TRIM: Thermal Auto-Compensation for Resistive In-Memory Computing

Dipesh C. Monga, *Graduate Student Member, IEEE*, Gaurav Singh, *Graduate Student Member, IEEE*, Omar Numan, *Student Member, IEEE*, Kazybek Adam, *Student Member, IEEE*, Martin Andraud, *Member, IEEE*, and Kari A. I. Halonen, *Member, IEEE*

***Abstract*—In-memory computing (IMC) has emerged as one of the most promising architectures to efficiently compute artificial intelligence tasks on hardware, particularly deep neural networks (DNNs). IMC can make use of analog computation principles alongside emerging nonvolatile memories (eNVM) technologies, potentially offering several orders of magnitude increased energy efficiency compared to generic processing units. Yet, the use of analog circuitry, potentially integrated with emerging technologies post-processed on top of silicon wafers, increases the susceptibility of hardware to a large spectrum of variations, for instance manufacturing, noise or temperature sensitivity. Hence, this susceptibility can hamper the large-scale deployment of IMC circuits into the market. To tackle the reliability of analog resistive-based IMC circuits regarding temperature variations, this article presents TRIM, a thermal on-chip auto-compensation method aimed at fully calibrating first-order temperature effects. TRIM is designed to maintain the computational accuracy of IMC cores in DNN applications over a wide temperature range, while being highly scalable and adaptable. In essence, the temperature compensation is realized through a complementary-to-absolute-temperature (CTAT) voltage reference integrated inside a voltage regulator and applied at the zero reference node of a multiplying digital-to-analog converter (MDAC), eliminating the need for external circuits or look-up table. The proposed methodology is demonstrated on a proof-of-concept 65 nm CMOS resistive IMC column. Measurement results showcase that the proof-of-concept auto-compensation system significantly enhances inference and multiply-and-accumulate (MAC) operation accuracy of any first-order resistive crossbar column, achieving inference accuracy recovery of 100% over a temperature range of –20 °C to 60 °C and a 91.3% improvement in MAC operation accuracy, with an area overhead of 2% and power overhead of < 0.02%.**

***Index Terms*—Compensation scheme, in-memory computing (IMC), multiply-and-accumulate (MAC), resistive crossbar, temperature compensation, ultralow power.**

Received 5 February 2025; revised 4 May 2025 and 20 June 2025; accepted 3 July 2025. Date of publication 8 July 2025; date of current version 22 January 2026. This work was supported in part by the Academy of Finland through the Project EHIR under Grant 13334487 and through the Project WHISTLE under Grant 332218. This article was recommended by Associate Editor W. Liu. *(Dipesh C. Monga and Gaurav Singh contributed equally to this work.) (Corresponding author: Dipesh C. Monga.)*

Dipesh C. Monga is with the Department of Electronics and Nanoengineering, Aalto University, 02150 Espoo, Finland, and also with the Integrated Circuits, VTT Technical Research Centre of Finland, 02150 Espoo, Finland (e-mail: Dipesh.Monga@aalto.fi).

Gaurav Singh, Omar Numan, Kazybek Adam, and Kari A. I. Halonen are with the Department of Electronics and Nanoengineering, Aalto University, 02150 Espoo, Finland (e-mail: Gaurav.Singh@aalto.fi; Omar.Numan@aalto.fi; Kazybek.Adam@aalto.fi; Kari.Halonen@aalto.fi).

Martin Andraud is with the Institute of Information and Communication Technologies, Electronics and Applied Mathematics ICTEAM, UCLouvain, 1348 Louvain-la-Neuve, Belgium (e-mail: Martin.Andraud@uclouvain.be).

Digital Object Identifier 10.1109/TCAD.2025.3586889

I. Introduction

IN TODAY'S technological landscape, there is a dramatic increase in AI-backed technologies and services, achieving significant milestones in automotive, health monitoring, wearables, and industrial applications. However, computing these AI-based solutions in embedded (or edge) applications requires power-efficient and reliable processors dedicated to AI tasks, called *AI accelerators* [1]. In that regard, most AI accelerators focus on deep learning, particularly deep neural networks (DNNs), which have become a de facto standard for most AI applications. A promising approach for efficient DNN acceleration relies on in-memory computing (IMC) architectures. IMC is particularly effective for DNNs, where multiply-and-accumulate (MAC) operations dominate the computational load, since it performs MAC operations directly in memory, using either digital or analog computing principles [2]. In that regard, analog IMC architectures use basic computation principles, relying on resistors or capacitors for multiplication, and current or charge sharing for accumulation [3]. As analog computing principles are more efficient than digital for low-resolution computation [4], analog IMC holds a significant potential for efficient DNN computation. Fig. 1 shows a typical IMC core composed of a matrix of (resistive) computing elements to accelerate DNN tasks. Inputs are converted into analog signals, and weights are stored as conductance values. Obtaining a variable conductance can be done using multiple resistive technologies, in CMOS or with emerging nonvolatile memories (eNVM) such as resistive RAMs (ReRAMs), Magnetic RAMs (MRAMs), or phase change memories (PCMs) [5], [6], [7]. At each crosspoint, inputs are multiplied by the conductance value, producing currents that are accumulated along the column. Subsequently, this current is converted into a digital signal by an ADC.

However, although multiple prototypes of resistive-based IMC cores have been presented, such implementations need to overcome various challenges to be used at a large-scale in a production line. Indeed, once embedded on the edge, accelerators must operate reliably across various temperature conditions, from internal device heat [8], [9] to external environmental shifts [10], [11]. In particular, temperature variations have been shown to negatively impact the accuracy of IMC cores, both locally at the circuit level [9], and globally for DNN inference [12], [13].

Thus, extensive research has been conducted to address the thermal sensitivity of resistive-based IMC systems using

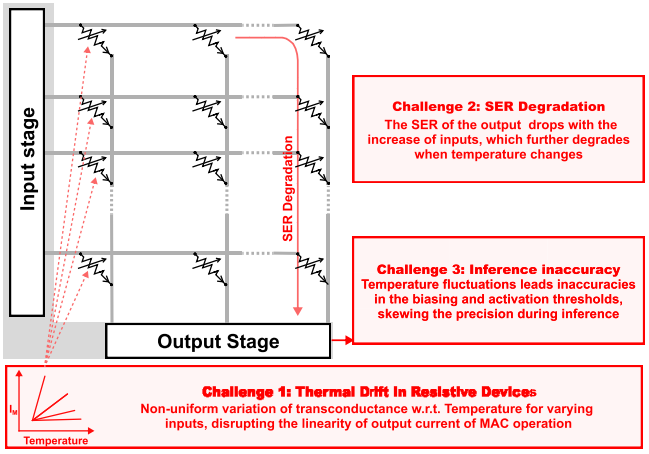


Fig. 1. Architecture of a resistive IMC core, highlighting the challenges faced due to temperature variation.

various approaches, e.g., [9], [10], [11], [12], [13], [14], [15], [16]. Compensation techniques span across model retraining for different temperatures [17], pairing computing tiles according to temperature [12], [16], modifying weights online [13], [18] or intervening directly at the hardware level by compensating the summed current at the bottom of the IMC column using various techniques [10], [14], [15], [16].

In this context, designing effective compensation techniques involves dealing with multiple challenges and complying with additional constraints. Regarding temperature variations, Fig. 1 illustrates the particular challenges of temperature variations in IMC cores, all intertwined: 1) the thermal drift of each resistive element, which causes errors accumulating in the bottom of the IMC column; 2) the degradation of the signal-to-error ratio (SER) due to increased variations in conductance; and 3) the accuracy degradation in DNN tasks during inference. Furthermore, temperature effects become more pronounced as the dimension of the DNN increases, challenging existing weight adjustment methods, e.g., [18], [19] and requiring more advanced techniques. Finally, any embedded compensation approach is designed with additional constraints, aiming to minimize overhead such as silicon area, power, time and additional design complexity. Hence, any compensation method should be effective while adding minimal overhead and complexity.

In light of these challenges, this work proposes a novel temperature compensation scheme for resistive-based IMC cores, based on a review of existing temperature compensation methodologies. This scheme aims to maintain reliable DNN inference across a wide temperature range while emphasizing simplicity and efficiency over complex training, algorithms, look-up tables (LUTs), power-intensive circuits, or intricate feedback and weight readjustment mechanisms typically present in state-of-the-art methods. As a result, the compensation overhead is greatly minimized while compensating for all first-order temperature variations. Our main contributions are listed as follows:

1) The proposed TRIM methodology is designed together with the output stage of the IMC core. It adjusts a regulating voltage dynamically with temperature, using a temperature-dependent voltage reference, and requires no additional hardware to be implemented.
2) The TRIM methodology is validated on a silicon proof-of-concept in 65 nm CMOS. Results show that TRIM enables the recovery of all first-order losses due to temperature from –20 °C to 60 °C.
3) The different self-compensation overheads are measured and compared with previous works. The proposed methodology requires an area overhead of 2% and power overhead of less than 0.02%.

The remainder of this work is organized as follows. Section II provides a literature review of existing temperature compensation methodologies and evaluates the remaining challenges. Section III details the system architecture of the baseline IMC system. Section IV details the TRIM compensation technique. Finally, Section V discusses the measurement results to demonstrate the TRIM methodology and compares these results with relevant state-of-the-art works.

II. Temperature Compensation of In-Memory Computing Devices

The large-scale integration of IMC cores using eNVMs faces challenges in reliability, robustness due to inherent variability, and endurance issues [20]. These challenges can significantly impact the accuracy of IMC cores, as well as the AI task running on these cores. There has been a large body of research work documenting the effect of temperature variations on SRAM-based IMC cores [15], resistive IMC architectures with high-density resistors [14] or emerging memories such as ReRAMs [16], [18] or PCMs [10]. First, this section explains the fundamental temperature-related challenges in IMC systems, reviews the state-of-the-art compensation methods, and outlines the remaining challenges, motivating our proposed approach.

A. General Considerations on Temperature Variations for IMC

Temperature-induced inaccuracies in IMC arrays can be categorized into three interconnected challenges, as shown in Fig. 1. These challenges span from device-level variations to system-level inference degradation, described below.

[Challenge 1—Thermal Drift in Resistive Devices (multiplication)]: Every resistive device used for MAC operations in the IMC array is sensitive to temperature. In a real computation scenario, fabrication variations will lead to a slightly different temperature dependence for each computing cell, introducing a small error in every multiplication operation. Depending on the technology used, the temperature sensitivity of the devices can vary. Hence, several models of temperature variations have been derived in the literature. A first method consists of considering first-order effects only, i.e., assuming a linear temperature dependence as a function of ambient temperature (T), given by

$$R(T) = R_0[1 + \alpha(T - T_R)] \quad (1)$$

where R_0 is the resistor's nominal value at room temperature T_R, and α is the temperature coefficient (TC) of resistance R. For CMOS-based IMC arrays, for instance, relying on SRAMs, resistance typically follows a well-defined temperature dependence. Hence, a linear compensation is typically sufficient [15], [21], [22]. However, certain resistor technologies exhibit nonlinear temperature dependencies. For instance, specific eNVM technologies like PCMs are known to have a strong nonlinear temperature dependence [10]. In this case, the temperature variations are modeled with an exponential or quadratic behavior. For exponential dependence, the resistance is given by

$$R(T) = R_0 \exp\left(\frac{K}{T}\right) \tag{2}$$

where K is a constant, and T is the absolute temperature. Similarly, a quadratic dependence can be represented as

$$R(T) = \alpha_2 T^2 + \alpha_1 T + \alpha_0 \tag{3}$$

where α_2, α_1, and α_0 are coefficients. Considering this nonlinear dependence, recent compensation methodologies also account for nonlinear effects, particularly in eNVM technologies [9], [10], [17], [18], [19], [20].

Yet, in terms of hardware, nonlinear compensation schemes often introduce significant implementation challenges. They may require additional circuit complexity, an increased computation overhead, and a precise characterization of higher-order dependencies, which may not always be practical in real-time IMC systems. Furthermore, for moderate temperature variations, the added precision of nonlinear compensation may not justify the associated increase in hardware and energy costs. Given these constraints, many IMC architectures can achieve acceptable compensation using first-order compensation, as a balance between accuracy and implementation feasibility. This is also motivated by the fact that, within practical operating ranges, temperature-induced resistance variations often exhibit a dominant linear trend.

[Challenge 2—SER Degradation (Accumulation)]: Accumulated temperature-induced resistance variations degrade the SER at the IMC column-level, defined as

$$\text{SER(dB)} = 10 \cdot \log_{10}\left(\frac{\sigma^2_{I_{RT}}}{\sigma^2_{\Delta I_T}}\right) \tag{4}$$

where $\sigma^2_{I_{RT}}$ is the output current variance at room temperature, and $\sigma^2_{\Delta I_T}$ is the variance induced by temperature deviations. Together with its dependence on input values, the SER degrades with temperature, as each MAC level variation will increase the overall error observed at the bottom of the column—the point where all currents from a crossbar column are summed. That is why a common technique is to compensate the current at the bottom of the column to reduce temperature variations [10], [14], [15].

(Challenge 3—DNN Inference Inaccuracy): The local impact of temperature on MAC operations can translate to significant inaccuracies at the model level. Our previous silicon experiments in a temperature-controlled chamber [13] showed that the accuracy of DNN decreases drastically outside a temperature range of 20 °C to 40 °C. For instance, on a digit classification task (MNIST), the accuracy can be degraded by 50%–70% from the baseline at 0 °C and 60 °C. This legitimizes the efforts toward innovative circuit designs and system architectures to improve the IMC's reliability.

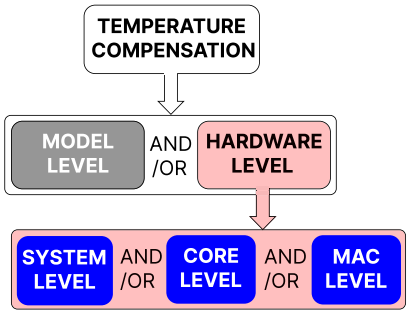


Fig. 2. Classification between different temperature compensation methodologies in the literature.

B. Existing Temperature Compensation Methods for IMC

Compensation methodologies differ in their approach, effectiveness, and adaptability to various memory technologies; therefore, a comparison is needed to assess their feasibility across different IMC core implementations. A summary of various compensation methodologies recently presented in the literature is listed in Table I. The methodologies are compared using the temperature compensation type (linear or nonlinear), the type of temperature dependence used, the type of memory (including SRAM), and according to the level of compensation, as explained in Fig. 2 and detailed in this section. Regarding the compensation level, a distinction is made between *model-level* compensation, where only the model is modified, and *hardware-level* compensation methods, where the hardware is designed to be less sensitive to thermal variation.

C. Model-Level Compensation

Typically, model-level compensation aims to limit any degradation in DNN accuracy due to temperature effects by directly retraining the DNN model according to different temperature parameters, without making any hardware changes. Yet, retraining the complete model would induce storing several sets of weights on-chip, which is infeasible in practice due to memory constraints. Instead, it has been proposed to update only the batch normalization parameters [17]. In DNN inference, batch normalization aims at rescaling DNN values in between two layers for better convergence. Here, several batch normalization parameters are learned, stored on-chip and used according to the current temperature. Finding the parameters is done using a knowledge distillation framework, injecting noise and variations in the initial model at different temperatures to train a set of smaller models, each with different normalization parameters. This technique maintains the same DNN accuracy across temperatures ranging from 25 °C to 55 °C.

TABLE I
COMPARISON OF TEMPERATURE COMPENSATION METHODS FOR IN MEMORY TECHNOLOGIES

REF	Level	Comp. Type	Memory	Element $\propto$	Efficiency	Compensation technique
[17]	Model	Non-linear	ReRAM	$e^{K/T}$ **(1)**	$15.89\% - 99\%$ **(3)**	PKD and BNA adapt the model for temperature variations, improving accuracy and efficiency.
[23]	System	Non-linear	Memristor	$e^{K/T}$ **(1)**	86.6% **(3)**	Uses global compensation & optimally maps matrix values to memristor conductances using s/w algorithms and peripheral circuit tuning.
[9]	System	Non-linear	ReRAM	$e^{K/T}$ **(1)**	$4.8\% - 58\%$	Uses system-wide task reordering. Temp.-aware adjustment, dynamic remapping, resilient training.
[19]	System	Non-linear	ReRAM	$e^{K/T}$ **(1)**	50% **(3)**	WRAP framework remaps critical weights to cooler subarrays, while WC shifts weights to lower conductance levels.
[20]	System	Non-linear	ReRAM	$e^{K/T}$ **(1)**	up to 95% **(3)**	THOR delays hot bank accesses and optimizes cache management to reduce temperatures and enhance lifetime.
[15]	Core	Linear	SRAM	T	$145\% - 154\%$	On-chip compensation stabilizes MAC operations by dynamically adjusting wordline voltage across temperatures.
[11]	Core	Linear	ReRAM	T **(2)**	87.5% **(3)**	RTC compensates resistance deviations, adjusts current amplification, and stabilizes weight mapping under temperature variation.
[21]	Core	Linear	NOR Flash	T	95% **(3)**	The compensation circuit adjusts conductance, threshold voltage, and coefficients to stabilize output voltage.
[14]	Core	Linear	SRAM w/ R-2R DAC	$e^{K/T}$ **(1)**	-	Modifies output current wrt temperature where the compensation current is proportional to change in temperature.
[13]	MAC	Linear	SRAM with R-2R DAC	$e^{K/T}$ **(1)**	100%-104%	Uses dynamic SRAM weight update for compensation, using on-chip temperature sensing and LUT characterization
[22]	MAC	Linear	NOR Flash	T **(2)**	-	Differential multiplier, gate coupling, and weight adjustment minimize temperature drift in computations.
[10]	MAC	Non-linear	PCM	$Te^{K/T}$	$46.7\% - 80\%$ **(3)**	First-order adjusts linearly, second-order uses nonlinear correction for enhanced PCM accuracy compensation.
[18]	MAC	Non-linear	ReRAM	$e^{K/T}$ **(1)**	4.8%-58%	HR3AM improves accuracy through bitwidth downgrading and thermal optimization via tile pairing.

(1) Assumption based on state conductance or referenced equations from corresponding literature; likely follows an Arrhenius-type relationship.
(2) Extracted from Figures/Tables in the corresponding literature.
(3) Estimated from results in the corresponding literature.

Generally, model-based applications are effective, but require an additional training overhead that can be long and complex. For instance, the knowledge distillation framework of [17], or a "chip-in-the-loop" infrastructure for training, where the model is tailored to each particular chip, requires a large infrastructure and dedicated retraining for each device.

D. Hardware-Level Compensation

Hardware compensation mechanisms add dedicated circuitry to calibrate the system independently of the AI task being run. Depending on where they intervene in the computing architecture, they can be decomposed into three sublevels, namely at the system, core, or MAC level.

1) System-Level Adaptation: System-level intends to preemptively reordering data, weights, or compute tasks in the system according to the conditions experienced by the hardware. They typically target a multicore IMC system, without modifying the IMC hardware itself. For instance, it is possible to assign computations based on the temperature of each processing core to reduce thermal variations [12], [16], [18]. Alternatively, IMC tiles can be dynamically paired according to their instantaneous temperature [18]. It ensures that the assigned task is not overheating, which may decrease the accuracy. This dynamic pairing can also happen at ReRAM cell-level [12], mapping larger weights to "colder" cells and smaller weights to "hotter" cells as much as possible to limit temperature variations, reordering weight rows in the array. A more fine-grained strategy acts at three levels in a ReRAM-based IMC system [16]: 1) decompose weights between positive and negative IMC arrays to minimize the temperature effects; 2) reorder the task on different IMC columns to limit the column-to-column variations; and 3) perform a fine-grained weight decomposition to further compensate remaining variations. The main downside of system-based compensation is that it can induce significant system overheads. For instance, the approach in [18] requires several reserve tiles, "cooling down" when necessary. In addition, catching local temperature changes may require additional compensation mechanisms [16], [18].

2) MAC-Level Adaptation: While system-based or core-based methods effectively manage workloads, they need to be paired with local adaptation strategies to comprehensively tackle temperature variations [16]. In this case, it is possible to intervene *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. For instance, the weight of each resistive element (or conductance) in the IMC core can be adjusted according to temperature [13], [18]. One possibility is to downgrade

the weights' bit-width of a ReRAM crossbar according to temperature [18], to map weights to an available conductance value, instead of keeping a misrepresented conductance, introducing errors. To avoid reducing the resolution, the weights can be directly rewritten according to an online temperature measurement [13]. In that case, the architecture should allow efficient on-chip weight storage, for instance, in SRAM cells.

Yet, MAC-level adaptation, like downgrading or rewriting weights, can include a significant hardware overhead. Downgrading bits [16] can reduce the effective resolution of the system. Changing individual weights [13] leads to a complete compensation of temperature variations from –20 °C to 60 °C, yet it applies mostly for SRAM-based architectures and not to any IMC core. Moreover, the writing process of eNVMs can be complex and lead to timing and power overheads.

3) Core-Level Adaptation: As a compromise between local and global adaptation, the compensation can happen at core level, typically at IMC column level. This is interesting as IMC hardware typically duplicates the same hardware for multiple columns. For that, a first possibility is to use one additional "dummy" column of the IMC core for temperature compensation [15]. In this case, the weights of the dummy columns can be chosen offline to compensate for temperature variations. A second approach consists of converting the MAC results at the bottom of the column and compensating for temperature variations with a dedicated digital block [10]. The compensation values are calculated with an analytical model built from reference devices and analytical equations, and allow compensation for first-order temperature effects. Alternatively, circuitry can be added to compensate for temperature-induced current offsets at the bottom of each IMC column [14], [16]. The compensation can be done with current mirrors [16] whose ratio is chosen with temperature, or a current generator with a variable TC [14] chosen to compensate for the TC of the column.

E. Summary of Challenges and Proposed Method

Although offering a good compromise, core-level methods at column-level have shortcomings. For instance, [18] requires precise current ratios for each column, while having only discrete compensation steps. Monga et al. [14] required an off-chip resistor to change the TC of the compensation, complicating the final integration of the system. Zang et al. [15] targeted only charge-based IMC (although the technique could be extended). Reference [10] is among the only methods that explicitly compensate for second-order effects, yet requiring a heavier offline characterization. Hence, this work proposes to solve current challenges of core-level methods by developing TRIM, a column-based self-compensation methodology requiring no external components, precise current ratios, or heavy offline characterization. Additionally, TRIM limits the compensation overheads by integrating the compensation loop directly into the IMC column output stage peripheral, which typically includes a summing amplifier. Finally, in specific nonvolatile memories like ReRAM, on-chip temperature variations, particularly hotspot formation, have a more pronounced effect than ambient temperature changes [5], [6], [12], [16], [20]. TRIM can be deployed as a matrix to sense and compensate for local processing elements with respect to hotspots. Since hotspots form dynamically depending on computation patterns, this method enables localized compensation without requiring predefined hotspot locations. While traditional approaches introduce significant overhead when applied at a fine-grained level, the compact size and low-power requirements of the implemented compensation logic allow for efficient deployment when compared to other compensation techniques.

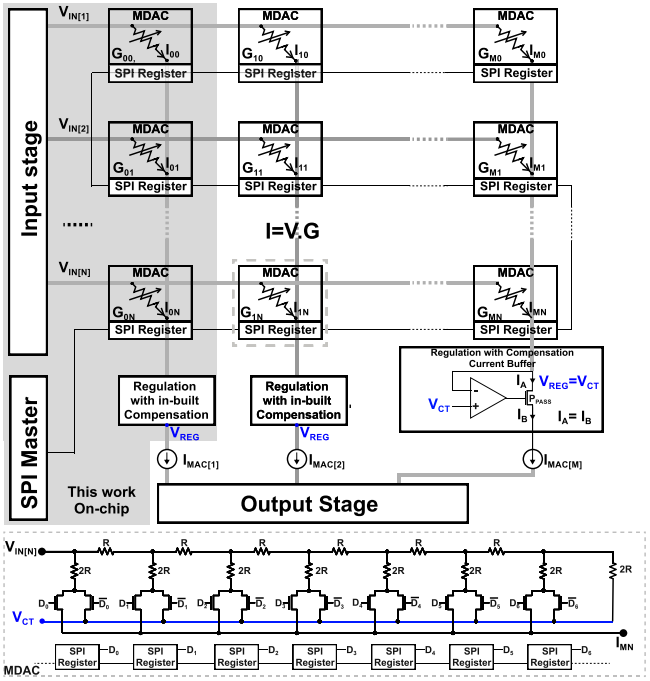


Fig. 3. System architecture of the demonstrated proof-of-concept resistive IMC core, and the proposed TRIM auto-compensation.

III. Baseline IMC System

Before introducing the TRIM method, we detail the baseline resistive-based IMC system designed in a 65 nm CMOS technology.

A. System Architecture

Fig. 3 illustrates the baseline resistive IMC core, in the form of a $N \times M$ crossbar, and the associated peripheral circuits. At the input stage, digital data is first converted to analog voltages ($V_{IN[1]}$ to $V_{IN[N]}$) by DACs. A sample-and-hold circuits ensure these analog inputs remain stable and precisely timed throughout inference. Each crosspoint in the array is implemented as an R-2R multiplying digital-to-analog converter (MDAC) weight cell, allowing configurable multibit precision for weight representation.

In this prototype, CMOS poly-resistors form the resistive elements. However, they can be replaced with eNVM technologies with a suitable adaptation of the peripheral circuits. While the overall structure remains similar, practical eNVM implementations require adapted peripheral circuits to handle

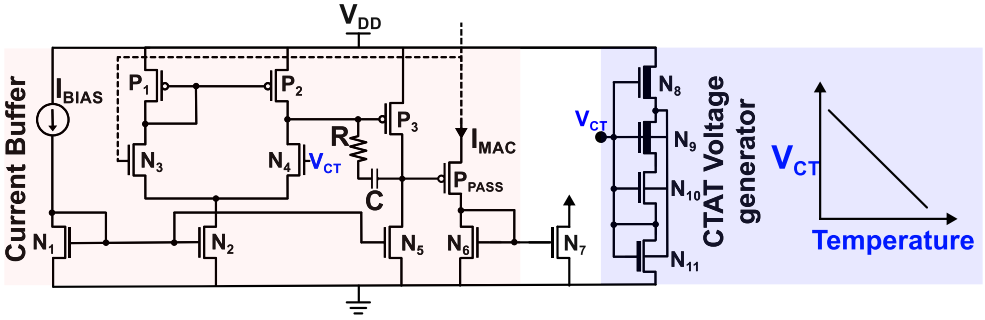


Fig. 4. Schematic of the integrated CTAT voltage reference generator and summation node regulator used for dynamic thermal compensation.

different operating and sensing conditions. After performing multiplications within the array, the resulting currents from each column are summed at a regulated node voltage (V_{REG}), stabilized by a voltage regulator before being digitized by an ADC.

B. Multiply and Accumulate Operation

1) Analog Multiplication: Each MDAC cell, shown in Fig. 3, performs multiplication by converting a digital weight into an equivalent conductance, and multiplying this conductance by the analog input voltage. Digital weights (D_0–D_6) are provided by an on-chip SPI interface, although in practical implementations, they could also be stored in SRAM cells.

The MAC output current for an N_B-bit and N number of MDAC cells at temperature T is expressed by

$$I_{\mathrm{MAC}}(T) = \sum_{i=1}^{N} \frac{V_{IN[i]} - V_{\mathrm{CT}}(T)}{R(T)} \cdot \frac{D_{[i]}}{2^{N_B}} \tag{5}$$

where $R(T)$ represents the temperature-dependent resistance of each MDAC element, V_{IN} is the analog input voltage, and $V_{\mathrm{CT}}(T)$ is the MDAC's zero reference node voltage, and D is the weight value (in integer format). Equation (5) shows how the precision of each MAC operation is influenced by temperature variations.

2) Analog Accumulation: The accumulation step sums the currents from all MDAC cells in every column ($I_{MAC1:M}$) at node V_{REG}. This node is held at the complementary-to-absolute-temperature (CTAT) reference voltage V_{CT} by an op-amp in negative-feedback, as shown in Fig. 3 and detailed in Fig. 4. Under feedback, the two-stage op-amp drives the gate of a pMOS pass transistor P_{PASS}, with its source connected to V_{REG} (N_3 gate, the op-amp's negative input) and its positive input (N_4 gate) fixed at V_{CT}. By continuously adjusting the gate voltage, the op-amp enforces $V_{REG} = V_{CT}$, making the pMOS source node a *virtual reference* with very low impedance. At the same time, any I_{MAC} current sinking into the P_{PASS} source is passed through and appears at its drain. Because the pMOS output current exactly tracks the input current under unity-gain feedback, it functions as a "current-controlled current source buffer" with high output impedance. This ensures that the summation node remains at a stable voltage while delivering I_{MAC} unchanged to the next stage, preserving linearity and accuracy across the crossbar. The compensated current output can be digitized by an output stage, such as an ADC, through an I-to-V conversion stage. Since the proposed compensation adjusts only offset voltages without affecting linearity, a PVT-invariant ADC can directly digitize the compensated output without dynamic reference adjustments.

IV. Proposed TRIM Compensation

A. Impact of Thermal Variations on IMC Arrays

In resistive-based IMC arrays, thermal variations directly influence the accuracy of MAC operations. As described in Section II, the resistive elements typically exhibit temperature-dependent behavior. Depending on the resistor's material, this temperature dependence can either be proportional-to-absolute-temperature (PTAT) or CTAT. In this work, CMOS polysilicon resistors exhibit a PTAT characteristic, where resistance increases linearly with temperature. The temperature-induced changes in the resistors' conductance cause an offset in the MAC results of the crossbar. Assuming a linear temperature dependence, the MDAC resistor values vary with ambient temperature as described in (1).

B. Proposed Compensation Technique: Theory

To offset the temperature-dependent errors introduced by PTAT resistors, the TRIM compensation technique employs a thermally adaptive CTAT voltage reference. As illustrated in Fig. 4, a dynamically varying voltage (V_{CT}) exhibiting CTAT behavior is integrated into both the MDAC zero-reference node and the column's regulated summation node. By dynamically adjusting this node, TRIM directly compensates the resistance-induced thermal variations in the IMC array.

1) Generation of the Compensation Voltage: The CTAT voltage (V_{CT}), can be approximated by first-order relationship

$$V_{\mathrm{CT}}(T) = K_{\mathrm{CT}} - \beta(T - T_{\mathrm{CT}}) \tag{6}$$

where K_{CT} is the intercept of the CTAT at a reference temperature T_{CT} and β represents the TC CTAT voltage. Fig. 4 shows this CTAT voltage variation w.r.t. temperature. By combining equations (1), (5), and (6) and differentiating I_{OUT} such that, $dI_{OUT}/dT = 0$, a condition for temperature stability is derived, allowing for temperature-induced resistance variations to be

compensated by appropriately adjusting β relative to α, and can be derived as

$$\beta = \frac{\alpha(V_{\text{IN}} - V_{\text{CT}})}{1 + \alpha T_{\text{CT}}}. \tag{7}$$

Since α is fixed for a particular type of resistance, and considering that the input voltage V_{IN} is within a fixed range, adjusting β appropriately relative to the TC α of resistance, it becomes feasible to compensate for temperature-induced resistance variations. As it can be seen in (7), β is independent of array size N (in (5)) or individual weight values; hence, this scheme can be used for virtually any IMC array size. This result is further confirmed through simulations with larger arrays, as discussed in the next section.

For small temperature variations around a reference temperature T_0, higher-order effects in temperature dependencies can be approximated as linear, allowing β to be expressed in a simplified form. For exponential dependence in (2), the compensation coefficient β is calculated as

$$\beta = \frac{(V_{\text{IN}} - V_{\text{CT}})K}{k_B T^2}. \tag{8}$$

Similarly, for resistors with a quadratic dependence, it is obtained

$$\beta = \frac{(V_{IN} - V_{\text{CT}})(2\alpha_2 T + \alpha_1)}{\alpha_2 T^2 + \alpha_1 T + \alpha_0}. \tag{9}$$

To simplify the compensation model for small deviations in nonlinear models around a reference temperature T_0, a first-order Taylor series expansion is applied to β in (8) and (9). For exponential dependence in (2), expanding β around T_0 using a first-order approximation gives

$$\beta(T) \approx \beta_0 + \frac{d\beta}{dT}\bigg|T = T_0(T - T_0) \tag{10}$$

where β_0 represents the compensation coefficient at T_0. Substituting from (2) and differentiating

$$\beta_0 = \frac{(V\,\text{IN} - V_{\text{CT}})K}{k_B T_0^2} \tag{11}$$

$$\frac{d\beta}{dT} = -\frac{2(V_{\text{IN}} - V_{\text{CT}})K}{k_B T_0^3}. \tag{12}$$

Thus, the linearized expression for $\beta(T)$ for exponential dependence is

$$\beta(T) \approx \frac{(V_{\text{IN}} - V_{\text{CT}})K}{k_B T_0^2}\left(1 - \frac{2(T - T_0)}{T_0}\right). \tag{13}$$

Similarly, for the quadratic dependence in (3), applying a first-order Taylor expansion at T_0. The resulting linearized form of $\beta(T)$ for the quadratic dependence is

$$\beta(T) \approx \frac{(V_{\text{IN}} - V_{\text{CT}})}{\alpha_2 T_0^2 + \alpha_1 T_0 + \alpha_0}\Bigg[(2\alpha_2 T_0 + \alpha_1) + \frac{2\alpha_2(\alpha_2 T_0^2 + \alpha_1 T_0 + \alpha_0) - (2\alpha_2 T_0 + \alpha_1)^2}{\alpha_2 T_0^2 + \alpha_1 T_0 + \alpha_0}\Bigg](T - T_0). \tag{14}$$

Using this approximation ensures that $\beta(T)$ exhibits a linear dependence with T, within a moderate temperature range even for nonlinear devices, simplifying the compensation mechanism. Although this approach reduces implementation complexity, higher-order temperature dependencies may introduce residual effects beyond the linear approximation. Therefore, the compensation can potentially account for second-order effects, making it applicable under moderate temperature variations.

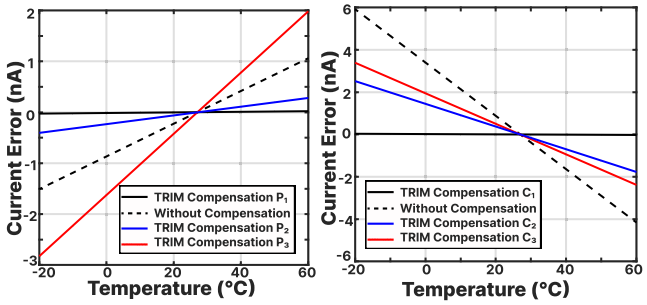


Fig. 5. I_{MAC} error for PTAT and CTAT resistors. The voltage generator adjusts the slope to counteract thermal variations.

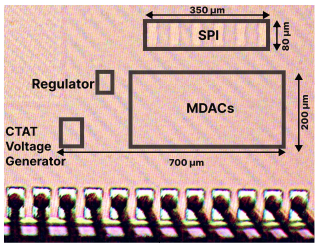
2) Reference Voltage Generation: A CTAT voltage V_{CT} is generated by an all-nMOS 4-transistor voltage reference generator, adjusting the current buffer's reference voltage and the MDAC's zero reference node dynamically, as shown in Fig. 4. This design, deriving insights from previous studies [24], [25], [26], employs nMOS devices with varying threshold voltages–zero threshold (N_8, N_9), low voltage threshold (N_{10}), and high voltage threshold (N_{11})–to produce a thermally sensitive voltage output controllable by aspect ratio and threshold voltage differences. The bodies of N_9, N_{10}, and N_{11} are connected to N_8's drain to reduce process variations. N_8 and N_9 are designed to address sensitivity to power supply variation V_{DD}, whereas N_{10} and N_{11} are tasked with generating the CTAT voltage, V_{CT}, described by

$$V_{\text{CT}}(T) = \underbrace{V_{TH_{11}} - \frac{m_{11}}{m_{10}} V_{TH_{10}}}_{\text{CTAT}} + \underbrace{m_{11} V_T \ln\left(\frac{K_{10} S_{10}(m_{10} - 1)}{K_{11} S_{11}(m_{11} - 1)}\right)}_{\text{PTAT}} \tag{15}$$

where V_{TH_i} and m_i denote the threshold voltage and subthreshold slope of transistor i (i.e., $i = 10, 11$), respectively; V_T represents the thermal voltage; $K_i = \mu C_{\text{OX}_i}$ (with μ as carrier mobility and C_{OX} as oxide capacitance); and S_i indicates its aspect ratio. By tailoring the S_i of M_{10} and M_{11}, the circuit's CTAT voltage characteristic can be optimized, as the first two terms of (15) are CTAT and the last term is PTAT. Therefore, the same circuit can also be used to compensate for CTAT resistors by generating a PTAT voltage, showing the versatility of the proposed compensation technique.

3) Aging Effects: Effects such as bias temperature instability and channel hot carrier effects cause aging in the circuit, for instance, shifting in threshold voltages over time [27]. To limit aging effects, the output of the CTAT generator is chosen to depend on the difference between two threshold voltages, instead of their absolute values, as shown in (15). This, in turn, provides more robustness against ageing-induced shifts, ensuring a better long-term stability of the reference voltage.

4) Simulation Results of the Proposed Compensation: Fig. 5 shows the simulated thermal behaviors of PTAT and CTAT resistors, illustrating how the compensation circuit can



Design Summary	
Technology	65nm
Weight/ Max. - Transconductance	7b/ 500 µS
Test Column size	10 MDACs
Compensation area overhead*	2%
Compensation power overhead*	<0.02%
*Overhead comes from the CTAT generation circuit	

Fig. 6. Microphotograph and a summary of the proof-of-concept system.

be applied to both types of resistors, demonstrating its versatility. The compensation circuit is based on (7), which adjusts the slope of the output current to achieve a thermally stable output. The black line represents the compensation achieved, effectively minimizing thermal-induced errors. The dotted line shows the error without compensation, where thermal variations cause significant current deviations. By adjusting the slope of the output current using different compensation settings (P_1-P_3 for PTAT and C_1-C_3 for CTAT), the voltage generator counteracts these variations, ensuring that the system can maintain stable performance over a wide temperature range.

V. Measurement Results and Discussion

To demonstrate the effectiveness of the TRIM inference scenario, an IMC system capable of performing multiple MAC operations in a 65 nm bulk CMOS technology is implemented in this work. Fig. 6 shows the design summary and the implemented system's microphotograph, composed of a column of 10 MDACs, an on-chip SPI, current buffers, and the CTAT voltage generator. The on-chip SPI is interfaced with an Arduino Due, 7-bit quantized inputs, weights, and biases from the LeNet-5 model trained on the MNIST dataset. Measurements were conducted on a custom PCB across –20 °C to 60 °C in Weisstechnik Labevent temperature chamber, employing a N6705C DC Power Analyzer and Keithley 6514 Electrometer for input and output monitoring. The model's inference accuracy is evaluated with 10,000 MNIST samples using a Python model for the MDAC column. Output currents, normalized to a 0-1 range, are stored in an LUT, which the Python model uses to determine convolution layer outputs. The subsequent sections present the measurement results, focusing on each previously identified challenges (Section II). It should be noted that the system is used to demonstrate the proposed TRIM methodology for linear temperature compensation, as first-order temperature dependencies are dominant in this architecture.

A. CTAT Voltage Generation

The CTAT voltage generator consumes power of 10 nW at 0.8 V supply and is functional between 0.6-1.3 V, outputs a voltage with 1850 ppm/°C TC, resulting in voltages of 600 mV at room temperature and 605 mV and 595 mV at –20 °C and 60 °C, respectively.

To evaluate the robustness of the proposed CTAT voltage generator under process variation, Monte Carlo simulations

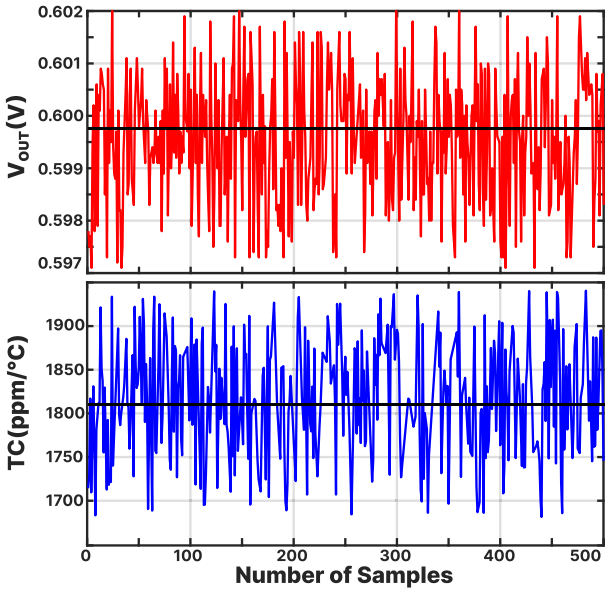


Fig. 7. Monte Carlo simulation of 500 samples of the integrated CTAT voltage generator for thermal compensation.

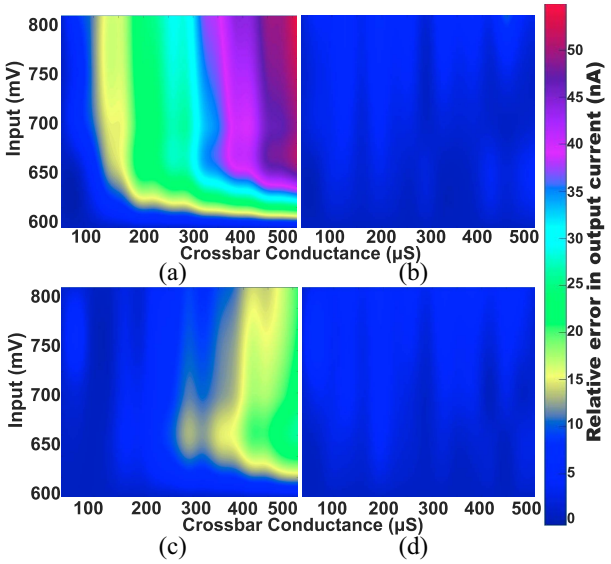


Fig. 8. Comparison of measured compensated versus uncompensated column outputs: output current error w.r.t to room temperature for varying input vectors across increasing crossbar conductance at extreme temperatures. (a) Uncompensated −20°C. (b) Compensated −20°C. (c) Uncompensated 60°C (d) Compensated 60°C

for 500 samples were performed. Fig. 7 depicts the variation in terms of output voltage (V_{OUT}) and TC. It can be seen that (V_{OUT} exhibits a mean value of 0.5995 V with a standard deviation of 0.002 V, corresponding to a $3\sigma/\mu$ process variation of 0.60%. The TC distribution across samples yielded a mean of 1816.08 ppm/°C with a standard deviation of 68.08 ppm/°C. validating the circuit's stability under simulated mismatch and variation scenarios.

B. Mitigating Thermal Drift in Resistive Devices

1) General Temperature Compensation: Fig. 8 illustrates the output current error's response to voltage changes across

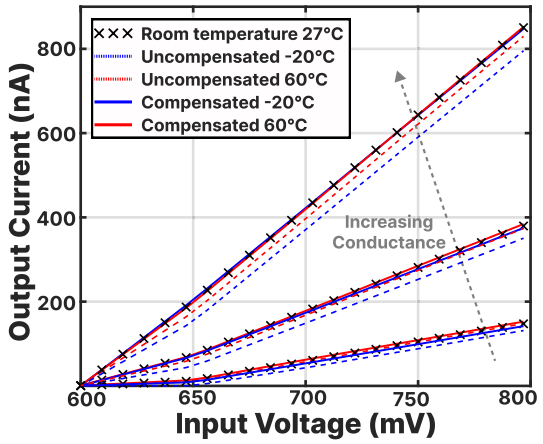


Fig. 9. MDAC output current variation across input voltages and temperatures, demonstrating the effect of increasing conductance.

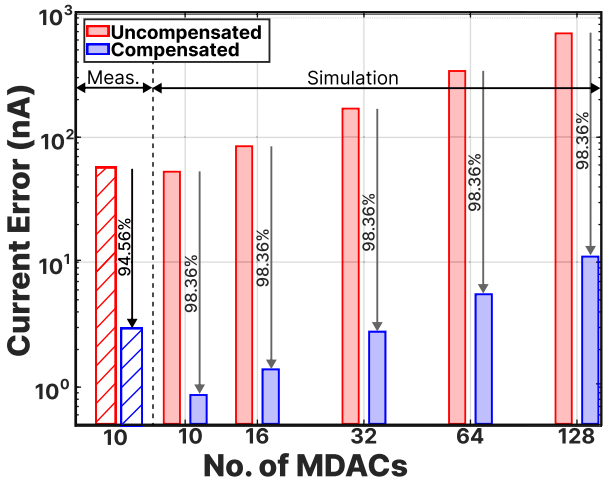


Fig. 10. Impact of increasing MDAC count on I_{MAC} error with and without compensation under extreme temperature variation relative to room temperature.

different conductance values in the resistive crossbar. At –20 °C, uncompensated outputs [Fig. 8(a)] show broad error up to 50 nA, with a wide color variation indicating substantial variability. The compensation significantly reduces the error dispersion, depicted by a more uniform color distribution [Fig. 8(b)], pointing to errors approaching near-zero levels. Similarly, at 60 °C, variability in uncompensated outputs [Fig. 8(c)] is markedly reduced in compensated conditions [Fig. 8(d)], demonstrating the compensation scheme's role in enhancing output current precision and stability across temperature and conductance levels.

2) Output Current Compensation: Fig. 9 extends the analysis of the temperature compensation effect on output current for individual conductance states, building on the overall improvements seen in previous results. The dashed lines represent uncompensated conditions at –20 °C and 60 °C, deviating noticeably from the room temperature baseline. The current offset due to temperature can increase up to 55 nA. After compensation is applied, the correction depicted by the solid lines aligns much closer to the room temperature. These results

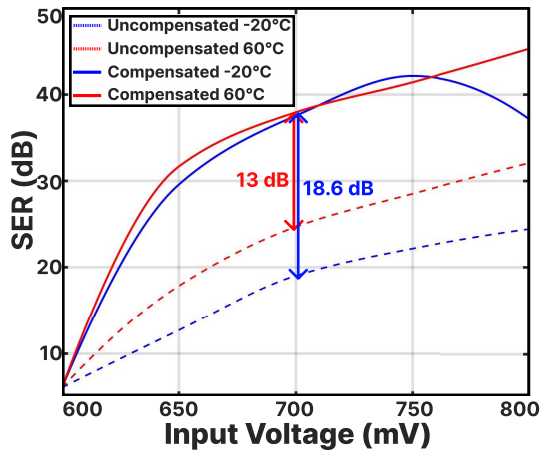


Fig. 11. SER variation with the input voltage, showcasing the effectiveness of compensation in mitigating temperature-induced variability.

demonstrate the effectiveness of the compensation scheme in reducing the thermal-induced error, thereby enhancing the precision and stability of output currents across a wide range of temperatures and conductance levels.

3) Extension of the Methodology for Bigger Arrays: Further, Fig. 10 shows simulation results for different column sizes. As can be seen, the current error becomes more pronounced in larger arrays (e.g., 128 MDAC cells), due to the accumulation of temperature-induced variations across all MAC units. Despite this, the proposed compensation technique consistently suppresses the thermal errors, maintaining an average reduction of approximately 98.36% across all evaluated sizes. These simulation results are also consistent with the measurement results obtained for the 10-MDAC case, where a 94.56% compensation efficiency was observed.

C. Improving SER Degradation

Fig. 11 shows the SER across varying input voltages. The output at room temperature is assumed to be the reference output. The error related to changes in output current with respect to temperature variations is measured according to (4). The measurements show that mid-input voltage of 700 mV, the compensated output achieves an SER increase up to 18.6 dB at –20 °C and 13 dB at 60 °C, demonstrating a substantial reduction in temperature-induced error. Higher SER values indicate an improvement in reducing temperature-induced errors. This results in a significant reduction in temperature-related errors and an overall enhancement in system accuracy.

D. Recovery of MAC Operation

Fig. 12 shows the MAC operation error distribution at column-level. The mean error reduces from 9.33% to –0.81% and the standard deviation from 17.73% to 10.41%, overall achieving a 91.3% improvement in MAC operation. The compensated output's errors are notably concentrated near the zero-error level, demonstrating that the temperature compensation circuit significantly reduces MAC operation deviations,

TABLE II
PERFORMANCE SUMMARY AND COMPARISON

		This Work		[11]		[18]		[15]		[13]	
	Technology (nm)	65		40		-		40		65	
	Topology	Current R-2R		ReRAM		ReRAM		SRAM		R-2R MDAC + 6T-SRAM	
	Validation Type	Measurement		Measurement		Measurement		Simulation		Measurement	
	Weight (bits)	7b		3b		2b		1b		7b	
	DNN Dataset	MNIST		CIFAR 10		MNIST		MNIST		MNIST	
	Temperature (°C)	-20	60	41.85	249.85	26.85	86.85	-40	120	-20	60
	SER Improvement @ mid V_{IN} (dB)	13	18.6	-	-	-	-	-	-	-	-
DNN Accuracy	**Baseline at 27°C(%)**	85		*		*		96.14		85	
	Before Compensation(%)	36	27	∼88	∼10	90	35	79.60	80.81	36	27
	After Compensation(%)	85	85	∼87	∼85	95	90	96.14	96.14	85	89
	Recovery(%)	100	100	-	-	-	-	100	100	100	104

* Relative Accuracy

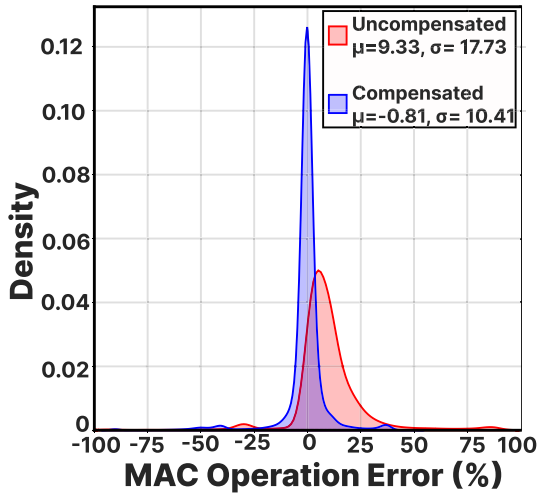


Fig. 12. Statistical distribution of measured MAC operation errors.

TABLE III
OVERHEAD COMPARISON WITH OUR PREVIOUS WORK

	This Work	**[14]**	**[13]**
Power	10nW	96 nW	14 μW
Area overhead	2%	0.5% and off-chip resistor	-
Latency overhead	Minimal with automatic control	Manual control, more time due to SPI adjustments	Moderate (involves recalibration with LUT)
Complexity	4 MOSFETs	11 MOSFETs, 1 resistor, SPI	Mixed-signal system with ADC and FSM

* Overheads calculated for a 10 MDAC column, with operating voltage of 1.2V

thus maintaining computational accuracy under diverse thermal conditions.

E. Recovery of Inference Accuracy

Inference accuracy was tested using the LeNet-5 model on the MNIST dataset with measurements done under varying temperatures. Here, it should be noted that the LeNet-5 architecture from [28] is implemented exactly with the same feature map, which, after quantization, had an accuracy of 85% when tested. While this accuracy is lower than some state of the art MNIST implementations, the primary focus is on analyzing the impact of temperature variations and the effectiveness of the compensation scheme. LeNet-5 is selected as a test case suitable for repeated simulations with different mappings and temperature conditions. The compensation techniques change the output of the MDACs prior to quantization or ADC sampling, independent of any model used. Therefore, any trained classifier, regardless of complexity, can operate on the compensated outputs.

The reported accuracy serves as a reference point to quantify the degradation and subsequent recovery, rather than as a benchmark for absolute performance. Without compensation, as shown in Table II, the deviations from the baseline temperature of 27 °C lead to 36% and 27% decrease in accuracy for extreme positive and negative temperature shifts, respectively. After applying the compensation, the accuracy is fully restored to the baseline level of 85%, demonstrating the effectiveness of the compensation in recovering performance over a wide range of temperatures.

F. Comparison and Overhead Evaluation

Table II compares the proposed system's performance designed for 7-bit weight, against state-of-the-art works, showing the accuracy improvements with thermal variations for extreme temperatures at –20 °C and 60 °C with compensation. The table shows how compensation elevates the inference accuracy back to its original value (100% recovery) across temperatures, showcasing the system's enhanced thermal robustness.

Another important point is to evaluate the compensation overheads, in terms of area, latency or power. For a fair comparison, Table III shows key differences between this work and our previous implementations of temperature compensation techniques, all demonstrated on a 10-column MDAC array. The proposed compensation block takes an area and power overhead of only 2% and 0.02%, respectively.

1) Power Overhead: The measured power overhead of TRIM remains constant even as the size of the IMC array increases, hence resulting in a decreasing relative overhead as

the system is scaled to 32, 64 or 128 MDACs per column. It arises primarily from the CTAT voltage generator, introducing only a 10 nW power overhead, which is significantly lower than the 96 nW reported in [14] and the 14 μW in [13].

2) Area Overhead: The 2% overhead of TRIM is slightly higher than the 0.5 reported in [14], yet removing the need for a large off-chip resistor. By integrating all components on-chip, this design avoids the external complexities of off-chip components, ensuring better system compactness. If the system is scaled to higher MDACs per column, the area overhead becomes increasingly negligible in proportion to the total system area.

3) Latency Overhead: Latency is another critical factor in temperature compensation systems, particularly for real-time applications. For that, TRIM features an automatic control mechanism, ensuring minimal time overhead. This contrasts with the manual control method in [14], which relies on SPI adjustments, leading to delays due to frequent changes. In [13], the latency overhead is moderate because recalibration processes involve updating weights using an LUT every once in a while. However, in a pipeline-based DNN/CNN implementation, different weights would need to be updated and recalibrated, leading to significant latency in the compute cycle. In contrast, for similar pipeline-based implementations, the TRIM technique introduces no additional latency.

VI. Conclusion

This study presents TRIM, a dynamic thermal auto-compensation technique in 65 nm CMOS technology, enhancing computational accuracy for resistive IMC systems across varying temperatures. The implementation effectively addresses and solves three critical challenges: the thermal sensitivity of first-order resistive elements, the SER degradation with temperature, and maintaining inference accuracy amidst temperature fluctuations. By integrating a CTAT voltage generator, the system adjusts dynamically to temperature variations, enhancing the accuracy of the MAC operation, as evidenced by a reduction in the mean error from 9.33% to –0.81%. Moreover, inference accuracy sees a substantial rise from 36% to 85% at –20 °C and from 27% to 85% at 60 °C, demonstrating the system's robustness against thermal changes. The architecture eliminates the need for LUTs, weight adjustments, or external circuits, ensuring stable operation with ultralow power and area overhead of < 0.02% and 2%, respectively. Furthermore, the TRIM compensation technique is scalable and can be applied to any column size, making it highly adaptable for larger or more complex IMC systems. The TRIM technique not only improves the thermal resilience of IMC systems but also supports their use in AI-centric edge applications, promising to enhance the adoption of DNNs in real-world scenarios.

References

[1] C. Silvano et al., "A survey on deep learning hardware accelerators for heterogeneous HPC platforms," 2025, *arXiv:2306.15552*.

[2] B. Murmann, "Mixed-signal computing for deep neural network inference," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 1, pp. 3–13, Jan. 2021.

[3] J.-S. Seo et al., "Digital versus analog artificial intelligence accelerators: Advances, trends, and emerging designs," *IEEE Solid-State Circuits Mag.*, vol. 14, no. 3, pp. 65–79, Mar.–May 2022.

[4] R. Sarpeshkar, "Analog versus digital: Extrapolating from electronics to neurobiology," *Neural Comput.*, vol. 10, no. 7, pp. 1601–1638, 1998.

[5] P. Chi et al., "PRIME: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, 2016, pp. 27–39.

[6] C.-X. Xue et al., "24.1 a 1Mb multibit ReRAM computing-in-memory macro with 14.6ns parallel MAC computing time for CNN based AI edge processors," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2019, pp. 388–390.

[7] M. Le Gallo et al., "A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference," *Nat. Electron.*, vol. 6, no. 9, pp. 680–693, 2023.

[8] M. Zhou, M. Imani, S. Gupta, and T. Rosing, "Thermal-aware design and management for search-based in-memory acceleration," in *Proc. 56th Annu. Design Autom. Conf.*, 2019, pp. 1–6.

[9] K. Smagulova, M. E. Fouda, and A. Eltawil, "Thermal heating in ReRAM crossbar arrays: Challenges and solutions," *IEEE Open J. Circuits Syst.*, vol. 5, pp. 28–41, Jan. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10416883

[10] I. Giannopoulos, M. Le Gallo, V. P. Jonnalagadda, E. Eleftheriou, and A. Sebastian, "Temperature compensation schemes for in-memory computing using phase-change memory," in *Proc. 2nd IEEE Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, 2020, pp. 286–290.

[11] Y. Ling et al., "Temperature-dependent accuracy analysis and resistance temperature correction in RRAM-based in-memory computing," *IEEE Trans. Electron Devices*, vol. 71, no. 1, pp. 294–300, Jan. 2024.

[12] M. V. Beigi and G. Memik, "Thermal-aware optimizations of ReRAM-based neuromorphic computing systems," in *Proc. 55th ACM/ESDA/IEEE Design Autom. Conf. (DAC)*, 2018, pp. 1–6.

[13] G. Singh, D. Monga, O. Numan, M. Andraud, and K. Halonen, "On-chip built-in self-calibration of thermal variations for mixed-signal in-memory computing," in *Proc. IEEE Eur. Test Symp.*, 2024, pp. 1–6.

[14] D. C. Monga, O. Numan, M. Andraud, and K. Halonen, "A temperature and process compensation circuit for resistive-based in-memory computing arrays," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2023, pp. 1–5.

[15] Q. Zang, W. L. Goh, F. Li, L. Lu, and A. T. Do, "Temperature compensation on SRAM-based computation in memory array," in *Proc. 19th Int. SoC Design Conf. (ISOCC)*, 2022, pp. 1–2.

[16] H. Shin, M. Kang, and L.-S. Kim, "A thermal-aware optimization framework for ReRAM-based deep neural network acceleration," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, 2020, pp. 1–9.

[17] J. Meng et al., "Temperature-resilient RRAM-based in-memory computing for DNN inference," *IEEE Micro*, vol. 42, no. 1, pp. 89–98, Jan./Feb. 2022.

[18] X. Liu, M. Zhou, T. S. Rosing, and J. Zhao, "HR3AM: A heat resilient design for RRAM-based neuromorphic computing," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, 2019, pp. 1–6.

[19] P.-Y. Chen, F.-Y. Gu, Y.-H. Huang, and I.-C. Lin, "WRAP: Weight RemApping and processing in RRAM-based neural network accelerators considering thermal effect," in *Proc. Design, Autom. Test Europe Conf. Exhibit. (DATE)*, 2022, pp. 1245–1250.

[20] M. V. Beigi and G. Memik, "THOR: THermal-aware optimizations for extending ReRAM lifetime," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. (IPDPS)*, 2018, pp. 670–679.

[21] D. Wei, J. Qu, Y. Zeng, and L. Qiao, "A compensation principle for the temperature sensitivity of in-memory computing arrays," in *Proc. 13th Non-Volatile Memory Syst. Appl. Symp. (NVMSA)*, 2024, pp. 1–6.

[22] X. Guo et al., "Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells," in *Proc. IEEE Cust. Integr. Circuits Conf. (CICC)*, 2017, pp. 1–4.

[23] M. Hu et al., "Dot-product engine for neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication," in *Proc. 53rd ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, 2016, pp. 1–6.

[24] C. Sawigun, X. Yang, and C. M. Lopez, "Ultra-low-power voltage references: Exploring picowatt-level design using CMOS and hybrid architectures," *IEEE Solid-State Circuits Mag.*, vol. 15, no. 4, pp. 50–57, Sep.–Nov. 2023.

[25] K. Yu, Y. Zhou, S. Li, and M. Huang, "A 23-pW NMOS-only voltage reference with optimum body selection for process compensation," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 11, pp. 4213–4217, Nov. 2022.

[26] D. C. Monga and K. Halonen, "A dual mode all NMOS 7-T temperature sensor and voltage reference for biomedical applications," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2024, pp. 1–5.
[27] W. Wang, V. Reddy, A. T. Krishnan, R. Vattikonda, S. Krishnan, and Y. Cao, "Compact modeling and simulation of circuit reliability for 65-nm CMOS technology," *IEEE Trans. Device Mater. Rel.*, vol. 7, no. 4, pp. 509–517, Dec. 2007.
[28] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

Dipesh C. Monga (Graduate Student Member, IEEE) received the M.Sc. degree (cum laude) in micro- and nanoelectronic circuit design from Aalto University, Espoo, Finland, in 2020, where he is currently pursuing the Ph.D. degree.

He is working as a Research Scientist with the Integrated Circuits Team, VTT Technical Research Centre of Finland, Espoo. He was a Visiting Doctoral Researcher with the Bio/CMOS Interfaces Laboratory, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, for five months. His main research interest is focused on power management units and analog-front-end interfaces for sensors.

Gaurav Singh (Graduate Student Member, IEEE) received the M.Sc. degree in VLSI from Amity University, Noida, India, in 2015. He is currently pursuing the Doctoral degree with the Department of Electronics and Nanoengineering, Aalto University, Espoo, Finland.

His research focuses on the development of low-power system-on-chip solutions for sensor data processing. His work particularly explores micro-processors, low-power SRAM memories, and serial interfaces to enhance communication and debugging. His research work also includes integrating RISC-V processors with compute in-memory cores as hardware accelerators, aimed at improving power efficiency in AI applications.

Omar Numan (Student Member, IEEE) received the M.Sc. degree in micro- and nano-electronic circuit design from Aalto University, Espoo, Finland, in 2020, where he is currently pursuing the Ph.D. degree.

His research interests include the design of analog and mixed-signal integrated circuits, with a focus on efficient resistive-based computing-in-memory macros. He has contributed to low-power, high-performance circuit design projects for emerging computing architectures and focuses on integrating memory and processing for next-generation AI accelerators.

Kazybek Adam (Student Member, IEEE) received the M.Sc. degree in electrical and electronic engineering from Nazarbayev University, Astana, Kazakhstan, in 2019. He is currently pursuing the Doctoral degree with the Department of Electronics and Nanoengineering, Aalto University, Espoo, Finland.

His research is in analog circuit design for AI accelerators with focus on sample and hold circuits, analog memories, and noise analysis.

Martin Andraud (Member, IEEE) received the Ph.D. degree from Grenoble University, Saint-Martin-d'Hères, France, in 2016.

He is an Assistant Professor of Microelectronics with UCLouvain, Louvain-la-Neuve, Belgium, and a Visiting Professor with Aalto University, Espoo, Finland. He was a Postdoctoral Researcher successively with TU Eindhoven, Eindhoven, The Netherlands, in 2016 and KU Leuven, Leuven, Belgium, from 2017 to 2019. From 2019 to 2024, he was an Assistant Professor with Aalto University. His research interests include the interface between edge AI, hardware/software co-design, testing, and reliability of custom ASIC for various AI accelerators, for instance, mixed-signal compute-in-memory architectures and alternative to deep learning models (probabilistic reasoning or hybrid AI).

Kari A. I. Halonen (Member, IEEE) received the M.Sc. degree in electrical engineering from the Helsinki University of Technology, Espoo, Finland, in 1982, and the Ph.D. degree in electrical engineering from Katholieke Universiteit Leuven, Leuven, Belgium, in 1987.

Since 1988, he has been with the Electronic Circuit Design Laboratory, Helsinki University of Technology (since 2011, Aalto University), Espoo. Since 1993, he has also been an Associate Professor and has been a Full Professor with the Faculty of Electrical Engineering and Telecommunications since 1997. He became the Head of the Electronic Circuit Design Laboratory in 1998. He was appointed as the Head of the Department of Micro- and Nanosciences, Aalto University from 2007 to 2013. He specializes in CMOS and BiCMOS analog and RF integrated circuits, particularly for telecommunication and sensor applications. He is the author or co-author of over 450 international and national conference and journal publications on analog and RF integrated circuits.

Prof. Halonen was a recipient of the Beatrice Winner Award in ISSCC Conference 2002. He has served as a TPC Member for ESSCIRC and ISSCC. He has been an Associate Editor of the IEEE JOURNAL OF SOLID-STATE CIRCUITS and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS, a Guest Editor of the IEEE JOURNAL OF SOLID-STATE CIRCUITS, and the Technical Program Committee Chairman of the European Solid-State Circuits Conference in 2000 and 2011.