IEEE Open Journal of
Circuits and Systems

Received 18 September 2023; revised 15 November 2023 and 17 January 2024; accepted 26 January 2024. Date of publication 30 January 2024; date of current version 23 February 2024.

Digital Object Identifier 10.1109/OJCAS.2024.3360257

Thermal Heating in ReRAM Crossbar Arrays: Challenges and Solutions

KAMILYA SMAGULOVA[1] (Member, IEEE), MOHAMMED E. FOUDA[2] (Senior Member, IEEE), AND AHMED ELTAWIL[1] (Senior Member, IEEE)

[1]Division of CEMSE, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia

[2]Rain Neuromorphics, Inc., San Francisco, CA 94110, USA

This article was recommended by Associate Editor T. Kawahara.

CORRESPONDING AUTHOR: M. E. FOUDA (e-mail: foudam@uci.edu)

This work was supported by the King Abdullah University of Science and Technology CRG Program under Grant URF/1/4704-01-01.

ABSTRACT The high speed, scalability, and parallelism offered by ReRAM crossbar arrays foster the development of ReRAM-based next-generation AI accelerators. At the same time, the sensitivity of ReRAM to temperature variations decreases R_{ON}/R_{OFF} ratio and negatively affects the achieved accuracy and reliability of the hardware. Various works on temperature-aware optimization and remapping in ReRAM crossbar arrays reported up to 58% improvement in accuracy and 2.39× ReRAM lifetime enhancement. This paper classifies the challenges caused by thermal heat, starting from constraints in ReRAM cells' dimensions and characteristics to their placement in the architecture. In addition, it reviews the available solutions designed to mitigate the impact of these challenges, including emerging temperature-resilient Deep Neural Network (DNN) training methods. Our work also provides a summary of the techniques and their advantages and limitations.

INDEX TERMS ReRAM, memristor, thermal heating, nonideality, resistive crossbar arrays, resistive hardware accelerators.

I. INTRODUCTION

THE RAPID progress in artificial intelligence (AI) is dictating new requirements for hardware accelerators. Modern computational processes are characterized by an abundance of dot-product operation and an extreme lack of storage space. In this regard, non-volatility, nanoscale size, and the ability to retain multiple states made *resistive switching materials (RSMs)* promising in designing energy-efficient high-density memory devices. Moreover, RSMs, and resistance random access memory (ReRAM) in particular, can act as synapses and allow the building of artificial neurons and even neural networks. Multiple ReRAM cells organized into crossbar arrays can perform vector-matrix multiplication (VMM) faster and more efficiently than von-Neumann-based architecture since ReRAM cells can store data values as conductance states and reduce data movement between separate memory and processing units [1].

Therefore, computing-in-memory (CIM) or processing-in-memory (PIM) analog and digital ReRAM-based accelerators such as ISAAC [2], PRIME [3], PUMA [4] and others are firmly entering modern electronics. In particular, ISAAC outperformed its fully digital counterpart DaDianNao with improvements of 14.8× in throughput, 5.5× in energy, and 7.5× in computational density [2].

Nevertheless, an intra-class comparison with state-of-the-art (SoTA) commercial accelerators shows that existing ReRAM-based accelerators have higher power density with non-uniform distribution [5]. On the other hand, it also leads to disproportional temperature distribution. Previous works showed that an increase in temperature has an impact on resistive switching behavior and the R_{ON}/R_{OFF} ratio of ReRAM cells [7], [8]. In turn, the change of the conductance states affects the accuracy of ReRAM-based hardware [9]. Moreover, the materials and dimensions of ReRAM cells can define the level of the hardware's sensitivity to temperature [10], [11]. A closer look at ReRAM-based architectures shows that heterogeneous parts of the accelerators demonstrate non-uniform power density distribution

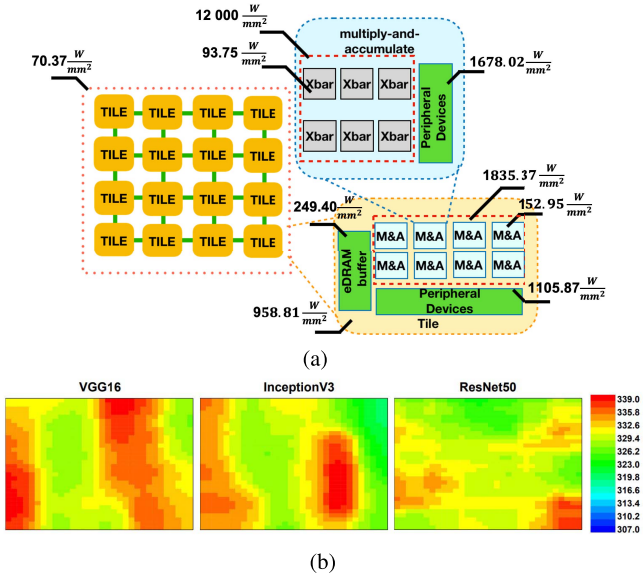

FIGURE 1. a) Power density of ISAAC-CE [5]; b) Steady-state temperature distributions of the same ReRAM chip running three different CNN models for inference: VGG16, InceptionV3, ResNet50 [6].

(as shown in Figure 1a for ISAAC) and consequently result in uneven temperature regions [5], [12]. Since power-hungry components have a higher rate of heat dissipation, they might interfere with the temperature and performance of surrounding elements. It was observed that an increase in temperature from 300K to 400K may reduce the accuracy of ReRAM-based hardware by up to 6× [9]. Moreover, due to different conductance values and input voltages, there might be a non-uniform temperature distribution within the ReRAM crossbar arrays, too. Figure 1b shows the steady-state thermal distribution in the same ReRAM chip during inference of VGG16, InceptionV3, and ResNet50 workloads for ImageNet dataset classification. As can be seen, the temperature difference between the models can reach up to 17.16K [6].

However, the majority of resistive hardware accelerators did not consider the thermal sensitivity of ReRAM cells in their design. The study of temperature impact on ReRAM-based architectures and the development of solutions to mitigate the problem started gaining attention only recently [8], [10]. This paper contributes in the following ways:

- We summarized the design and performance challenges of ReRAM-based hardware caused by temperature increase;
- We reviewed existing solutions developed to address the identified challenges;
- We categorized methods designed to mitigate the impact of temperature and analyzed their advantages and shortcomings compared to each other;
- Finally, based on the solutions discussion, we highlighted the key takeaways.

The rest of the paper is organized as follows: Section II provides information on ReRAM crossbar arrays and SoTA ReRAM-based neural accelerators. Section III discusses the thermal challenges in ReRAM-based hardware caused by temperature increase, and Section IV introduces existing techniques developed to address these challenges. Finally, Section V provides a summary discussion of the presented solutions.

II. EXISTING RESISTIVE NEURAL ACCELERATORS

A. RERAM CROSSBAR ARRAYS

ReRAM is a non-volatile memory device with conducting filament (CF) material sandwiched between top and bottom electrodes [13]. *Resistivity switching* (RS) of ReRAM cells from High Resistance State (HRS) to Low Resistance State (LRS) and vice versa can be controlled via connection and disconnection of the CF. Typically, ReRAM devices operate in *read* and *write* modes. During write mode, current or voltage pulses of certain amplitude, polarity, and duration are applied to the ReRAM to program its state. Sensing the ReRAM state is performed during read mode via applying voltages and currents of a specific range.

ReRAM's high speed and scalability, power efficiency, nanoscale size, and ability to retain a value in a non-volatile manner sparked interest in ReRAM-based resistive crossbar array (RCA) architectures. RCAs can serve either as non-volatile memory devices for storing data or as CIM architectures for performing VMM or accelerating neural networks. In the latter application, a ReRAM cell acts as a synaptic weight $w_{i,j}$ of a neural network with a neuron output $y_j = \sum_{i=1}^{N} w_{i,j} \times x_i$. According to Kirchoff's current law (KCL), the output current of each column in RCA is equal to a weighted summation of the input voltages, $I_j = \sum_{i=1}^{N} G_{i,j} \times V_i$. The weights are mapped to the RRAMs' conductances, G, while the inputs are mapped to the applied voltages, V, This property forms the basis of many ReRAM-based accelerators [1].

B. SOTA RERAM ACCELERATORS

The typical architecture of many-core bank- or tile-based resistive hardware accelerators comprises ReRAM crossbar arrays and various peripheral circuits and interconnects. Two of the first many-core ReRAM-based accelerator designs were ISAAC [2], and PRIME [3]. ISAAC has many-core architecture with tiles connected via network-on-chip (NoC). Compared to fully digital neural network accelerator DaDianNao [14], utilization of RCAs in ISAAC for VMM operation allowed ISAAC to reduce energy by 5.5× and increase throughput and computational density by up to 14.8× and 7.5×, respectively [2]. PRIME consists of banks that are connected via bus interconnect and uses RCAs for both data storage and VMM. Both accelerators support only the inference phase with 16-bit precision. Figure 2 shows the hierarchical structure of a ReRAM-based accelerator, including a node, processing tile (PT), processing unit (PU), and RCA.

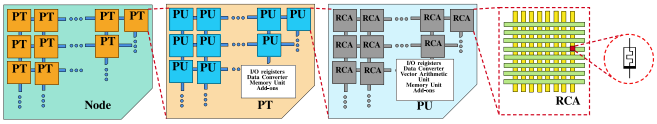

FIGURE 2. Hierarchical architecture of a ReRAM-based CIM accelerator (2D planar design): from a node to a resistive cell.

Communication between on-chip and off-chip components in multi/many-core platforms takes place via interconnects. Currently, ISAAC and PRIME serve as baseline models for the majority of SoTA ReRAM neural accelerators. Subsequent architectures such as AEPE [15], PUMA [4], Newton [16], and others have mainly aimed to decrease the power consumption of the peripheral circuits by modification of the resolution of ADC and DAC circuits or optimization of the weight mapping. In addition, PRIME-based PipeLayer [17] and ISAAC-based AtomLayer [18] and PANTHER [19] architectures provide support for on-chip training phases.

Heterogeneous on-chip and off-chip components in multi/many-core platforms should be placed to ensure high signal transmission speed/rate, small area, and low power. Traditional two-dimensional integrated circuits (2D ICs) are no longer feasible for this task, and active research is being conducted in the fields of 2.5D/3D stacking [20]. A through-silicon via (TSV) (also called an active TSV-interposer) technology allows bonding several dies in a face-to-face (F2F), face-to-back (F2B) and back-to-back (B2B) manner. TSV is used in 2.5D/3D die-stacking, including popular commercial technologies like Micron's Hybrid Memory Cube (HMC) and Hynix's High Bandwidth Memory (HBM). However, TSV does not scale well as the technology node size shrinks. Recently proposed monolithic three-dimensional (M3D) integration, also called 3D sequential integration, allows integration of ICs on top of each other on a single silicon substrate [21].

Improvement in bandwidth and power can also be achieved by stacking 2D planar ReRAM crossbar arrays into horizontal 3D ReRAM (H-ReRAM) or horizontal cross-point architecture (HCPA). There is also a vertical 3D ReRAM (V-ReRAM) design known as a vertical cross-point architecture (VCPA). Here, multiple devices are fabricated at the sidewall of horizontally running word-lines (WL) and a vertically oriented bit-line (BL). Both H-ReRAM and V-ReRAM allow scaling the ReRAM device size down to $4F^2/n$ where n is the number of stacked layers [22]. Generally, 3D die-stacking and 3D stacking of ReRAM arrays also lead to higher power densities and thermal problems.

In Figures 3a and b, SoTA ReRAM neural accelerators are compared against commercial accelerators, including Google TPUv4 [23], GraphCore C2 [24], Groq [25], Nvidia A100 [26] and H100 [27]. All accelerators use 16 floating point precision operation. An intra-class comparison shows that the power density of the commercial accelerators is always less than $0.5W/mm^2$, whereas the

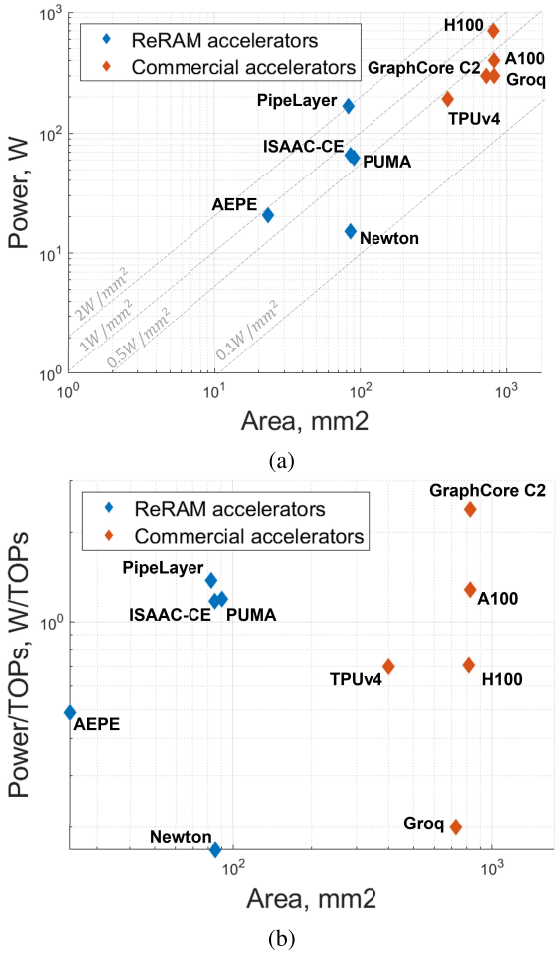

FIGURE 3. The state-of-the-art ReRAM-based and commercial accelerators: a) Power density; b) Power density per operation per second.

power density of the majority of resistive accelerators is above the bound and reaches $2W/mm^2$ in the case of PipeLayer. Figure 3b represents power spent per operation per second versus area. The best performance are yield by ReRAM-based Newton and commercial H100 and Groq.

III. CHALLENGES

The increase in the computational capacity of information processing systems is accompanied by continuous device size shrinking and technological advancements such as three-dimensional stacking, die integration, and packaging. But they also create thermal challenges and lead to unwanted performance degradation. This applies in particular to ReRAM technology since all processes that take place during RS are thermally activated and can be described using an Arrhenius dependence [10]. In this section, we identified challenges that limit the design and operation of ReRAM-based hardware caused by thermal disturbance. Overall, these challenges can be grouped into challenges associated with device-level reliability and thermal design constraints at the system level.

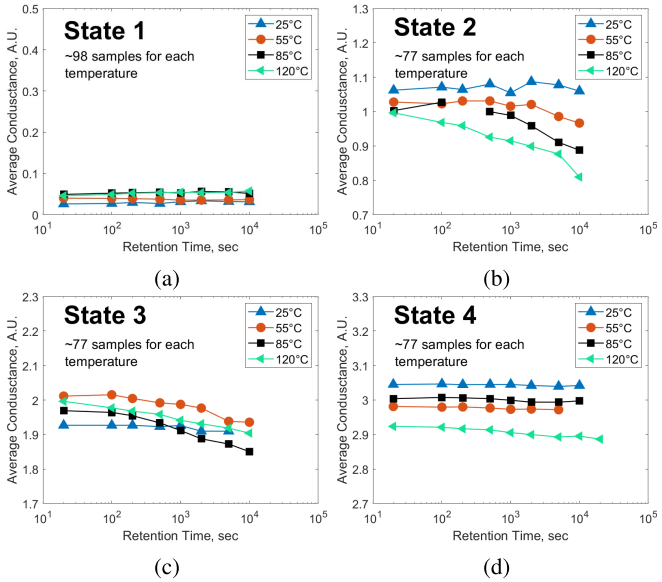

FIGURE 4. Measured average conductance of 2-bit HfO_2-based array over time at different temperature for the 4 possible states [30].

A. PERFORMANCE CHALLENGES

1) CHALLENGE 1: STATIC AND DYNAMIC RETENTION

A study of the conduction mechanism in RSMs has shown that room temperature affects the readout margin of a device [7]. In particular, a stable bipolar switching behavior in TiN/HfO_2/Ti/TiN ReRAM on $0.25\mu m$ complementary metal–oxide–semiconductor (CMOS) technology is observed within the temperature range 213–413K. However, further temperature increase leads to a proportional decrease of R_{OFF}/R_{ON} ratio and data loss. To describe the temperature effect on ON-state and OFF-state, a quantum point-contact (QPC) framework was used [28]. According to the QPC model, the ON-state shows a metallic characteristic, and resistance can be modeled as:

$$R_{ON} = R^0_{ON}[1 + \rho(T - T_0)] \tag{1}$$

where R^0_{ON} is the resistance measured at temperature T_0= 293K; and temperature coefficient $\rho = 3{\times}10^{-2}\ 1/K$.

Another test was conducted on 2-bit 256×256 1T1R HfO_2-based array 90nm technology. The temperature was varied from 300K to 395K, and the measured *static retention* characteristics were used to update a model [29], resulting in the following equations [30]:

$$\Delta\mu = \mu(t) - \mu_{init} = A_{avg} \times \log t \tag{2}$$

$$\Delta\sigma = \sigma(t) - \sigma_{init} = B_{var} \times \log t \tag{3}$$

where t is retention time; μ is the average conductance of the state and σ is its standard deviation; and A_{avg} and B_{avg} are the conductance drift rates that depend on temperature. According to the observations, the intermediate states of ReRAM are more susceptible to thermal effect due to a weak filament as shown in Figure 4 a-d [29], [31] Here, State 1 exhibits High Resistance State (HRS), whereas States 2-3 are in Low Resistance State (LRS). Resistance of HRS decreases over time due to neutral oxygen vacancy aggregation, and this process is irreversible, whereas resistance of LRS increases due to a gradual dissolution of the conducting filament [32]. Therefore, the sign of coefficients A and B depend on the initial state.

In addition to a static retention variation, there is a *dynamic retention* variation caused by temporal temperature changes. Dynamic retention can be modeled as the sum of the static variations at each temperature step [33].

2) CHALLENGE 2: ENDURANCE

The *expected shortest lifetime (ESL)* of ReRAM is around eight years. From prior works [34], the dependence of endurance of temperature variation can be expressed via write latency t_w [8]:

$$\text{Endurance} \approx (t_w/t_0)^{U_F/U_S - 1} \tag{4}$$

where t_0 is a constant that depends on the device, and U_F and U_S are the activation energy for the failure mechanism and the switching mechanism, respectively. For non-volatile devices, the typical ratio of $\frac{U_F}{U_S}$ varies from 2 to 4. From the analytical model, it was derived that increasing the temperature from 300K to 330K decreases t_w from 50 *ns* to 30 *ns* and reduces device endurance [8]. Moreover, high rates of SET-RESET also increase the temperature and decrease the average ReRAM lifetime. Surprisingly, low temperature also has a negative impact on ReRAM as it hinders recovery of a broken filament [35].

B. DESIGN CHALLENGES

1) CHALLENGE 3: THERMAL CROSS-TALK EFFECTS

The repeated SET-RESET switching cycles in a ReRAM device generate Joule heat, which may also affect the performance of surrounding devices. To quantify the thermal effect, a $Cu/TaO_x/Pt$ crossbar array with a neighboring line pitch of between 150 μm and 185 μm was studied. The heated ("aggressor") cell deteriorates the neighboring unheated ("victim") cell with the degradation factor D, which can be found as follows:

$$D = 1 - \frac{M_x(heated)}{M_x(unheated)} \tag{5}$$

where M_x is the maximum number of SET-RESET switching cycles of a 'marginal' memory cell required to become volatile. The term "marginal" means that the device is used as a temperature-sensitive probe. Testing of around 100 "marginal" devices (the current is set to I_{cc}=10 μA and voltage ramp rate rr=1.1V/s.) showed that the M_x of unheated TaO_x-based unheated device is around 13 and afterward the device demonstrates unstable performance [11].

In two-dimensional (2D) crossbar arrays, the degradation factor depends on the presence of a shared electrode, its material and size, and the remoteness of the unheated device from a heated cell. The study also showed that the first

neighbor cell suffers the highest degree of degradation. In particular, in $Cu/TaO_x/Pt$ crossbar arrays, degradation of the first neighbor along the Pt electrode was about D=67%; along the Cu electrode, it was 80%. In the case of non-shared electrodes, degradation of the first diagonal neighbor was D=19%. The thermal effect increases with the downscaling of the pitch size and spacing between them. It becomes a huge issue in commercial ReRAM devices that are 1000× smaller than the studied device [11].

Another parameter that can be used to evaluate thermal cross-talk in a crossbar array is the time t_s required to reach a thermal steady state - a condition where the temperature within the ReRAM device or/and crossbar array reaches a balanced temperature and remains stable over time under the same operating conditions. Maintaining a thermal steady state is essential in electronic devices since it ensures predictable and desirable operation of the device. For an individual device with feature size $80nm$, t_s is 5 ns, which is less than the required RESET time. But for 1D1R cell in $1\times1\times1$ array, t_s is around $50ns$ and steady-state temperature is equal to 500K, whereas, in a $3\times3\times3$ block array, t_s is $500ns$ and the temperature is 605K. Therefore, the thermal model of a single device should be extended. In [10], the authors presented two different "worst case" scenarios - one in a typical crossbar array structure and the other in a crossbar array with shared WL/BL represented in [10, Figs. 3a and 3b]. In the first case, when ReRAM cells were reset from LRS to HRS by applying a reset pulse, thermal heat propagated along the vertical direction and disturbed the unprogrammed layer. In the second case, the configuration allows erasing/programming at different layers of the crossbar. This time, heat from neighboring cells propagated in both vertical and horizontal directions and disturbed unprogrammed cells.

2) CHALLENGE 4: DIE-STACKING

ReRAM crossbar arrays can be stacked into heterogeneous structures using 2.5D and 3D integration technologies. These include TSV-based interposer and monolithic integration. The common interfacing methods in TSV-based integration are HBM and HMC. Such multiple die-stacking offers numerous advantages over 2D geometry scaling, including shorter interconnect, reduced latency, higher density, and smaller footprint [36], [37].

However, due to the different thermal densities of the components, stacked architectures suffer from inter-die thermal coupling and hotspots. Consequently, die-stacking leads to accuracy degradation and reliability challenges, including retention, thermal cross-talk, and endurance. For instance, in a 2.5D stacking design the temperature in the ReRAM banks reaches up to 344K and decreases their lifetime close to or below ESL. In a 3D interposer stacking design, the vertical heating temperature rises up to 380K and reduces ReRAM lifetime below 2.6 years [8]. In terms of die-to-die interconnections, M3D design has less area overhead compared to TSV-3D, but it is more sensitive to temperature. At ten years, the accuracy drop in M3D-air architecture was 53%, whereas, in TSV-3D, it was 10% [38].

3) CHALLENGE 5: LIMITED SCALING POTENTIAL

ReRAM, among other NVM technologies, is known for having the smallest size, around $4F^2$. Typically, a single ReRAM size is below $10nm$. Although ReRAM miniaturization allows saving power and area, scaling down the feature size (F) in devices such as NiO ReRAM from $100nm$ to $30nm$ node can lead to an increase of temperature from around 400K up to 1800K. In addition, miniaturization enhances the thermal cross-talk issue [10].

The thermal reaction model from [39] was utilized to study the behavior of saturated temperature in various ReRAM devices at low resistivity (10 $\mu\Omega$ cm), medium resistivity (50 $\mu\Omega$ cm) and high resistivity (100 $\mu\Omega$ cm) [40]. In the analysis, the reset voltage was set at 0.5V, and the thickness of the oxidation membrane was 200nm. The radius of the conductive filament of ZnO, TiO_2, WO_3 and HfO_2 was varied from $10nm$ to $100nm$. Overall, the conduction mechanism in ReRAM is mainly defined by the material and the geometry of CF and electrodes in a metal-insulator-metal (MIM) structure. Most popular ReRAMs can be classified into conductive bridge random access memory (CBRAM) and metal oxide ReRAM (OxRRAM).

4) CHALLENGE 6: RERAM CELL RESOLUTION

The precision of the weights significantly affects the accuracy of the output results. DNN training and inference on conventional GPU platforms are done using 32-bit floating-point precision. In the case of high-resolution ReRAM cells, there is a need for fewer crossbar arrays, which benefits in lower latency and better accuracy [41]. However, ReRAM cells have limited states and suffer from low precision. Prior work demonstrated that a 16-bit-wide fixed-point number representation is adequate for classification problems [42] and was used in a majority of the early ReRAM-based accelerators [2], [4]. The recent state-of-the-art works demonstrated that the lowest recommended bit width is 8 bits or above [43]. Besides, a 4-bit ReRAM cell is more susceptible to temperature variation than a 2-bit ReRAM cell due to its having a larger number of states [44]. Moreover, intermediate states are more vulnerable to heat than the states close to electrodes [30]. As mentioned earlier, an increase in temperature leads to a decrease of the G_{on}/G_{off} ratio and lowers the noise margin (NM) [44]. In particular, utilization of an 8-bit cell instead of a 2-bit cell decreases the number of required resistive crossbar arrays by 75%, but it also leads to a 64× NM drop [45]. Therefore, numerous ReRAM-based accelerators have adopted a weight-composing scheme [2], [4] with an increased number of arrays and additional power consumption and latency [5].

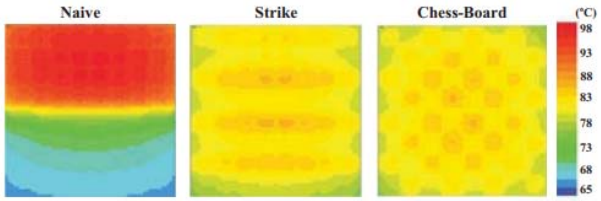

FIGURE 5. Steady-state temperature distributions of the bottom ReRAM layers: naive, strike, and chess-board allocation [46].

5) CHALLENGE 7: INPUT DISTRIBUTION

The amplitude and frequency of input signals contribute to the power density and speed of the hardware, respectively. An increase of input voltages leads to higher power consumption and generates heat [12]. The higher operating frequency decreases execution time, but thermal noise causes limitations in the frequency scaling of ReRAM-based designs [44] and an increase of frequency leads to accuracy decrease [6], [45]. The degradation gets more severe as the depth of neural networks grows. For instance, at 1GHz, the accuracy of VGG-19 drops by 50% compared to 5% in LeNet-5. Adding residual connections improves accuracy [44]. At frequency 100MHz and temperatures of 300K and 400K, the accuracy of ResNet20 was equal to around 88.2% and 88.1%, respectively. When frequency increased to 1GHz, at the same temperature conditions, accuracy decreased to 83.5% and 80.5%, respectively [45].

IV. EXISTING SOLUTIONS

The nanoscale size and non-volatile nature of RSMs allow the implementation of small-size and energy-efficient computational hardware components based on ReRAM crossbar arrays. On the other hand, dense architecture design increases temperature susceptibility and negatively affects reliability since an active ReRAM cell in a crossbar array causes thermal disturbance (TBD) in neighboring victim cells [10]. Therefore, the accuracy of a ReRAM simulation model plays a vital role in the validation of the ReRAM-based hardware design before its fabrication. One of the ways to overcome the conductance drift problem in RCAs is to refresh the cells' states frequently, but it requires additional power consumption [33].

In [46], authors tested three memory allocation schemes in a 3D crossbar array, namely "strike," "chess-board" and "naive", for the same model and identified that in all cases, the bottom ReRAM layers were the hottest. In addition, as can be seen from Figure 5, the peak temperature in the "strike" and "chess-board" schemes are 363K in contrast to 371K in the original "naive" scheme. Moreover, the average temperature of the ReRAM crossbar array was also reduced. Overall, these demonstrate that, to a certain extent, the thermal distribution can be controlled by static allocation schemes. In this section, we present the existing SoTA thermal-aware remapping and optimization solution designed for ReRAM-based memory and PIM accelerators.

A. SOLUTION 1: THOR

The goal of the *thermal-aware optimization for extending ReRAM lifetime (THOR)* is to keep the temperature of the ReRAM banks below a threshold temperature to ensure a lifetime above ESL. THOR consists of THOR - Lazy Access (THOR-LA) and THOR - Smart Access (THOR-SA) schemes, which can work both together and independently from each other. THOR-LA delays requests to hot banks and thus allows their cooling during idle periods. The delays are implemented by extending the memory controller (MC) to four queues: Normal read/write and Lazy read/write. THOR-SA reduces the number of accesses to hot arrays.

Nevertheless, it allows overall system power reduction by 5.5% and ReRAM lifetime enhancement by 2.06× the baseline design with a normal read queue, a lazy read queue, a normal write queue, and a lazy write queue [8].

B. SOLUTION 2: DEEPSWAPPER

Hybrid DRAM/NVM memory systems benefit from the lower access latency of DRAM and the high capacity of ReRAM. However, data migration between two memory types is costly due to the need for metadata storage. Existing swapping schemes are based on prediction tables and do not consider the temperature effect.

DeepSwapper is a novel deep learning-based page swap management scheme for hybrid DRAM/ReRAM memory. Instead of lookup tables, it uses a Long Short-Term Memory (LSTM) recurrent neural network (RNN) to predict future memory access patterns. This hardware-managed framework consists of two main components: an LSTM-Based Address Predictor and a Temperature-Aware Swap Management Unit. Evaluation results showed that the ReRAM lifetime was enhanced by 1.87× that of other schemes [8].

C. SOLUTION 3: TADMSIMA

Thermal-Aware Design and Management for Search-based In-Memory Acceleration (TADMSIMA) is a thermal-aware data allocation scheme that utilizes steady-state and dynamic thermal management (DTM) techniques [46]. In the first stage, static program analysis is used to estimate the number of ReRAM banks and their power consumption based on the type of application program, the size of the dataset, the architecture, and the operating frequency. Then, banks are classified as high power-consuming and low power-consuming. For thermal-aware mapping, a two-phase design space exploration method based on a genetic algorithm is applied.

The proposed system was validated on two search-based applications - hyperdimensional computing and database query processing. The experimental setup included ten encoding-search ReRAM bank pairs to store and compute data. According to the results, the steady-state temperature was reduced by at least 15.3K and the lifetime of

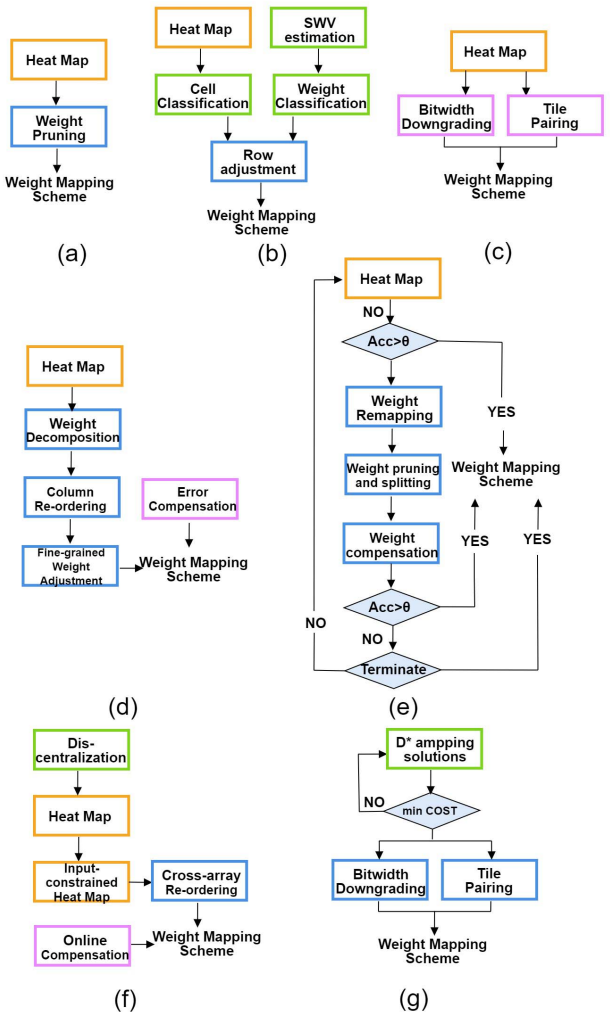

FIGURE 6. Thermal-aware Solutions for acceleration of DNN and IMC: a) Temperature-aware weight adjustment (TAWA) scheme; b) Thermal-aware Optimizations of ReRAM-based Neuromorphic Computing Systems (TARA); c) A Heat Resilient Design for RRAM-based Neuromorphic Computing (HR^{3}AM); d) Thermal-aware optimization framework for accelerating DNN on ReRAM (TOPAR); e) Weight RemApping and Processing in RRAM-based Neural Network Accelerators (WRAP); f) Thermal-aware layout optimization and mapping methods for resistive neuromorphic engines (TALOMRNE); g) Placement strategy in AccuReD.

the ReRAM device was extended by 57.2% on average. The dynamic temperature management provided 17.6% performance improvement compared to other SoTA methods.

D. SOLUTION 4: TARA

One of the first works that addressed the impact of temperature on computational accuracy in a ReRAM-based neuromorphic computing system proposed *weight remapping* based on *temperature-aware row adjustment (TARA)* [9]. Its performance is compared to a baseline architecture with random mapping and an architecture with *temperature-aware weight adjustment (TAWA)*. TAWA is based on weight pruning, as proposed in [47], and its implementation scheme is illustrated in Figure 6a. Here, weights mapped to hot cells are pruned, and the neural network is retrained again.

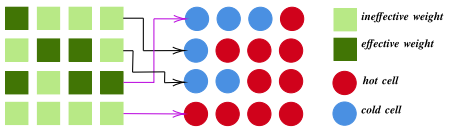

FIGURE 7. Weight mapping in TARA.

The schematic of TARA shown in Figure 6b was designed for Micron's HMC architecture. Here, diode thermal sensors were placed at the center and left side of each row of the ReRAM array - the hottest spots due to the close location of the analog-to-digital (ADC) converter and memory. At each epoch time, the temperature of the rows was approximated and classified. If the estimated temperature of a row was higher than the threshold temperature equal to 330K, the ReRAM crossbar row was considered hot; otherwise, cold. In addition, rows of neural networks were classified as *effective* and *ineffective* using a metric called Summed Weight Variations (SWV) and predefined threshold β:

$$SWV_{pq} = \sum_{j=0}^{m} |w_{pj} - g_{qj}| \tag{6}$$

where w_{pj} is the weight at the weight matrix location (p, j) and g_{pj} is the corresponding conductance of the ReRAM cell located at (q, j) in the ReRAM crossbar array. If $SWV > \beta$, rows are effective; otherwise, they are ineffective.

Increase of row temperature from 340K to 360K leads to an accuracy decrease from 61.7% to 23.4%. Therefore, at the final stage of the scheme, effective rows were mapped to the ReRAM crossbar array, avoiding hot rows, as in Figure 7. Evaluation of the *thermal-aware row adjustment* on a two-layer neural network in NeuroSim demonstrated an increase of the system accuracy by up to 39.2%.

E. SOLUTION 5: HR^{3}AM: A HEAT RESILIENT DESIGN FOR RRAM-BASED NEUROMORPHIC COMPUTING

Conversion of neural network weight w into conductance state of ReRAM cell G can be done based on the equation below:

$$G = \alpha \times w + \beta \tag{7}$$

where parameter $\alpha = \frac{G_{max} - G_{min}}{w_{max} - w_{min}}$ is used to scale a weight w within a range of $[G_{min}, G_{max}]$ and parameter $\beta = G_{max} - \alpha \times w_{max}$ is used to remove negative weights.

It was observed that a 1° increase of temperature in ReRAM-based architecture leads to an overall performance decrease of 0.9%. In order to decrease the negative impact of heat on ReRAM-based CNN accelerators, the HR^{3}AM design (Figure 6c) utilizes a *bitwidth downgrading* technique (HR^{3}AM-BD) and *tile pairing* (HR^{3}AM-TP) [6]. To do this, the HR^{3}AM system monitors temperature distribution in the ReRAM chip dynamically using temperature sensors. If the temperature is above the threshold (330K), a heat-resilient weight adjustment is applied:

$$G_{new} = \frac{1}{2^N} \times (\alpha \times w + \beta) \tag{8}$$

where G_{new} is the new conductance state and N is the number of shifted bits so that:

$$V_o = V_i^T \times G_{new} \times R_S \times 2^N = \left(V_i^T \times G_{old} \times R_S/2^N\right) \times 2^N. \tag{9}$$

where V_o is the output voltage; V_i is the input voltage; and G_{old} is the old weight.

In addition to this HR^{3}AM-BD, thermal distribution can be reduced by the introduction of *master* and *slave* tiles. The overheated (master) tile is paired with a cooled-down idle (slave) tile in such a way that the output of the master tile is read from even-index columns $V_{out}^m = \{v_0, v_2, \ldots v_{2N}\}$ and the output of the slave tile is read from odd-index columns $V_{out}^s = \{v_1, v_3, \ldots v_{2N+1}\}$. This decreases the number of functioning cells in a crossbar array and thus reduces power consumption. The pairing mode is represented by *a pairing bit* and *a master/slave bit* in crossbar arrays. The design was tested on a small two-layer network for MNIST classification and larger networks such as VGG16, ResNet50, and InceptionV3 for ImageNet classification. The obtained results showed 4.8%-58% improvement compared to the baseline model, which has no thermal optimization. In addition, HR^{3}AM showed better accuracy by 4.3%–41.8% over TARA [9].

F. SOLUTION 6: TOPAR

To reduce average temperature and temperature variance between ReRAM arrays in DNN accelerators, a *thermal-aware optimization framework for accelerating DNN on ReRAM (TOPAR)* has been proposed [48]. It consists of three-stage offline thermal optimization and online thermal-aware error compensation, as shown in Figure 6d.

There are $2^N - V$ ways to decompose an N-bit weight value V to positive and negative arrays. To reduce the temperature in the ReRAM chip, the first step of the offline stage performs a thermal-aware weight decomposition (TOPAR-I). In other words, TOPAR-I aims to identify a decomposition case with the smallest sum of partial weights. The next step, a thermal-aware column reordering (TOPAR-II), shuffles the order of the column pairs in positive and negative ReRAM arrays. This changes the weight and temperature distribution in ReRAM arrays and does not affect the computational output. The final step in offline optimization (TOPAR-III) is a fine-grained weight adjustment if there are more than two decomposition cases in TOPAR-I. It is performed sequentially, starting from the top-left position of the crossbar array. TOPAR-III aims to reduce the cost difference between positive and negative arrays. At an online stage, TOPAR improves ReRAM endurance by up to 2.39× and preserves inference accuracy.

G. SOLUTION 7: WEIGHT REMAPPING AND PROCESSING IN RRAM-BASED NEURAL NETWORK ACCELERATORS (WRAP)

A weight remapping and processing (WRAP) framework adopted the *subarray-based* approach rather than dealing with each weight individually [49]. This helped to reduce computational complexity accuracy while mitigating thermal issues to maintain the system. Figure 6e shows the flow of WRAP, which is based on three algorithms: weight remapping (WR); weight pruning and splitting (WPS); and weight compensation (WC).

At the initial stage, the framework receives parameters of DNN model hardware such as ReRAM cell resolution and ReRAM array size and maps weights to the accelerator. Afterward, it retrieves a heatmap of layers and estimates the accuracy of the system *Acc*. If the latter is below a predefined threshold level θ, three subarray-based algorithms are applied to remap the weights until the estimated accuracy is above the threshold accuracy. *Weight Remapping (WR)* algorithm remaps critical weights to relatively cool subarrays. If some of the subarrays are still hot, *Weight Pruning and Splitting (WPS)* algorithm generates several unused subarrays by pruning less-critical weights, and then critical weights are mapped to released subarrays. It was observed that deep layers are less sensitive to pruning than shallow layers. The process is terminated when the "prune ratio" in WPS is zero. *Weight Compensation* technique shifts conductance levels in order to decrease the impact of the temperature if WR and WPS methods cannot help. The framework was evaluated on VGG8, VGG11, ResNet34, and AlexNet for CIFAR-10 classification with less than 2% inference accuracy loss [49].

H. SOLUTION 8: TALOMRNE

Thermal-aware layout optimization and mapping methods for resistive neuromorphic engines (TALOMRNE) introduced a new layout that implies decreased temperature distribution by dis-centralizing components of the accelerator. In addition, TALOMRNE (Figure 6f) also noted that previous works emphasize only the weight itself and do not consider the input distribution that contributes to the final power consumption of the system. In addition, it adopted the cross-array swapping method of input-contained weights W_{bias}.

The method validation was done for two SNN models transformed from VGG9 and VGG11 on CIFAR-10 and CIFAR-100 datasets, respectively. The method allowed reducing the peak temperature up to 10.4° and improved the endurance by up to 1.72× [12].

I. SOLUTION 9: ACCURED

AccuRed is a heterogeneous ReRAM-GPU-based architecture for CNN training and inference. Its compute-intensive layers are mapped to ReRAM arrays, whereas precision-critical layers are mapped on GPUs. In addition, AccuRed performs a thermal-aware placement strategy (Figure 6g) based on a joint performance–thermal-aware mapping and a thermal reference cell (TRC) to reduce temperature impact. For effective mapping and high accuracy, AccuReD applies the Multiobjective Optimization (MOO) technique. The combined objective can be expressed as follows:

$$D^* = MOO(D, OBJ = U(d), T(d)) \tag{10}$$

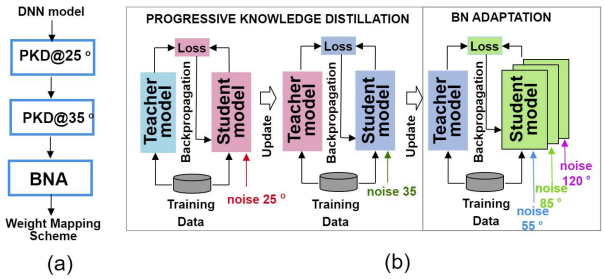

FIGURE 8. a) Temperature-Resilient ReRAM-based In-Memory Computing for DNN Inference (TRRIMC); b) Implementation of PKD and BNA in TRRIMC.

where D^* is the set of Pareto optimal mapping solutions among CNN layer mapping D; d is candidate mapping; T is the temperature objective combining both horizontal and vertical heat flow models; *MOO* is a multiobjective optimization solver; and *OBJ* is the set of all objectives (latency and temperature). AMOSA [50] was used as the MOO solver to minimize *cost*, which is a function of latency and temperature. Performance- and thermal-aware mapping of L layers of CNN to N processing elements (GPUs and ReRAMs) ensures low temperature and high performance. The CNN layers are classified as high-power (HP) and low-power (LP) layers. During one pipeline stage, the HP layer requires the ReRAM to be active for more than 50% and placed near the sink. LP layers can be placed farther from the sink. A comparison of TSV-based and M3D designs showed that M3D integration of AccuRed has superior thermal characteristics and allows more CNN layers without sacrificing accuracy. AccuRed outperformed conventional GPUs by 12× on average and can be further scaled up [44].

J. SOLUTION 10: TRRIMC

Noise injection during DNN model training is one of the ways to increase the robustness of ReRAM-based accelerators when it comes to temperature variations. In addition, knowledge distillation [51] can improve the performance of the model accelerated on the hardware. However, such a model recovers only in certain conditions and fails in other scenarios. To improve the generality of the model, authors [33] proposed a novel training algorithm - progressive knowledge distillation (PKD) - and thermal-aware batch normalization adaptation (BNA). The schematic of the *Temperature-Resilient RRAM-based In-Memory Computing for DNN Inference (TRRIMC)* is shown in Figures 8a and 8b.

In PKD, a clean, low-precision model is used as a teacher model. In the initial phase, a student model, duplicated from a teacher model, is trained with low-temperature noises. Then, the trained model acts as a new teacher model, and a new duplicated student model is trained with higher temperature noises and so on. During BNA, weights and learnable parameters of DNN are frozen, and further training of 16-bit fixed-point batch normalization (BN) parameters Y_{BNA} with noise injection at different temperature T scenarios is performed:

$$Y_{BNA} = w_T \times \frac{Y - \mu_T}{\sigma_T} + b_T \tag{11}$$

where Y is the output preactivation; w is the weight; μ is the mean within the batch and σ is its standard deviation. The proposed PKD+BNA method allows recovering the accuracy of the 2-bit ResNet on the CIFAR-10 for more than 30% and of the 4-bit ResNet-18 on TinyImageNet for more than 60%. The primary advantage of the scheme is the absence of the need to reprogram the initial ReRAM weights.

V. DISCUSSION

Table 1 lists the proposed solutions for ReRAM-based memory devices and accelerators. Each of them aimed to overcome certain challenges discussed in Section III such as the recovery of computation accuracy, extending the lifetime of ReRAM cells, and reduction of power consumption. *Solutions 1-3* in Section IV are designed for NVM memory architectures. They implement a thermal-optimal data allocation in order to retain the state and extend the lifetime of the cells. Due to the cumulative effect on the output of crossbar arrays, the ReRAM conductance state drift caused by thermal heat severely affects the performance of the ReRAM-based DNN accelerators rather than ReRAM-based storage devices. Therefore, the main aim of *Solutions 4-8* in Section IV is the recovery of the accuracy of the system during the acceleration of DNN and CIM tasks. Moreover, early works considered mainly steady-state temperature distribution cases, whereas recent works propose methods to control runtime temperature variations, too. The proposed temperature-adjustment schemes for RCAs can be divided into two categories: 1) temperature-aware optimization and remapping, and 2) temperature-resilient training of the DNN model.

The goal of thermal-aware optimization and remapping is to mitigate the impact of high temperature and to create a uniform temperature distribution in a ReRAM crossbar array. The optimization and remapping take place at different levels of granularity: weight level, group of weights (row-/column-wise) level, subarray level, array level, and tile level. The initial stage of these solutions requires the creation of a thermal profile, typically obtained from the limited number of temperature sensors located around ReRAM arrays. In HMC, it is the center and left side of the rows [9]. Afterward, various weight mapping optimization techniques are applied. These methods are provided in Table 2. The offline stage involves temperature-aware training and/or optimization steps prior to deployment on the ReRAM-based hardware. Besides, the temperature distribution in chips might dynamically change with time. The online stage includes measures designed to react to dynamic changes in temperature on the fly.

One of the basic methods of temperature-aware optimization and remapping is *weight pruning (WP)*. WP can be applied on either effective or ineffective weights.

TABLE 1. Thermal-aware ReRAM layout optimization solutions and the challenges. Missing challenges are not addressed yet.

Solution (Year)	Application	Architecture	Ch1	Ch2	Ch4	Ch6	Ch7	Description	Setup and Tools
THOR (2018)	NVM memory	2.5D/3D interposer		✓				THOR-LA has four queues: Normal read, Lazy read, Normal write, and Lazy write. 1. ReRAM banks are classified as hot and cold banks based on the sensed temperature. 2. Read and write requests to hot banks are delayed, allowing to cool them. Lifetime enhancement by 1.70×. Power reduction by 6.7%. THOR-SA has two queues: Normal read, Normal write. 1. Six bits are added to cache tags (two bits to show rank ID and four bits to identify bank number). 2. LLC are sampled to sets with hot and cold banks and maintained in the least recently used (LRU) order. 3. Two hit counters: hot hit counter and cold hit counter. 4. Maintains a temperature-aware policy to keep cache lines from hot banks longer in the LLC and reduce the number of future accesses to hot banks. Lifetime enhancement by 1.36×. Power reduction by 4.6%.	1. gem5 simulator integrated with NVMAIN + CACTI + DESTINY 2. Requires additional hardware
DeepSwapper (2019)	Hybrid NVM/DRAM memory	N/A		✓				1. Seq2seq LSTM: A sequence of past LLC miss addresses is used to predict a sequence of future LLC miss addresses. 2. A beam-search decoder is used to improve LSTM. Endurance improvement by 1.87×.	1.gem5 with Ramulator 2. A two-layer depth LSTM model with 128 and 64 hidden units 3. P100 NVIDIA GPU
TADMSIMA (2019)	Search-based hyperdimensional computing and database query processing	HMC-like	✓	✓				1. Static program analysis is used to estimate the number of ReRAM banks and their power consumption based on the type of application program, the size of the dataset, the architecture and the operating frequency. 2. Banks are classified as high power-consuming and low power-consuming. 3. For thermal-aware mapping, a two-phase design space exploration method based on genetic algorithm is applied. Steady-state temperature reduction by at least 15.3° and ReRAM lifetime enhancement by 57.2% on average.	1. McPAT 2. CACTI 3. HSPICE 4. HotSpot 5. Ni/HfO_2/PT and Ti/TiO_2/Pt
TARA (2018)	CIM/DNN inference (a 2-layer NN on MNIST)	HMC	✓					Baselines: neuromorphic hardware with random mapping scheme (accuracy 54.1%); TAWA (accuracy 62%). 1. Temperature collection and estimation. 2. Classification NN weight rows to effective and ineffective based on SWV. 3. Temperature-aware row adjustment, e.g. avoiding mapping effective rows to hot ReRAM cells. Accuracy improvement by 23.8% compared to the baseline; by 14.3% more than the TAWA scheme.	1. 1-bit per ReRAM 2. $[G_{off}; G_{on}]$=[3.07nS, 38.4nS] 3. NeuroSim
HR^{3}AM (2019)	CNN inference (1. a small two-layer NN on MNIST 2. VGG16, ResNet50 and InceptionV3 on ImageNet)	HMC	✓			✓		Baseline: neuromorphic hardware with random mapping scheme, TARA. 1. HR^{3}AM-BD: temperature-aware row adjustment using bit-width downgrading technique. HR^{3}AM-BD aims to improve accuracy. (Suitable for DNN with large cell resolution.) 2. HR^{3}AM-TP: the tail pairing method is when some operations from hot tiles are performed on the idle tile. HR^{3}AM-TP aims to reduce the temperature of the chip. (But extra tiles lead to a loss in parallelism.) Decrease in maximum temperature by 6.2°; decrease in average temperature for the entire chip by 6°; Accuracy improvement by 4.8%–58% over the baseline and 4.3%–41.8% over TARA.	1. Based on ISAAC 2. Requires temperature sensors, registers and comparators 3. A downgrade bit and control logic 4. Adjustment of shift-and-add and encoding circuits 5. Reserved idle tiles (10% of all tiles) 6. Tensorflow
TOPAR (2020)	DNN inference (ResNet18, ResNet50, VGG-16, neural collaborative filtering (NCF), a 2-layer stacked LSTM)	N/A		✓				Baseline: HR^{3}AM Offline optimization: 1. TOPAR-I: a thermal-aware weight decomposition. 2. TOPAR-II: a thermal-aware column reordering. 3. TOPAR-III: a fine-grained weight adjustment. Online optimization: 1. Restoring distorted current-sum results with the thermal-aware error compensation. Improved endurance up to 2.39×.	1. Based on ISAAC 2. 2-bit ReRAM cell 3. 64×64 array 4. HotSpot thermal simulator 5. Pytorch 6. 8-bit weights 7. Endurance 4.14×10^8 8. Synopsys Design Compiler
WRAP (2022)	CIM/DNN inference (VGG 8, VGG11, Alexnet and ResNet34)	3D (4-layer HMC-like)	✓			✓		Baseline: HR^{3}AM Subarray-based approach saves computational resources. 1. WR: avoiding mapping of important weights to higher temperature subarrays. 2. WPS: less critical weights are pruned, which frees some subarrays. 3. WC: bitwidth downgrading on subarrays. Accuracy loss is less than 2%; and less than 1% loss with compensation.	1. Based on ISAAC 2. $[G_{off}; G_{on}]$=[3.07nS, 38.4nS] 3. Pytorch 4. 4-bit, 6-bit, 8-bit weights 5. Pruning ratio 40–50% (best results)
TALOMRNE (2022)	SNN	N/A		✓	✓		✓	Takes into consideration input distribution and utilizes a cross-array mapping method. 1. Layout optimization by dis-centralizing high-density components. 2. Thermal-aware weight reordering considering input distribution and weights value. The average power range decreases by 20% in *Conv* and by 15% in *FC* layers. Endurance improvement by 1.30× and 1.72× in VGG-11 and VGG-9, respectively.	1.Based on ISAAC 2. Block size 128×128 3. Input voltage [0;0.9]V 4. [R_{on};R_{off}]=[5kΩ; 500kΩ] 5. Hotspot 6. Endurance 4.14×10^8
AccuReD (2020)	Hybrid NVM/GPU; CNN training/ inference (VGG-19)	M3D			✓	✓		Aims to reduce *cost*, which is a function of temperature and CNN pipeline latency. 1. Used a thermal reference cell (TRC) and multicell reference array. 2. MOO and thermal-aware mapping. AMOSA used as the MOO solver.	1. Based on AccuReD 2. GPGPU-Sim 3. PytorX
TRRIMC (2021)	CIM/DNN training (a 2-bit ResNet-18 on CIFAR-10 and TinyImageNet)	N/A						Baseline: 2-bit ResNet-18 on the CIFAR-10 with perodic refreshing after 30 s of operation. 1. PKD training. 2. Thermal-aware BNA. Accuracy is > 90% until 10^4 s with reduced refreshing frequency by 250×.	1. 90nm prototype chip 2. 2-bit HfO_2 ReRAM cell 3. 256×256 ReRAM array 4. NeuroSim 5. For BNA: temperature sensors and BN multiplexer

Ch: Challenge; N/A: not reported

TABLE 2. Methods applied to decrease temperature effect in ReRAM-based CIM accelerators.

Approach	Method	Implementation	Level	Solution	Offline	Online	Advantages	Disadvantages
Temperature-aware optimization and weight remapping	Weight pruning (WP)	pruning weights of hot ReRAM cells and retraining of DNN	array	TAWA	✓	✗	hot ReRAM cells are excluded from utilization	requires extra training
	Row adjustment (using the SWV metric)	effective weights are mapped to cold ReRAM array rows	array	TARA	✓	✗	row level; able to recover accuracy of DNN model	the ambient temperature might increase and degrade performance
	Bit-width downgrading	weight is shifted from a temperature sensitive conductance state to a conductance state with less sensitivity and the obtained multiplication result is shifted back	weight	HR^{3}AM	✓	✓	reduces thermal effect on weights with high conductance	cannot be used for weights with low resolution
	Tile pairing	pairing of overheated and cooled-down idle tiles	array	HR^{3}AM	✓	✓	reduces the average temperature	requires extra tiles/crossbar arrays
	Weight decomposition (WD)	searches for the smallest sum of partial weights among (2^N - V) cases for N bit value V	weight	TOPAR	✓	✗	no additional training required; finding an efficient way of weight decomposition and mapping into positive and negative arrays	possible temperature variance between negative and positive arrays
	Column reordering	shuffling order of columns without affecting the computational output	array	TOPAR	✓	✗	changes distribution of weights in an array and the temperature variance between them	the process is complicated due to a group of positive and negative arrays
	Fine-grained weight adjustment	WD with minimum cost upon the result of column reordering; performed sequentially	array	TOPAR	✓	✗	reduces temperature variation between positive and negative arrays	limited to weights that have more than two thermal-optimized decomposition cases
	Error compensation	current mirror circuits are used for compensation of the current-sum results	column	TOPAR	✗	✓	helps to restore the accuracy dropped due to the rise of ambient temperature	additional area for transistors
	Weight remapping (WR)	avoiding mapping important weights of shallow layers to hot subarrays	subarray	WRAP	✓	✗	subarray level; able to recover accuracy of DNN model	the ambient temperature might increase and degrade performance
	Weight pruning and splitting (WPS)	less-critical (ineffective) weights are pruned and critical (effective) weights are mapped to unused subarrays	subarray	WRAP	✓	✗	pruning frees several subarrays for mapping critical weights	"prune ratio" could become 0 before accuracy reaches above the threshold
	Weigh compensation (WC)	weight is shifted from a temperature sensitive conductance state to a conductance state with less sensitivity and the obtained multiplication result is shifted back	weight	WRAP	✓	✗	reduces thermal effect on weights with high conductance	the compensated result is close to the original
	Cross-array reordering	weight reordering with consideration of input signal and weight value	cross-array	TALOMRNE	✓	✗	considers input distribution; has larger solutions space since re-ordering is performed on rows or columns between arrays	typically requires many iterations
	Performance- and thermal-aware mapping	CNN layers are classified as high power (HP) and low power (LP) HP are mapped near the sink; LP are mapped farther from the sink	array	AccuReD	✓	✗	implementing multiply-and-accumulate MAC operations on ReRAM rather than GPU is more energy-efficient	exploration of the best mapping in the mapping space can be time-consuming
Temperature-resilient training	Progressive knowledge distillation (PKD)	aims to minimize discrepancy between teacher and subsequent student models trained at different levels of temperature and noise	array	TRRIMC	✓	✗	the model is trained to be robust when it comes to static and dynamic temperature fluctuations	typically for a short operating time; can lead to overall accuracy degradation
	Batch normalization adaptation (BNA)	training of batch normalization (BN) parameters with noise-injection while the weight and learnable parameters of the NN remain the same	array	TRRIMC	✓	✗	improves robustness when it comes to temperature and hardware compatibility	adds extra BN parameters

In [9], effective weights that were mapped to hot cells are pruned, and NN retrained again. This avoids critical weights being mapped to hot ReRAM cells and maintains accuracy. In [49], ineffective weights are pruned to free space in the arrays. Then, critical weights are remapped. In addition to weight pruning, at weight level, techniques such as *weight decomposition (WD)*,*bit-width downgrading (BD)* and *weight compensation (WC)* are available. These approaches are based on the ReRAM feature that implies that high conductance states are more vulnerable to an increase in temperature. In the cases when weights of different polarities are implemented using negative and positive crossbar arrays, the WD technique searches for the decomposition case so that the temperature distribution in both arrays is uniform and as low as possible. For further optimization, a *fine-grained weight adjustment* method can be applied. BD is the dynamic thermal management method and is used when the temperature is above 330K. In this technique, weight adjustment is performed by shifting bits and, therefore, can be applied only on weights with high resolution. Generally, BD and WC are the same operations; both methods shift conductance states and restore multiplication results, but WC is applied only to weights that were not protected by the WR and WPS techniques.

Other ways to change temperature distribution in ReRAM crossbar arrays include *swaping rows or columns* within the same arrays ("in-array") or between arrays ("cross-array") in order to decrease temperature variation. These techniques can be applied at a row and/or column, subarray, array, or tile level. In [6] authors proposed to utilize the *tile pairing* method to split weights into two tiles. Pairing hot tiles with idle tiles decreases the average temperature since both of them work in low-power mode. The temperature-aware training of the DNN model implies the resilience of the trained model to the ambient temperature change of a given range. In *Solution 10*, the DNN model was trained with noise injection considering possible temperature fluctuations. Such model remains resilient to temperature variations for a certain period of time after mapping to the hardware and does not require retraining and reprogramming of the states [33]. One of the ways to improve the PKD method is to implement via injection lower noise levels to fully connected layers of CNN, as they are found to be more sensitive to noises [45].

Due to heterogeneous materials and components, computing systems have non-uniform temperature distribution across their architecture. "Hot spot" regions and average temperature of the die can differ from five to ten orders of magnitude. Therefore, the hottest regions determine the overall reliability of the hardware [52]. Although the majority of the proposed optimization techniques were designed for HMC-like 3D configurations, it was noticed that remapping techniques did not take into account the impact of heat from neighboring "aggressor" cells in horizontal, vertical, and diagonal directions. Moreover, in addition to information from thermal sensors, consideration of input distribution, the ReRAM cell's feature size, RCA proximity to ADC, DAC, and eDRAM, and their pitch lengths would improve the weight reordering algorithms. Such a close relationship between temperature, ReRAM technology, architecture design, and performance suggests that one of the best ways of developing thermal-aware and robust design should be solved as a MOO problem as in the case of AccuReD in *Solution 9*. As mentioned earlier, AccuReD is a heterogeneous ReRAM/GPU platform that supports both inference and training. Unlike with the majority of other accelerators, the presence of full-precision GPU in AccuReD allows execution of *Normalization* (V-norm) and *SoftMax* layers and achieves near-GPU accuracy. Along with the pipeline latency and model accuracy, its weight mapping strategy takes into account vertical and horizontal heat flows as objectives. The authors also highlight that the MOO design and optimization problem can include other objectives and be solved by different MOO solvers. Besides, *Solution 8* suggested optimizing the layout by dis-centralizing the hot components like ADCs, DACs, and eDRAM. Despite the seeming advantages, such implementation requires additional research since it leads to other challenges, e.g., reconsideration of routing and latency. Most importantly, the new layout may be incompatible with adopted chip fabrication standards.

In [53], the authors proposed electrical-thermal co-design of a multitier CIM accelerator based on heterogeneous 3D integration (H3D) using TSV. Here, the number and diameter of TSVs were varied to find an optimal point between system performance and thermal disturbance. Besides, the number of tiers in the 3D structure was also considered a variable parameter. In [44], TSV-based 3D design allows four tiers, and M3D integration has up to eight tiers when a threshold temperature is set to 373K, and therefore a preference is given to M3D due to faster heat dissipation. On top of that, one of the recent works [54] proposes benefiting from temperature and using natural biomaterials for manufacturing sustainable and pollution-free temperature-controlled ReRAM devices. These can be applied for the production of temperature-controlled sensors and detectors as well as medical treatment devices.

VI. TAKEAWAYS

Temperature-aware data allocation in ReRAM storage devices was developed to increase its reliability. Early works also applied temperature-aware weight mapping techniques on ReRAM-based DNN acceleration to restore ReRAM states and system accuracy. In addition to weight reordering in resistive crossbar arrays, recent Solutions started considering other design features, such as thermal cross-talk issues and the impact of the input distribution. Attempts to train temperature-resilient models were also made. It should be noted that thermal-aware remapping optimization designed for ReRAM can also be successfully employed in other types of non-volatile memory (NVM) technologies such as spin transfer torque magnetoresistive random access memories (STT-MRAM), Phase Change Memory (PCM), Ferroelectric RAM (FeRAM).

Thermal stability in ReRAM can also be achieved by utilization of certain materials, their composition, and structures. For instance, applying the active layer nitrogen doping technique on HfO_2-based ReRAM allows maintaining its operation at temperatures up to 550 °C [55]. Another way is increasing thickness of a buffer layer of ReRAM [56]. There is a lack on research in this area to enable RRAM without the need for off-chip solutions such as using microfluidic cooling layers [57] for decreasing the effect of temperature on the hardware, and emerging devices-based designs in particular.

As demonstrated in Section IV, the proposed thermal remapping techniques benefit in improved device performance, such as enhanced data retention, extended lifespan, and increased energy efficiency. However, implementing thermal optimization methods also brings complexity and overhead into the framework, which is typically discussed only briefly. For instance, remapping effective and ineffective weights to crossbar arrays allowed increasing accuracy by up to 39.2% for a two-layer neural network with an area overhead of up to 5% [9]. The first challenge is associated with integrating temperature sensors into ReRAM crossbar array. These include the need for material compatibility, minimal disturbance, low power consumption, and high reliability. Calibration of the thermal sensors and real-time temperature monitoring can also be sophisticated. Secondly, performing thermal control requires additional data movement and may cause data loss. Most of the proposed solutions [6], [8], [49], [58] were designed for *inference* of different workloads on ISAAC configuration which uses *naive and straightforward weight mapping* [2]. However, recent designs of accelerators [4], [18] introduced weight *reuse mapping* and support of *training phase* [4], [17], [18] that highlights the need for reevaluation of the weight remapping methods. Therefore, another issue is the scalability of the proposed methods. Moreover, the weight update process requires additional power consumption. Besides, the review has shown that there is a lack of simulation tools and frameworks designed to emulate temperature dependence in ReRAM devices, especially considering their inherent variability. In particular, proposed frameworks used thermal simulators such as 3D-ICE [59] and HotSpot [60] which still requires chip validation. In other words, temperature dynamics should be included in the ReRAM model. To sum up, the choice of thermal optimization strategy depends on the material, application, and given design constraints.

REFERENCES

[1] M. Hu et al., "Memristor-based analog computation and neural network classification with a dot product engine," *Adv. Mater.*, vol. 30, no. 9, Jan. 2018, Art. no. 1705914.

[2] A. Shafiee et al., "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," *ACM SIGARCH Comput. Architect. News*, vol. 44, no. 3, pp. 14–26, Aug. 2016.

[3] P. Chi et al., "PRIME: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory," *ACM SIGARCH Comput. Architect. News*, vol. 44, no. 3, pp. 27–39, Aug. 2016.

[4] A. Ankit et al., "PUMA: A programmable ultra-efficient memristor-based accelerator for machine learning inference," in *Proc. 24th Int. Conf. Architect. Support Program. Lang. Operat. Syst.*, 2019, pp. 715–731.

[5] K. Smagulova, M. E. Fouda, F. Kurdahi, K. N. Salama, and A. Eltawil, "Resistive neural hardware accelerators," *Proc. IEEE*, vol. 111, no. 5, pp. 500–527, May 2023.

[6] 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.

[7] C. Walczyk et al., "Impact of temperature on the resistive switching behavior of embedded HfO_2-based RRAM devices," *IEEE Trans. Electron Devices*, vol. 58, no. 9, pp. 3124–3131, Sep. 2011.

[8] M. V. Beigi, "Thermal-aware optimizations for emerging technologies in 3d-stacked chips," Ph.D. dissertation, Dept. Comput. Eng., Northwestern Univ., Evanston, IL, USA, Dec. 2019.

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

[10] P. Sun et al., "Thermal crosstalk in 3-dimensional RRAM crossbar array," *Sci. Rep.*, vol. 5, no. 1, Aug. 2015, Art. no. 13504.

[11] M. S. Al-Mamun, M. K. Orlowski, and L. N. N. Thanh, "Reliability degradation of resistive switching memory cells due to thermal crosstalk," in *Prime Archives in Electronics* . Hyderabad, TS, India: Vide Leaf, 2020.

[12] C. Zhang, Y. Ma, and P. Zhou, "Thermal-aware layout optimization and mapping methods for resistive neuromorphic engines," in *Proc. 27th Asia South Pac. Design Autom. Conf. (ASP-DAC)*, 2022, pp. 50–55.

[13] Y. Chen, "ReRAM: History, status, and future," *IEEE Trans. Electron Devices*, vol. 67, no. 4, pp. 1420–1433, Apr. 2020.

[14] Y. Chen et al., "DaDianNao: A machine-learning supercomputer," in *Proc. 47th Annu. IEEE/ACM Int. Sympos. Microarchitect.*, 2014, pp. 609–622.

[15] S. Tang et al., "AEPE: An area and power efficient RRAM crossbar-based accelerator for deep cnns," in *Proc. IEEE 6th Non-Volatile Memory Syst. Appl. Symp. (NVMSA)*, 2017, pp. 1–6.

[16] A. Nag et al., "Newton: Gravitating towards the physical limits of crossbar acceleration," *IEEE Micro*, vol. 38, no. 5, pp. 41–49, Sep./Oct. 2018.

[17] L. Song, X. Qian, H. Li, and Y. Chen, "PipeLayer: A pipelined ReRAM-based accelerator for deep learning," in *Proc. IEEE Int. Symp. High Perform. Comput. Architect. (HPCA)*, 2017, pp. 541–552.

[18] X. Qiao, X. Cao, H. Yang, L. Song, and H. Li, "AtomLayer: A universal ReRAM-based CNN accelerator with atomic layer computation," in *Proc. 55th ACM/ESDA/IEEE Design Autom. Conf.*, 2018, pp. 1–6.

[19] A. Ankit et al., "PANTHER: A programmable architecture for neural network training harnessing energy-efficient ReRAM," *IEEE Trans. Comput.*, vol. 69, no. 8, pp. 1128–1142, Aug. 2020.

[20] J. H. Lau, *Semiconductor Advanced Packaging*. Berlin, Germany: Springer, 2021.

[21] Y. Cheng, X. Guo, and V. F. Pavlidis, "Emerging monolithic 3D integration: Opportunities and challenges from the computer system perspective," *Integration*, vol. 85, pp. 97–107, Jul. 2022.

[22] B. Hudec et al., "3D resistive RAM cell design for high-density storage class memory—a review," *Sci. China Inf. Sci.*, vol. 59, no. 6, pp. 1–21, Jun. 2016.

[23] (Google Technol. Co., Mountain View, CA, USA). *Google Demonstrates Leading Performance in Latest MLPerf Benchmarks*. Accessed: Aug. 10, 2021. [Online]. Available: https://cloud.google.com/blog/products/ai-machine-learning/google-wins-mlperf-benchmarks-with-tpu-v4

[24] D. Lacey, (Graphcore Semicond. Co., Bristol, U.K.). *Updated Graphcore IPU Benchmarks*. Accessed: Aug. 10, 2021. [Online]. Available: https://www.graphcore.ai/posts/new-graphcore-ipu-benchmarks

[25] L. Gwennap, "Groq rocks neural networks." Jan. 2020, Accessed: Aug. 10, 2021. [Online]. Available: https://groq.com/wp-content/uploads/2020/04/Groq-Rocks-NNs-Linley-Group-MPR-2020Jan06.pdf

[26] C. Campa, C. Kawalek, H. Vo, and J. Bessoudo, (Nvidia Softw. Co., Santa Clara, CA, USA). *Defining AI Innovation With NVIDIA DGX A100*. Jul. 2021. [Online]. Available: https://developer.nvidia.com/blog/defining-ai-innovation-with-dgx-a100/
[27] J. Choquette, "NVIDIA hopper H100 GPU: Scaling performance," *IEEE Micro*, vol. 43, no. 3, pp. 9–17, May/Jun. 2023.
[28] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nat. Electron.*, vol. 1, no. 6, pp. 333–343, 2018.
[29] P.-Y. Chen and S. Yu, "Compact modeling of RRAM devices and its applications in 1T1R and 1S1R array design," *IEEE Trans. Electron Devices*, vol. 62, no. 12, pp. 4022–4028, Dec. 2015.
[30] W. Shim, J. Meng, X. Peng, J.-S. Seo, and S. Yu, "Impact of multilevel retention characteristics on RRAM based DNN inference engine," in *Proc. IEEE Int. Reli. Phys. Symp. (IRPS)*, 2021, pp. 1–4.
[31] C. Wang, H. Wu, B. Gao, T. Zhang, Y. Yang, and H. Qian, "Conduction mechanisms, dynamics and stability in ReRAMS," *Microelectron. Eng.*, vols. 187–188, pp. 121–133, Feb. 2018.
[32] W.-M. Chung et al., "A study of the relationship between endurance and retention reliability for a HFO_X-based resistive switching memory," *IEEE Trans. Device Mater. Rel.*, vol. 20, no. 3, pp. 541–547, Sep. 2020.
[33] 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.
[34] D. B. Strukov, "Endurance-write-speed tradeoffs in nonvolatile memories," *Appl. Phys. A*, vol. 122, no. 4, pp. 1–4, Mar. 2016.
[35] A. V. Fadeev and K. V. Rudenko, "To the issue of the memristor's HRS and LRS states degradation and data retention time," *Russian Microelectron.*, vol. 50, no. 5, pp. 311–325, 2021.
[36] Y. Yu and N. K. Jha, "Energy-efficient monolithic three-dimensional on-chip memory architectures," *IEEE Trans. Nanotechnol.*, vol. 17, no. 4, pp. 620–633, Jul. 2018.
[37] K. Dhananjay, P. Shukla, V. F. Pavlidis, A. Coskun, and E. Salman, "Monolithic 3D integrated circuits: Recent trends and future prospects," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 3, pp. 837–843, Mar. 2021.
[38] A. Kaul, Y. Luo, X. Peng, S. Yu, and M. S. Bakir, "Thermal reliability considerations of resistive synaptic devices for 3D CIM system performance," in *Proc. IEEE Int. 3D Syst. Integr. Conf. (3DIC)*, 2021, pp. 1–5.
[39] Y. Sato, K. Kinoshita, M. Aoki, and Y. Sugiyama, "Consideration of switching mechanism of binary metal oxide resistive junctions using a thermal reaction model," *Appl. Phys. Lett.*, vol. 90, no. 3, Jan. 2007, Art. no. 033503.
[40] T. D. Dongale et al., "Investigating the temperature effects on resistive random access memory (RRAM) devices," 2016, *arXiv:1602.08262*.
[41] H.-Y. Kao, S.-H. Huang, and W.-K. Cheng, "Design framework for ReRAM-based DNN accelerators with accuracy and hardware evaluation," *Electronics*, vol. 11, no. 13, p. 2107, Jul. 2022.
[42] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1737–1746.
[43] J. Zhou et al., "DyBit: Dynamic bit-precision numbers for efficient quantized neural network inference," 2023, *arXiv:2302.12510*.
[44] B. K. Joardar, J. R. Doppa, P. P. Pande, H. Li, and K. Chakrabarty, "AccuReD: High accuracy training of CNNs on ReRAM/GPU heterogeneous 3-D architecture," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 40, no. 5, pp. 971–984, May 2021.
[45] X. Yang et al., "Multi-objective optimization of ReRAM crossbars for robust DNN inferencing under stochastic noise," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, 2021, pp. 1–9.
[46] M. Zhou, M. Imani, S. Gupta, and T. Rosing, "Thermal-aware design and management for search-based in-memory acceleration," in *Proc. 56th ACM/IEEE Design Autom. Conf.*, 2019, pp. 1–6.
[47] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–9.
[48] H. Shin, M. Kang, and L.-S. Kim, "A thermal-aware optimization framework for ReRAM-based deep neural network acceleration," in *Proc. 39th IEEE/ACM Int. Conf. Comput.-Aided Design*, 2020, pp. 1–9.
[49] 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 Eur. Conf. Exhib. (DATE)*, 2022, pp. 1245–1250.
[50] S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb, "A simulated annealing-based multiobjective optimization algorithm: AMOSA," *IEEE Trans. Evol. Comput.*, vol. 12, no. 3, pp. 269–283, Jun. 2008.
[51] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[52] A. L. Moore and L. Shi, "Emerging challenges and materials for thermal management of electronics," *Mater. Today*, vol. 17, no. 4, pp. 163–174, May 2014.
[53] X. Peng, A. Kaul, M. S. Bakir, and S. Yu, "Heterogeneous 3-D integration of multitier compute-in-memory accelerators: An electrical-thermal co-design," *IEEE Trans. Electron Devices*, vol. 68, no. 11, pp. 5598–5605, Nov. 2021.
[54] B. Sun, G. Zhou, T. Yu, Y. Chen, F. Yang, and Y. Zhao, "Multi-factors-controlled ReRAM devices and their applications," *J. Mater. Chem. C*, vol. 10, pp. 8895–8921, May 2022.
[55] J. Park, E. Park, and H.-Y. Yu, "Active layer nitrogen doping technique with excellent thermal stability for resistive switching memristor," *Appl. Surf. Sci.*, vol. 603, Nov. 2022, Art. no. 154307.
[56] H. He, Y. Tan, C. Lee, and Y. Zhao, "Ti/HfO2-based RRAM with superior thermal stability based on self-limited TiOx," *Electronics*, vol. 12, no. 11, p. 2426, 2023.
[57] Y. Zhang, A. Dembla, Y. Joshi, and M. S. Bakir, "3D stacked microfluidic cooling for high-performance 3D ICS," in *Proc. IEEE 62nd Electron. Compon. Technol. Conf.*, 2012, pp. 1644–1650.
[58] T. Abbey, C. Giotis, A. Serb, S. Stathopoulos, and T. Prodromakis, "Thermal effects on initial volatile response and relaxation dynamics of resistive RAM devices," *IEEE Electron Device Lett.*, vol. 43, no. 3, pp. 386–389, Mar. 2022.
[59] A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza, "3D-ICE: Fast compact transient thermal modeling for 3D ICS with inter-tier liquid cooling," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, 2010, pp. 463–470.
[60] K. Skadron, M. R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, and D. Tarjan, "Temperature-aware microarchitecture: Modeling and implementation," *ACM Trans. Architect. Code Optim. (TACO)*, vol. 1, no. 1, pp. 94–125, 2004.