistive element (or conductance) in the IMC core according to temperature. For instance, the least significant bit of the weight can be downgraded if the core temperature goes across a predefined threshold [9]. Yet, the compensation is limited to one bit of change in the weights.

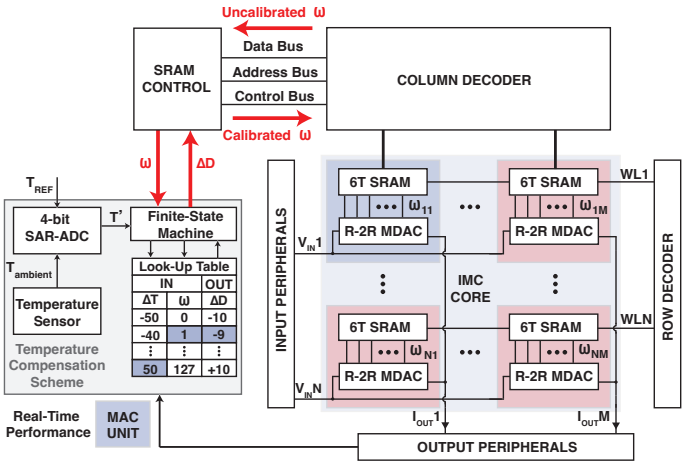

Fig. 2: Overall system architecture of the proposed temperature compensation scheme seamlessly integrated with a mixed-signal IMC core.

B. Proposed contribution

This work addresses the highlighted limitations by: 1.) developing a fully integrated, real-time BISC for mixed-signal IMC cores to handle both local and global temperature variations without requiring complex algorithms; 2.) minimally impacting the IMC core's computation path by updating stored weights rather than analog currents/voltages; and 3.) validating the calibration algorithm across a broad temperature range with silicon devices.

III. Proposed architecture

Fig. 2 depicts the proposed architecture for thermal compensation, focused on *localized* and *hardware-efficient* thermal management solutions. The scheme exploits existing circuitry within the IMC core as much as possible and is tailored to preserve the accuracy of MAC operations. The key principle is to directly adjust the weight values of the MAC units stored in SRAM cells according to the temperature variations sensed by an on-chip sensor. This section highlights the key system functionalities.

A. System architecture

The overall system is composed of an IMC core and a BISC circuitry which adjusts the stored weights directly online. This precise adjustment is achieved through the integration of a moderate resolution SAR ADC (e.g., 4-bit), a Finite State Machine (FSM), and a Look-up Table (LUT).

a) IMC core: The IMC core features a crossbar array of size (N × M). At each intersection of this crossbar lies a mixed-signal MAC unit to multiply a vector of input voltages V_{IN_i} with their respective weights W_{ij} stored in a SRAM memory within each unit. The multiplication is realized with an N_W-bit R-2R Multiplying Digital-to-Analog Converter (MDAC) and N_W 6T-SRAM cells, where N_W represents the precision of the weights in the MAC unit. Hence, weights are digitally represented as conductance values $W_{ij} = G_{ij} = 1/R_{ij}$, encoded in the MAC units at each (i, j) position in the array (see details in Section III-B). Then, the output current I_{out} of each unit is accumulated along each column, representing the results of the MAC operations.

b) Input/output peripherals: Input and output peripherals are crucial for IMC cores, handling tasks around MAC operations such as data conversion and sampling, distributing data across the memory crossbar, and interfacing with external devices like digital processors. A critical output component is the Summing Amplifier (SA), particularly Transimpedance Amplifiers (TIA), crucial for converting output currents from memory columns into voltages for further processing, including activation functions, pooling, and ADC conversions.

c) SRAM control: The SRAM control block and the row/column decoders are integral part of an effective system management, responsible for programming, writing, and reading the weight values in the N_W 6T-SRAM cells of each MAC unit. The row and column decoders accurately select the specific cells within the core. The SRAM cells are arranged using a half-butterfly configuration to optimize data routing and SRAM read/write efficiency by providing an efficient data transfer pathway crucial for parallel processing tasks. The half-butterfly layout is chosen for its balance between high performance and manageable complexity in the memory layout.

d) Thermal BISC block: The proposed BISC method incorporates a temperature sensor font-end, as well as an FSM and a LUT for the temperature compensation. The sensor front-end can accurately detect a wide temperature range, from -40°C to 80°C, with a granularity of 7.5°C. The proposed calibration method is particularly suited for integration into existing systems, requiring minimal alterations, thereby offering a practical solution in rapidly evolving technological environments (see details in Section IV).

B. MAC units

Fig. 3 illustrates the proposed MAC unit based on a R-2R MDAC topology. It is similar to a typical R-2R DAC topology, yet has the reference voltage as a variable input. As a complete integration of emerging resistive memories with state-of-the-art CMOS technologies is still a work in progress, the R-2R MDACs are realized with CMOS resistors. Prospectively, these CMOS resistors can be replaced with high-density resistor technologies that enable MΩ resistor values in compact spaces comparable to standard SRAM cells [12]. Using R-2R MDACs allows to obtain linear and programmable features, offering a solution to the typically fixed and non-linear behaviour of ReRAM cells. Indeed, this architecture enables to perform

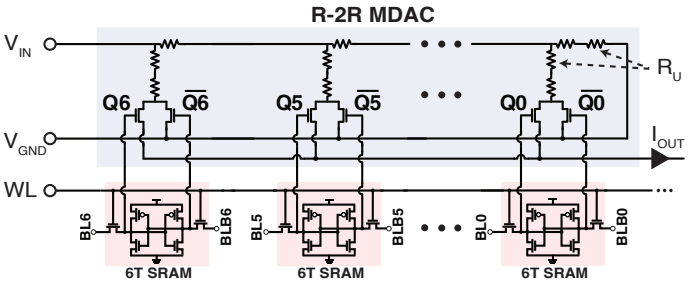

Fig. 3: Schematic of R-2R MDAC interfaced with 6T-SRAM cells as an integrated component of a mixed-signal MAC unit.

an analog multiplication while storing weights digitally in SRAM cells. The intrinsic linearity of the R-2R ladder topology ensures a direct and proportional relationship between digital weights and analog outputs, making it suitable for applications requiring frequent weight adjustments and high precision. In the context of temperature compensation, this dual approach not only streamlines the design but also significantly boosts the accuracy and efficiency of temperature compensation. Digitally storing weights results in more precise and stable calibration, a key factor in maintaining consistent performance across varying temperatures.

The output current of an MDAC cell is given by

$$I_{OUT} = \frac{V_{IN} - V_{GND}}{R_U} \cdot \frac{\sum_{k=0}^{N_W-1} 2^k \cdot D_k}{2^{N_W}}, \tag{1}$$

where V_{GND} is the virtual ground voltage, and R_U is the unit resistance. The second term in the equation represents the decimal equivalent of the digital weight (D) relative to the converter's resolution, where D_k is the binary value at k^{th} bit position. Hence, the consistency and predictability of R-2R MDACs make them a suitable case study for testing the proposed temperature compensation method.

C. 5-T Temperature sensor

The temperature sensing front-end relies on an on-chip all NMOS, 5-T (Transistor) based temperature sensor, based on a modification of [13], as shown in Fig. 4. The circuit generates a complementary to absolute temperature (CTAT) voltage as an output w.r.t change in temperature. It consists of three types of transistors: a native transistors M_{1A}, M_{1B} with almost zero threshold voltage (V_{TH}), a thin-oxide transistor M_2 with low V_{TH} and two thick-oxide transistors M_3 and M_4 with high V_{TH}. A supply voltage of V_{DD} is provided to the circuit, and the temperature dependant voltage output is obtained at V_{OUT}. All the transistors operate in sub-threshold region to optimize power consumption. M_1 consists of stacked identical transistors M_{1A} and M_{1B} to obtain a good line sensitivity. M_3 and M_4 are used as an active load to provide a temperature-dependent output voltage. The output voltage value is the sum of V_{DS} of both the transistors. The V_{OUT} is given by

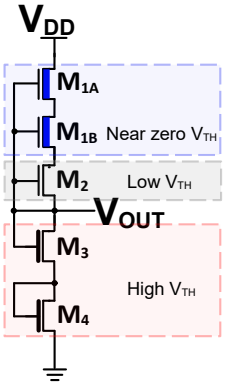

Fig. 4: Schematic of implemented all NMOS 5-T temperature sensor circuit as a key component of the proposed BISC.

$$V_{OUT} = V_{TH_{4,3}} - \frac{m_{4,3}}{m_2} V_{TH_2} + m_2 V_T ln \frac{C_{OX_2}(m_2 - 1)(W_2/L_2)}{C_{OX_4}(m_{4,3} - 1)(W_4/L_{4,3})}, \quad (2)$$

where W and L are the width and length of the device, μ is the mobility, C_{OX} is the gate oxide thickness, m is the subthreshold slope, V_T is the thermal voltage of the transistors, which are denoted by the subscript. From (2), it can be inferred that by selecting the aspect ratio of M_2, M_3 and M_4, the desired CTAT voltage characteristic can be achieved from the proposed circuit.

IV. Thermal Compensation Methodology

A. Compensation algorithm

The proposed temperature compensation algorithm, detailed in pseudocode in Algorithm 1, includes key components such as the sensor front-end (temperature sensor and ADC) and calibration circuitry (FSM and LUT). Upon initiating the BISC with a threshold, the system acquires the temperature, and the digital FSM evaluates the necessity for compensation. Compensation is triggered by a temperature change of at least 7.5°C, corresponding to a voltage shift of 6.25mV, as outlined in the steps below:

- **Detection and triggering:** The FSM continuously tracks the ADC's output for significant temperature changes, ΔT. Upon detection, it initiates the calibration process for each layer during inference on the IMC core, transitioning to the compensation state to execute calibration.
- **Calibration execution:** In this phase, the system temporarily deactivates all MAC units except one selected cell, which undergoes a weight sweep from 0 to 2^{N_W} - 1, with N_W being the MAC unit's bit count. Output values at the new temperature are noted and compared to reference values recorded at a standard room temperature of 27°C on silicon samples. This comparison is essential for evaluating the binary weight change (ΔD) variation.
- **ΔD calculation:** Temperature changes create different changes in ΔD values, which are recorded in a LUT for thermal adjustment. An on-chip microprocessor or a similar digital logic can update this LUT in real time. This update happens only once for each temperature change in the IMC core. ΔD is normalized to the least significant bit (LSB), defined as 1 LSB = $I_{MAX}/2^{N_O}$. Here, I_{MAX} is the MAC unit's maximum output current, and the output precision, N_O is the sum of the MAC unit's input precision N_I and weight precision N_W.
- **Weight adjustment:** For each non-zero weight in the IMC core, the FSM generates the SRAM controls to update the MAC unit's weight: $W'_{ij} = W_{ij} - \Delta D \times 1$ LSB for the respective ΔT.

Algorithm 1 Online Temperature Compensation Algorithm

```
Initialize the system and compile the LUT.
Monitor ADC output for significant ΔT, w.r.t. 27°C.
if Temperature change ΔT ≥ 7.5°C then
    for each DNN layer do
        for each SRAM weight stored in the array do
            Read the weight W_ij for each SRAM word stored
            at address A_ij (SRAM Read Cycle).
            if W_ij = 0 then
                Skip to address A_ij + 1
            else
                Find ΔD for W_ij at A_ij from LUT.
                Update the weight, W'ij = W_ij - (ΔD × 1LSB)
                (SRAM Write Cycle).
            end if
        end for
    end for
    Calculate the temperature-compensated layer output.
end if=0
```

As a key feature, the LUT enhances online re-compensation efficiency by storing ΔD for each weight across the temperature range. If BISC detects a known temperature, the FSM skips recalibration, directly adjusting weights with ΔD from the LUT, avoiding unnecessary IMC core or microprocessor interactions. This approach minimizes redundant calibrations, improving long-term system efficiency and reducing power consumption.

V. Measurement Results

Fig. 5 shows the microphotograph and layout of the test circuit in 65 nm CMOS technology, including a column of MAC units, a Summing Amplifier (SA), and a temperature sensor. The dimensions are 232 μm × 38 μm for each MAC unit, 202 μm × 532 μm for the column, 71 μm × 47 μm for the temperature sensor, and 57 μm × 36 μm for the SA. Weights are programmed via an on-chip interface, with MAC units powered by individual buffered supplies from AD8656 buffers and a Keysight N6705C DC Power Analyser for precise control. The temperature is tested in a controlled environment in the Weisstechnik Labevent chamber ranging from -20°C to 60°C.

A. Building the compensation scheme

Data from a silicon prototype is collected by measuring a single MAC unit's output current across temperatures (-20°C

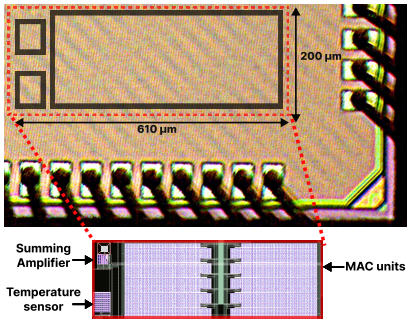

Fig. 5: Microphotograph and layout of the test circuit.

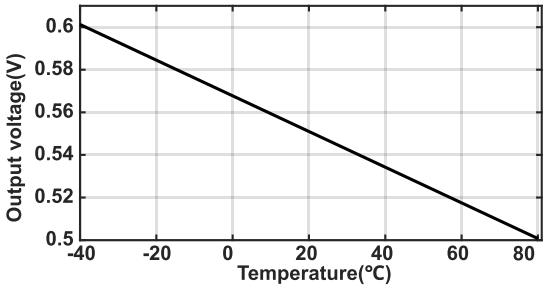

Fig. 6: Measured temperature sensor output.

to 60°C, in 7.5°C steps) and input voltages (1V to 1.1V). This information is used to create a database that quantifies output currents for each quantization level across all 7-bit inputs and weights, with these currents normalized across the range. This database is further used to generate MAC unit multiplication results at various temperatures and constructing an LUT model for the BISC approach. To handle the extensive data from testing the 7-bit MAC unit's full input voltage and weight range, a strategy focusing on the input dynamic range's midpoint simplifies calibration, i.e., fixing the input voltage to be in the middle of the range. This practical and robust approach simplifies calibration, enabling complete on-chip compensation without the complexity of a large dataset.

B. Evaluation of the BISC scheme

Fig. 6 shows the measurement results of the temperature sensor, which provides a temperature coefficient of 1550 ppm/°C. The designed circuit consumes a power of 16.54 nW under an operating voltage of 1.2 V.

Fig. 7 shows the impact of temperature compensation on a single MAC unit across temperatures from -20°C to 60°C, using a full scale of 7-bit weights and input values with 0.787 mV quantization step. The output current is recorded for all these values at various temperatures. The figure reports the deviations in LSB from the 27°C reference baseline. Results reveal broad error distribution without compensation, with a mean (μ = -5.08) and a standard deviation (σ = 2.09), indicating systematic temperature-related errors and high variability. The proposed compensation method significantly corrects these errors, nearly eliminating the mean error (μ = 0.09) and reducing the standard deviation to below 1 LSB (σ = 0.98), proving its effectiveness in minimizing MAC errors caused by temperature fluctuations.

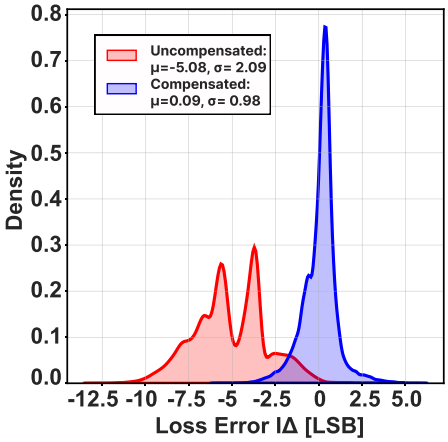

Fig. 7: The density of the MAC accuracy (in LSB) measured across various ambient temperatures before/after compensation.

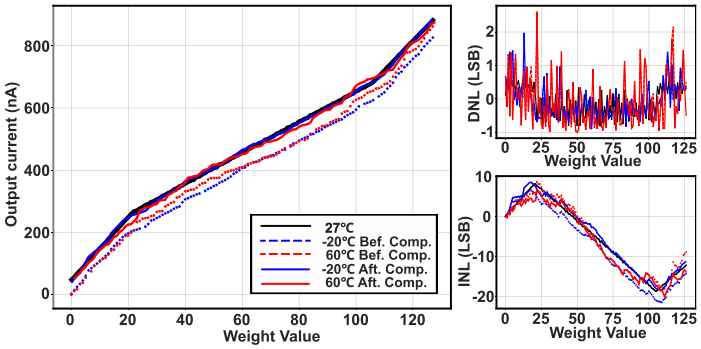

Fig. 8: MAC unit transfer function and linearity before/after temperature calibration.

Fig. 8 evaluates MAC unit linearity across temperature variations, which affect conductance and multiplication accuracy, thus impacting linearity. Uncompensated data show negative offsets due to temperature changes in a specific MAC unit. Post-compensation, response curves closely align with the 27°C reference, indicating preserved linearity across temperatures. Additionally, the figure presents DNL and INL of MAC responses before and after compensation, showing consistent non-linearity errors and confirming the compensation method maintains system linearity without adding errors.

Table I compares this work with prior works [6], [9], evaluating a typical DNN workload. Data extracted from silicon measurements, incorporating the proposed compensation strategy into TensorFlow, assesses inference accuracy across temperatures using LeNet-5 architecture on the MNIST dataset for handwritten digit classification. Training involved 60,000 images, and inference was tested on a separate set of 10,000 images, with accuracy determined from 10,000 predictions.

Although the impact of temperature is severe on the accuracy, it can be recovered by the proposed methodology, as initially illustrated in Fig. 1. This recovery process is further evidenced by data in Table I, which shows that deviations in temperature from the baseline of 27°C result in accuracy losses of 68.2% and 57.6% for significant positive and negative temperature changes,

TABLE I: COMPARATIVE ANALYSIS OF THE PROPOSED COMPENSATION METHOD WITH LITERATURE

		[6]	[9]	This Work
Technology		40nm	-	65nm
Cell Type		6T-SRAM	ReRAM	R-2R MDAC + 6T-SRAM
In/W/Out precision		1b/1b/8b	1b/2b/8b	**7b/7b/7b**
Validation		Simulation	Silicon	Silicon
DNN Benchmark		MNIST	MNIST	MNIST
Accuracy without compensation	-20 °C	83.00%*	-	36%
	27 °C	90.97%	90%	85%
	60 °C	87.36%*	55%	27%
Accuracy with compensation	-20 °C	96.14%	-	85%
	27 °C	96.14%	95%	85%
	60 °C	96.14%	92%	89%
Loss relative to 27°C				
Before compensation	-20 °C	8.76%	-	57.6%
	60 °C	3.90%	38.9%	68.2%
Recovered accuracy relative to 27°C				
With compensation	-20 °C	100%	-	100%
	60 °C	100%	96.84%	104.7%

∗ Linearly extrapolated from the original work.

respectively. Remarkably, the proposed methodology facilitates a substantial accuracy recovery of 104.7% and 100% relative to the baseline (room temperature). Despite the baseline accuracy being lower compared to other designs, it can be significantly enhanced during the training phase. More importantly, the recovery measured against this baseline accuracy demonstrates the proposed approach's effectiveness in maintaining consistent performance across different temperatures. During testing, it was observed that the system's performance marginally improves when the temperature ranges from 30°C to 40°C, as shown in Fig. 1. We hypothesize that this increase in temperature may uniformly enhance the accuracy of the MAC units' outputs in the IMC core, resulting in better DNN performance compared to room temperature. However, relying on ambient temperature changes for optimization is impractical due to temperature variability, emphasizing the need for a compensation method to ensure accuracy across various temperatures.

C. Analysis of calibration overheads

In addition to the calibration performance, overheads in time, power, memory, and area, also need to be analyzed. In terms of time, there is an overhead due to rewriting part of the weights into the memory. If most weights are updated, a 30% increase in weight update times can occur. Yet, this is mitigated by the infrequent need for compensation updates and the prevalence of zero-value weights. For instance, updating the weights only every 100 access in the core reduces the time overhead to less than 0.5%. In terms of power consumption, the compensation logic adding approximately 14µW, which is a marginal amount compared to the system's overall power demand of a few mW. In terms of memory, the requirements for the LUT used in temperature compensation depend on the IMC core size lead. For instance, the memory increase is 112.5% for a 32x32 IMC core, equating to a total memory of 1.904KB from an initial 0.896KB, whereas for a larger 128x128 IMC core, the memory overhead is only 7%. Area overhead follows a similar trend than memory, with a 3.04% increase for the 32×32 core and reduced to 0.856% for 128×128 core. Overall, this underscores the efficiency of scaling up the core size in reducing the proportional impact of the compensation circuit.

VI. Conclusion

This work proposes an effective on-chip Built-in Self-calibration methodology against temperature variations for mixed-signal In-Memory Computing arrays. It can be realized fully on-chip, does not change the computation path to preserve the performance of the system and does not require complex algorithms to determine the compensation values. The proposed methodology is validated with a silicon prototype, effectively limit the error below 1 LSB across a wide temperature range.

Acknowledgments

This work is partially supported by Academy of Finland projects EHIR (grant 13334487) and WHISTLE (grant 332218).

References

[1] 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.

[2] W. Haensch, T. Gokmen, and R. Puri, "The Next Generation of Deep Learning Hardware: Analog Computing," *Proc. IEEE*, vol. 107, no. 1, pp. 108–122, 2019.

[3] J.-s. Seo, J. Saikia, J. Meng, W. He, H.-s. Suh, Anupreetham, Y. Liao, A. Hasssan, and I. Yeo, "Digital Versus Analog Artificial Intelligence Accelerators: Advances, trends, and emerging designs," *IEEE Solid-State Circuits Magazine*, vol. 14, no. 3, pp. 65–79, 2022.

[4] Le Gallo, et al., "A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference," vol. 6, no. 9, pp. 680–693.

[5] M. Valad Beigi and G. Memik, "THOR: THermal-aware Optimizations for extending ReRAM Lifetime," in *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2018, pp. 670–679.

[6] Q. Zang *et al.*, "Temperature Compensation on SRAM-Based Computation in Memory Array," in *2022 19th International SoC Design Conference (ISOCC)*, 2022, pp. 1–2.

[7] D. C. Monga, O. Numan, M. Andraud, and K. Halonen, "A Temperature and Process Compensation Circuit for Resistive-based In-memory Computing Arrays," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2023, pp. 1–5.

[8] H. Shin, M. Kang, and L.-S. Kim, "A Thermal-aware Optimization Framework for ReRAM-based Deep Neural Network Acceleration," in *2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, 2020, pp. 1–9.

[9] X. Liu, M. Zhou, T. S. Rosing, and J. Zhao, "HR3AM: A Heat Resilient Design for RRAM-based Neuromorphic Computing," in *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 2019, pp. 1–6.

[10] I. Giannopoulos *et al.*, "Temperature Compensation Schemes for In-Memory Computing using Phase-Change Memory," in *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2020, pp. 286–290.

[11] J. Meng, W. Shim, L. Yang, I. Yeo, D. Fan, S. Yu, and J.-s. Seo, "Temperature-Resilient RRAM-Based In-Memory Computing for DNN Inference," *IEEE Micro*, vol. 42, no. 1, pp. 89–98, 2022.

[12] W. Choi *et al.*, "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)*, 2021, pp. 23.1.1–23.1.4.

[13] K. Yu, Y. Zhou, S. Li, and M. Huang, "A 23-pW NMOS-Only Voltage Reference With Optimum Body Selection for Process Compensation," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 11, pp. 4213–4217, 2022.