SCIENCE CHINA
Information Sciences

CrossMark

• RESEARCH PAPER •

June 2026, Vol. 69, Iss. 6, 162404:1–162404:11
https://doi.org/10.1007/s11432-025-4708-9

Thermal analysis and management for RRAM-based computing-in-memory chips

Awang MA, Bin GAO*, Ruihua YU, Qingtian ZHANG, Peng YAO, Jianshi TANG, He QIAN & Huaqiang WU

School of Integrated Circuits, Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China

Received 25 June 2025/Revised 19 September 2025/Accepted 3 December 2025/Published online 8 May 2026

Abstract In recent years, artificial intelligence (AI) has experienced rapid development, and high performance computing (HPC) has raised increasingly higher demands for hardware computational capacity. Resistive random-access memory (RRAM)-based computing-in-memory (CIM) technology is expected to overcome the bottleneck of memory wall and provide HPC solutions. However, CIM chips face critical thermal challenges, including severe hotspot formation and thermally induced performance degradation, due to increasing power density and strong data-space coupling effects. Existing thermal management solutions designed for conventional digital chips are not directly applicable to CIM architectures. In this work, we propose a comprehensive framework for thermal analysis and management tailored to CIM chips. Targeted strategies are developed across the design, pre-operation, and operation stages. During the design stage, it is essential to mitigate thermal-induced accuracy degradation by adopting optimized design strategies. During the pre-operation stage, we propose a latency-thermal co-optimization (LTCO) strategy for static thermal management. By combining LTCO with a genetic algorithm to optimize the neural network mapping scheme, we reduce the hotspot temperature by 6.5°C and the temperature standard deviation by 5.3°C, without increasing the latency. During the operation stage, we develop a dynamic thermal management (DTM) strategy tailored for RRAM-based CIM chips, considering their unique architecture and the coupling between data and space. The evaluation results show that when thermal management is triggered, the combination of LTCO and DTM achieves more than a 10% performance improvement over DTM alone.

Keywords RRAM-based CIM chips, thermal analysis, static thermal management, latency-thermal co-optimization, dynamic thermal management

Citation Ma A W, Gao B, Yu R H, et al. Thermal analysis and management for RRAM-based computing-in-memory chips. Sci China Inf Sci, 2026, 69(6): 162404, https://doi.org/10.1007/s11432-025-4708-9

1 Introduction

AI technologies, such as large language models (LLMs), autonomous driving, and intelligent assistants, have developed rapidly in recent years [1–3], leading to significantly higher demands for computational capabilities in hardware devices [4]. However, the development of integrated circuits has been constrained by bottlenecks such as the memory wall and power wall, making it challenging to sustain computational growth through traditional scaling down of circuit dimensions [5–7]. In the conventional von Neumann architecture, the physical separation of memory and computation results in significant power consumption and latency consumption due to frequent data transfers, a phenomenon known as the memory wall [8–10]. Currently, the latest advancements in chip technology include the use of 2.5D and 3D advanced packaging techniques, exemplified by products like NVIDIA's A100 and H100 [11,12]. These technologies help reduce the physical distance between computation and memory, alleviating the limitations imposed by the memory wall [13–15]. Furthermore, the computing-in-memory (CIM) technology based on emerging non-volatile memory (eNVM) devices, such as resistive random-access memory (RRAM), integrates memory and computation functions within a single device [16–18]. A lot of studies have demonstrated in-memory computing on RRAM-based CIM chips, showcasing their great potential for future high-performance and energy-efficient computing systems [19–25]. Moreover, CIM technology is considered one of the most promising solutions to overcome the memory wall [26].

However, the power wall significantly limits the performance of CIM chips. Studies indicate that chips typically operate at temperatures between 85°C and 100°C, and even a 10°C increase beyond this range, operational reliability of electronic chips can decrease by as much as 50% [27]. Although CIM technology offers higher energy efficiency

* Corresponding author (email: gaob1@tsinghua.edu.cn)

© Science China Press 2026

compared to traditional chips [28, 29], it still faces the power wall constraint in high-performance computing applications, and the power density of CIM chips has been increasing rapidly in these years [30]. During the calculation process of matrix-vector multiplication, the RRAM arrays will be fully opened in parallel [31–34], causing high local power and high local temperatures. And thermal issue becomes even more severe under 2.5D and 3D advanced packaging technologies [35]. Without effective thermal management technology, CIM chips will be damaged by high temperatures. Meanwhile, RRAM-based CIM technology falls within the analog computing paradigm, and the precision and reliability of RRAM-based CIM chips are sensitive to temperature [36, 37]. Furthermore, edge applications represent an important scenario. When used outdoors, the high environmental temperature will significantly increase the difficulty of thermal management. Therefore, it is of great importance to find effective thermal management technology for RRAM-based CIM chips.

There are various thermal management solutions, for example, dynamic voltage and frequency scaling (DVFS) [38], task allocation [39], and memory access control [40]. CIM chips have a fundamentally different computational architecture, and the thermal management solutions designed for digital chips are not suitable for the thermal management of CIM chips. Existing related studies mainly focus on thermal modeling for RRAM devices and thermal simulation for CIM chips. Some compact models have been proposed for RRAM devices, such as the compact model about temperature coefficient of RRAM conductance [36], retention characteristics of RRAM [37] and an endurance model that accounts for temperature effects [41]. Some existing studies focus on thermal modeling of CIM chips and evaluation of temperature effects on chip inference accuracy and reliability [42–44]. Thermal management for search-based in-memory acceleration has also been proposed [45], when RRAM-based hybrid memory cube is used for storage of data vector. However, thermal management solutions for RRAM-based CIM chips are still missing.

In this work, we present a unified framework aimed at addressing the thermal challenges of RRAM-based CIM chips, focusing on issues such as hotspot formation, thermally induced performance loss, and data-space coupling. Targeted optimization strategies are devised for three critical phases: chip design, pre-operation, and operation. During the design phase, thermal-induced accuracy degradation is mitigated through architectural optimizations. In the pre-operation phase, a latency-thermal co-optimization (LTCO) method is introduced, where a genetic algorithm is applied to optimize neural network mapping. This approach lowers the hotspot temperature by 6.5°C and reduces the temperature standard deviation by 5.3°C, without incurring additional latency. For the operation phase, a dynamic thermal management (DTM) strategy is developed, specifically adapted to the unique characteristics and data-space interaction of RRAM-based CIM architectures. The evaluation results show that when thermal management is triggered, the combination of LTCO and DTM achieves more than a 10% performance improvement over DTM alone.

2 Background

The human brain performs highly complex computations with remarkable energy efficiency, utilizing 10^{11} neurons and 10^{15} synapses [46]. As shown in Figure 1(a), RRAM is similar to biological synapses in that it can change its state through ion movement and process analog signals [27, 34]. RRAM is one type of eNVM device with a simple sandwich structure, including a top electrode (TE), a resistive switching layer, and a bottom electrode (BE). It can also be integrated into the back end of CMOS technology. The RRAM used in this study is the $TiN/HfO_x/TaO_y/TiN$ memristor. As shown in Figure 1(b), parallel computations can be performed based on Ohm's law and Kirchhoff's law. We designed a multi-level architecture for the RRAM-based CIM chip used in modeling. As Figure 1(b) shows, the CIM chip includes nine tiles, an I/O interface and DTM modules. These modules are all connected via the data bus. The chip exchanges data with the external device via the InOut interface module and performs thermal management through the DTM module. Every tile houses four processing elements (PEs) and corresponding peripheral modules, including SRAM buffer, pooling, activation function, shift and adder, tile controller, and more. Each PE consists of an RRAM array crossbar with peripheral circuits, such as digital-to-analog converters (DACs) and analog-to-digital converters (ADCs), and others. Each PE is equipped with a temperature sensor. Joule heat is primarily dissipated into the environment through the upper side and the thermal resistance is set at 4.92 $cm^2{\cdot}K/W$, while the four sides and bottom dissipate very little heat and thermal resistance is approximated as adiabatic. The ambient temperature is assumed to be 26.85°C. In addition, the die size is 10 mm × 10 mm. The die material is silicon. And the operating clock frequency is 100 MHz. The thermal simulations were performed with our in-house tool developed on the MATLAB platform, as detailed in [36].

In recent years, CIM technology has been widely studied and many chips have been developed [47–52]. CIM chips have evolved from multiple array-level demos to macro-level demos and are now progressing toward full system in one chip designs. CIM technology has the potential to address the memory wall, making it a popular choice for AI

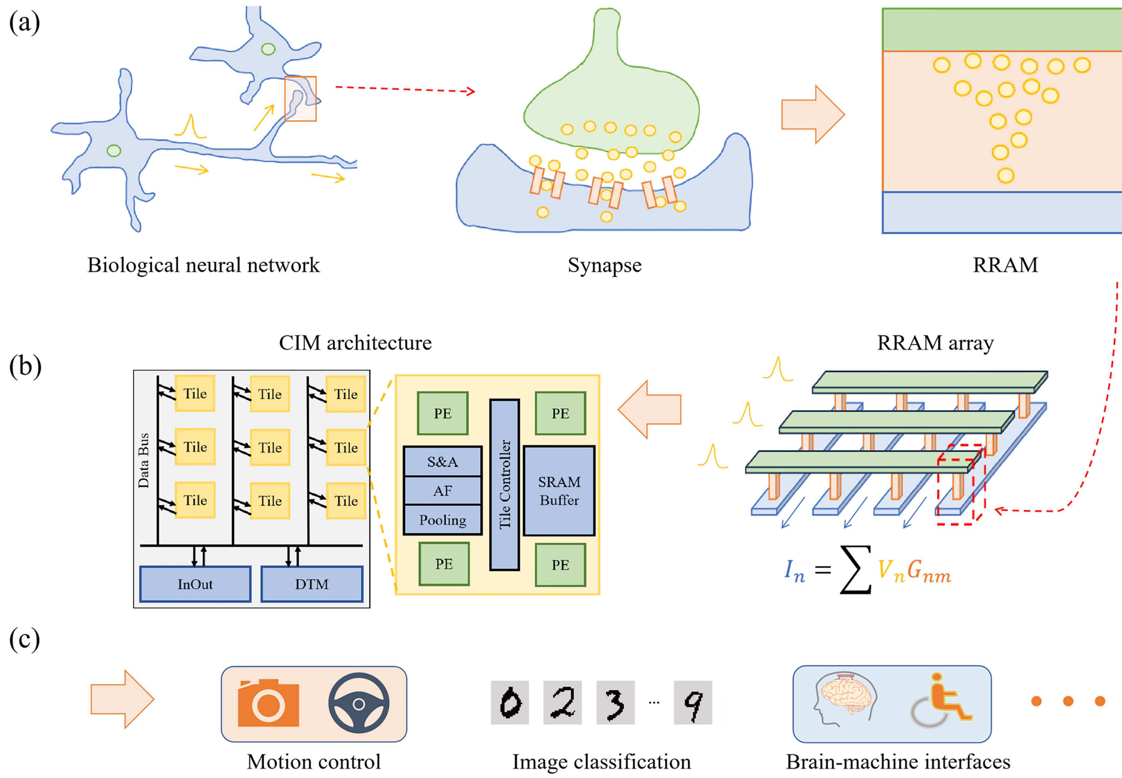

Figure 1 (Color online) Principles and applications of RRAM-based CIM chips. (a) Motivations for using RRAM in neuromorphic computing. RRAM devices can process analog signals similar to neural synapses. (b) RRAM array and CIM architecture. Based on Kirchhoff's law and Ohm's law, RRAM arrays can perform parallel computations just like the human brain. (c) Applications of RRAM-based CIM chips, including motion control, image classification, brain-machine interfaces, and more.

Table 1 Parameters of RRAM devices.

Parameter	Value
Material	$TiN/HfO_x/TaO_y/TiN$
Conductance range	2–20 µS
Conductance states	16

Table 2 Parameters of the CIM chip.

Parameter	Value
Die size	10 mm × 10 mm
Material	Silicon
Clock frequency	100 MHz
Thermal resistance	4.92 cm^2·K/W
Ambient temperature	26.85°C

applications (Figure 1(c)), including motion control [53], image classification [54, 55], image reconstruction [56, 57], brain-machine interfaces [58], homomorphic encryption [59], physical reservoir computing [60], and more.

The CIM chip in this work uses HfO_x-based analog RRAM devices and the corresponding parameters are listed in Table 1. Conductance range of RRAM device is between 2 and 20 µS, and each RRAM device has 16 independent conductance states. Thermal simulation parameters of the chip are listed in Table 2. The parameters for the RRAM device and the RRAM-based CIM chips used in this work are extracted from a real 28 nm CIM chip, as referenced in [36]. While variations in materials and architectures may lead to differences in specific parameters for RRAM-based CIM chips, the approach proposed in this study is broadly applicable.

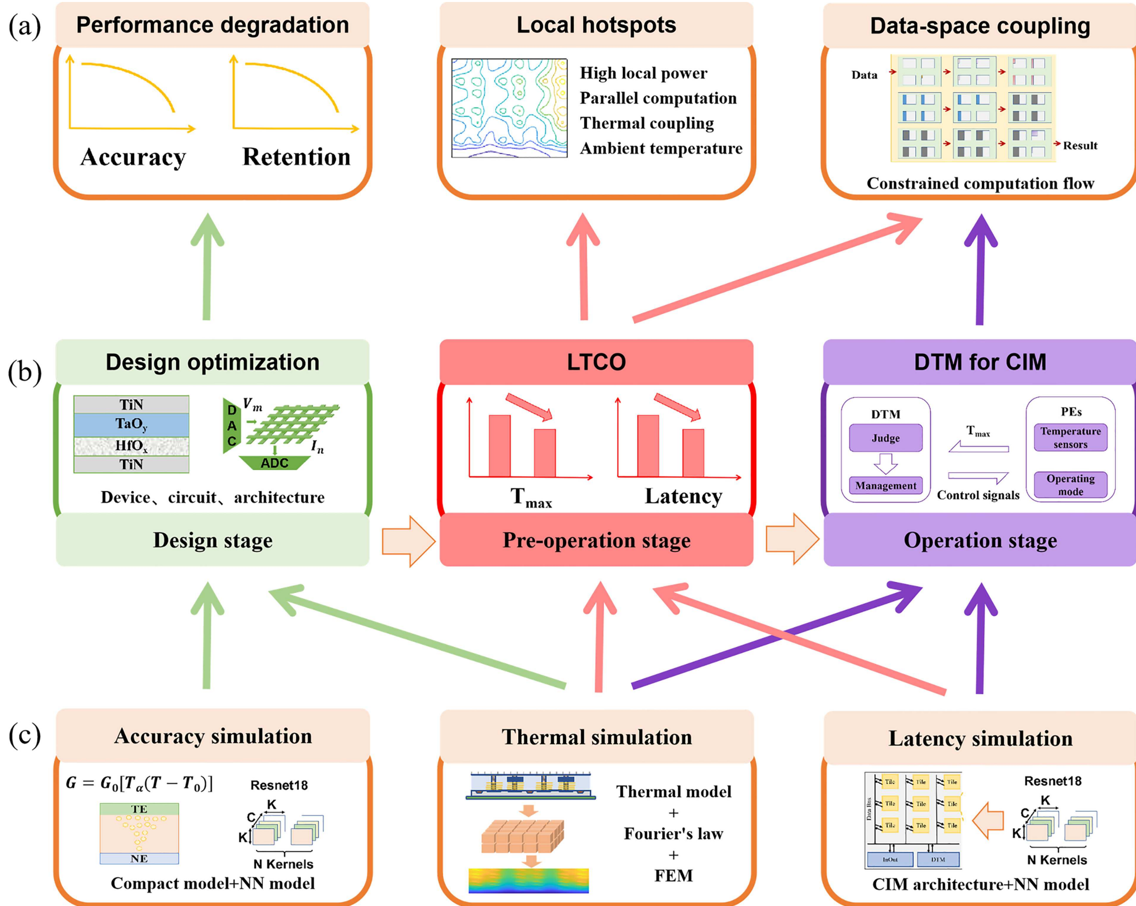

Figure 2 (Color online) Framework of thermal analysis and management. (a) The thermal issues in RRAM-based CIM chips are categorized into three aspects, including local hotspots, thermally induced performance degradation and data-space coupling. (b) Proposed thermal optimization and management methods across three stages: design stage, pre-operation stage, and operation stage. (c) The core modules involved in the optimization schemes include accuracy simulation, thermal simulation, and latency simulation.

3 Framework of thermal analysis and management

In recent years, the CIM technology has developed rapidly. However, the thermal issues in RRAM-based CIM chips are becoming increasingly severe. First, the temperature distribution within the RRAM-based CIM chip is non-uniform. As a typical PE in the CIM architecture, the RRAM array is surrounded by ADCs and DACs. On the one hand, PEs account for over 70% of the power consumption, yet occupy less than 20% of the chip area, resulting in high local power and high local temperatures in the PE area. On the other hand, thermal crosstalk between PEs leads to even higher local temperatures. Second, both the characteristics of RRAM devices and the overall chip performance are sensitive to temperature. The conductance of the analog RRAM device drifts with temperature. In particular, when analog devices are used to represent multiple bits, the conductance drift leads to degradation of the computing accuracy of CIM chips. High temperature also reduces the retention time of RRAM devices. As a result, the computing accuracy of CIM chips decreases with time under high temperatures. Finally, data-space coupling exists in the computing of CIM chips. The parameters of the neural network are mapped to different RRAM arrays on the chip before inference computation begins. The data stored in each RRAM array typically does not change during computation. Furthermore, the entire neural network must complete its computation across the chip before the final results can be obtained. Data-space coupling significantly complicates thermal management.

To address these issues, we propose a comprehensive framework for the thermal analysis and management of RRAM-based CIM chips. As shown in Figure 2, the framework is composed of three sections. Figure 2(a) identifies the major challenges encountered by current CIM chips, including local hotspots, thermally induced performance degradation and data-space coupling. Figure 2(b) proposes targeted solutions across the three key phases of chip development, including the design, pre-operation, and operation stages. Figure 2(c) summarizes the primary simulation modules utilized for each solution outlined in Figure 2(a).

The proposed framework integrates accuracy, thermal, and latency simulators. The accuracy simulator is imple-

mented as a neural network program trained on the Python platform, which is used to evaluate inference accuracy. The thermal simulator is an in-house tool developed on the MATLAB platform for three-dimensional thermal analysis of RRAM-based CIM chips. More detailed descriptions of the thermal modeling framework for RRAM-based CIM chips can be found in our previously published work [36]. The latency simulator, also developed on the MATLAB platform, estimates task execution time by considering factors such as data flow, bandwidth, and neural network architecture.

During the design stage, thermal optimization of the CIM chip performance can be performed based on inference accuracy simulations and thermal simulations. The compact model of RRAM devices describes how device characteristics vary with temperature and enables the integration of temperature distributions from thermal simulations with neural network parameters mapped onto CIM chips. This allows analysis of device- and chip-level performance degradation caused by temperature. Design optimization is then conducted at multiple levels—including device, circuit, and chip architecture—to mitigate heat-induced performance loss.

Lowering the operating temperature can reduce conductance drift and thus mitigate accuracy loss. In our previous study [44], reducing the temperature from 87°C to 77°C decreased the accuracy loss by 5.8%; nevertheless, the overall degradation remained at 61.7%. To address this, prior studies [36,44] proposed thermal-aware and ADC quantization correction techniques, which reduced accuracy loss to below 1.4%. In addition to the discussions in [36,44] regarding the impact of temperature-induced conductance drift on inference accuracy and corresponding correction methods, our earlier study [37] investigated the effect of elevated temperatures on RRAM retention characteristics and system-level inference accuracy. Given these detailed prior studies, this paper focuses on the refinement of thermal management architecture, where temperature-induced accuracy loss and corresponding optimization strategies are considered during the design stage.

During the pre-operation stage, both latency and thermal simulations can be jointly optimized to achieve the best performance in terms of both latency and temperature. The emergence of hotspots in RRAM-based CIM chips is primarily attributed to the high local power consumption of PEs and thermal crosstalk between them. Typically, mapping schemes optimized solely for latency tend to concentrate power consumption, leading to higher local hotspots. Conversely, schemes focused purely on thermal optimization often neglect latency considerations, resulting in unacceptably high delays. In this work, we propose an LTCO strategy, which integrates latency and thermal simulations using a genetic algorithm to achieve a balance between latency and thermal performance.

During the operation stage, thermal and latency simulations are used to design effective dynamic thermal management strategies. The difficulty of thermal management in RRAM-based CIM chips during actual operation mainly stems from data-space coupling—the parameters mapped onto the chip remain fixed in location during computation, and a full pass over all data is required to produce results. When the chip temperature becomes excessively high and thermal management actions must be taken, significant performance loss often occurs. To address this issue, we have developed a DTM strategy specifically for RRAM-based CIM chips. This approach enables thermal management with minimal performance degradation when managing overheating scenarios.

4 Static thermal management

4.1 Introduction of LTCO

In the pre-operation stage, we need to map parameters of the NN onto RRAM arrays on the CIM chip. During the inference operation stage, the parameters mapped on RRAM arrays are not adjusted, as such adjustments would consume significant time. The formation of hotspots in RRAM-based CIM chips largely results from the high localized power consumption of PEs and the associated thermal crosstalk. To achieve optimal latency, the parameters of each layer of the neural network are usually mapped to nearby RRAM arrays. Latency-optimized mapping strategies often exacerbate this issue by concentrating power in specific regions, thereby intensifying local hotspots. On the other hand, strategies aimed solely at thermal optimization tend to disregard latency constraints, leading to unacceptable increases in delay. To address these challenges, we propose an LTCO strategy for static thermal management, aimed at reducing hotspot temperatures and improving temperature uniformity across the CIM chip. Built upon a genetic algorithm framework, LTCO simultaneously considers both latency and thermal performance in RRAM-based CIM chips. This approach effectively avoids the latency penalties often associated with purely thermal-driven optimization, while also mitigating the severe hotspot issues that arise from latency-focused strategies. Furthermore, we have tailored the algorithm to accommodate the unique characteristics of RRAM-based CIM architectures, ensuring rapid convergence and practical applicability. The flow of the LTCO method is displayed in Figure 3(a). A classical ResNet-18 [61] is used to classify the CIFAR-10 dataset, demonstrating the

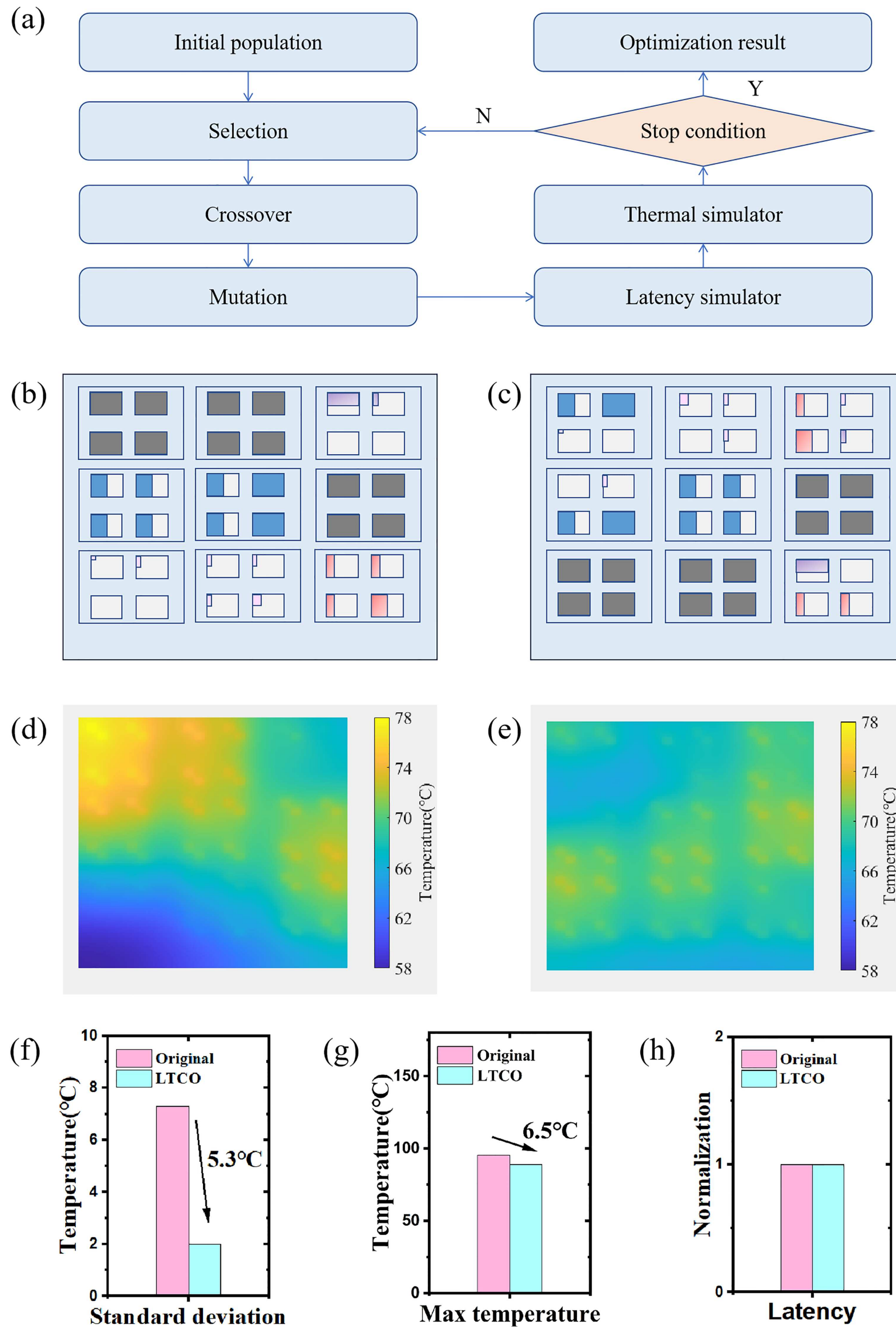

Figure 3 (Color online) The flow and evaluation results of static thermal management LTCO. (a) The flow of the LTCO method; (b) the weight arrangement of one mapping scheme before optimization; (c) the weight arrangement of the optimized mapping scheme; (d) the temperature map of the mapping scheme before optimization; (e) the temperature map of the optimized mapping scheme; (f) comparison of the standard deviation of temperature; (g) comparison of maximum temperature; (h) comparison of latency.

application of this method.

During the initialization phase, we analyze the neural network's parameter characteristics and generate an initial set of mapping schemes. As different layers, each with varying parameter sizes, are assigned to different PEs, the number of active ADCs, DACs, and RRAM devices changes accordingly, resulting in power consumption differences across PEs. A key design principle is to map layers with large numbers of parameters into a single PE whenever possible. For instance, in ResNet-18, three layers contain particularly large parameter sets, each requiring distribution across four PEs. To minimize latency, it is preferable to place all PEs associated with a single layer within the same tile when feasible.

The subsequent optimization process involves selection, crossover, and mutation operations. Mutation is per-

formed by swapping the positions of two PEs mapped by different layers, while crossover exchanges the positions of multiple PEs simultaneously. In each iteration, either a crossover or a mutation is conducted. Considering the architecture of RRAM-based CIM chips and the characteristics of the ResNet-18 model, layers occupying the same number of PEs and exhibiting similar power consumption are grouped together. Swapping their positions typically yields little thermal benefit but may negatively impact latency. To expedite convergence, layers with high power consumption and spanning multiple PEs are first fixed in position, while adjustments are made only to the remaining layers.

Following initialization, we perform both latency and thermal simulations. Each candidate mapping is evaluated using our developed thermal simulator to determine the maximum temperature across the chip. Latency simulation assesses communication delays based on bus bandwidth and total data throughput. To further accelerate the optimization, latency simulation is conducted first. If a candidate mapping results in latency exceeding a predefined threshold, thermal simulation is skipped and the associated mutation is discarded, retaining the previous mapping.

Finally, a decision is made after each iteration. If a mapping scheme achieves a lower maximum temperature without increasing latency, it is accepted; otherwise, the previous scheme is preserved. Iterations are terminated if no improvements are observed over several consecutive attempts. The final optimized mapping scheme is then selected from the resulting candidates, along with its corresponding temperature map and estimated latency. The overarching goal of the LTCO approach is to optimize the temperature distribution of the RRAM-based CIM chip without introducing additional latency. To further enhance convergence efficiency, the optimization process incorporates the power distribution characteristics of neural network layers and the symmetry of the chip architecture.

4.2 Evaluation results

To evaluate the effectiveness of the proposed LTCO method, we employ a classical ResNet-18 model to classify the CIFAR-10 dataset. An initial set of weight mapping schemes is generated, with one example shown in Figure 3(b).

Prior to optimization, weights are mapped sequentially according to the layer order. This approach achieves relatively low latency. However, it leads to unfavorable thermal characteristics. Specifically, layers with high power consumption are placed adjacent to each other, as are layers with low power consumption, resulting in a maximum chip temperature of 95.3°C, as illustrated in Figure 3(d). On one hand, this temperature exceeds the safe operating limit of 85°C. On the other hand, the thermal map reveals significant non-uniformity, creating large temperature gradients that could induce mechanical stress and jeopardize the reliable operation of the CIM chip.

After applying the LTCO optimization, the weight distribution is adjusted to jointly consider latency and thermal performance. As depicted in Figure 3(c), layers with higher power consumption are spread more evenly across the chip and placed closer to low-power digital regions, while still maintaining the proximity of adjacent layers to preserve low latency. Although further dispersing high-power layers could further reduce the peak temperature, it would cause an unacceptable increase in latency. Post-optimization, the temperature distribution becomes significantly more uniform. As shown in Figure 3(e), compared to the initial mapping, the LTCO method reduces the standard deviation of the temperature by 5.3°C (Figure 3(f)) and lowers the maximum temperature by 6.5°C (Figure 3(g)), all without increasing the overall latency (Figure 3(h)). Specifically, during static thermal management, only the parameter mapping strategy is adjusted, while the total power consumption, boundary conditions, and chip material properties remain unchanged. The primary change occurs in the spatial distribution of power across PEs. Before optimization, PEs with high power density tend to cluster together, as do PEs with low power density. This clustering results in local hotspots and significant on-chip temperature gradients. Following optimization, while maintaining latency constraints, the power distribution among PEs becomes more uniform, effectively reducing hotspot temperatures and mitigating on-chip thermal variations.

Further insights into the construction of the LTCO approach are as follows. Three key factors contribute to the feasibility of latency-thermal co-optimization. First, if the number of available arrays on the chip exceeds the number required by the neural network, the redundant arrays can be strategically utilized to optimize thermal distribution without affecting latency. Second, the CIM architecture exhibits inherent symmetry, where tiles are interconnected via a common bus. Thus, altering the data mapping among tiles can change the thermal profile without impacting communication latency. Third, although the four PEs within each tile are equivalent from a latency perspective, they differ thermally. This asymmetry allows additional flexibility in optimizing the temperature distribution. Together, these factors enable effective temperature optimization without sacrificing latency in RRAM-based CIM chips. Additionally, in this work, our primary focus is on optimizing the temperature profile, with latency constrained to remain unchanged. If desired, the LTCO convergence conditions could be modified to require simultaneous reductions in both latency and temperature, allowing for further overall performance improvements.

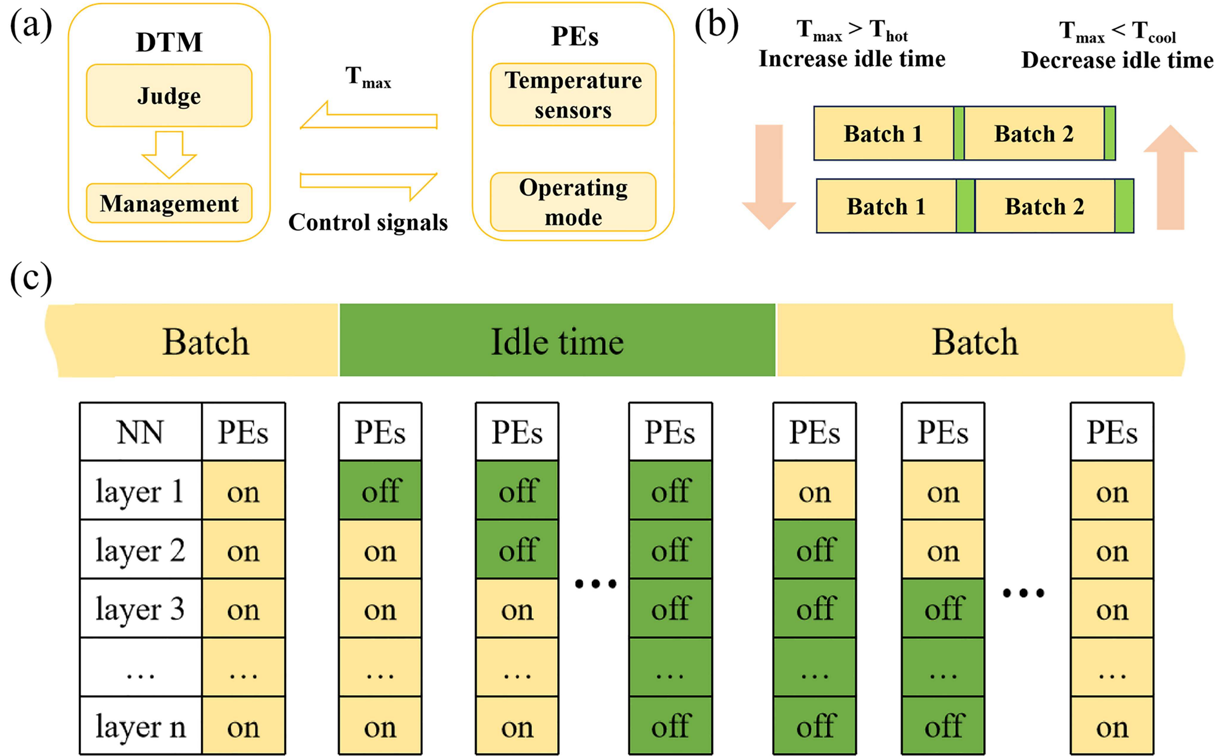

Figure 4 (Color online) The method of dynamic thermal management. (a) Architecture of DTM; (b) schematic diagram of idle time control; (c) the switching control of PEs.

5 Dynamic thermal management

5.1 Introduction of DTM for CIM chips

During the operational stage, when the maximum temperature of RRAM-based CIM chips exceeds the safe operating threshold, effective thermal management becomes critical to maintain reliable chip function while minimizing performance degradation.

Traditional thermal management techniques for digital chips typically involve adjusting voltage and frequency settings or redistributing workloads across computational cores to balance temperature profiles. However, such methods are unsuitable for CIM chips. Modifying voltage and frequency during operation would lead to significant performance losses. And CIM chips generally operate in a fixed mode, executing the same task repeatedly without dynamic workload variation.

RRAM devices offer unique advantages in this context. As non-volatile memory elements, RRAMs retain stored data even when the chip is powered off and restarted. Moreover, RRAM-based CIM chips often process discrete tasks, such as image recognition, where controllable time intervals between tasks are acceptable and do not compromise task accuracy or system throughput.

Leveraging these characteristics, we propose a DTM technique specifically designed for RRAM-based CIM chips, as illustrated in Figure 4(a). Each PE is equipped with a temperature sensor, and temperature readings are transmitted via the data bus to a centralized DTM unit. Two temperature thresholds are defined: T_{hot}, representing the upper safe temperature limit (typically set to 85°C), and T_{cool}, a lower threshold (typically 80°C). These thresholds are adjustable to prevent excessive system response to minor temperature fluctuations.

As shown in Figure 4(b), if the measured maximum temperature (T_{max}) exceeds T_{hot}, the system increases the idle interval between two adjacent batches of computation. This effectively reduces the average power consumption and lowers the maximum temperature of the chip. Conversely, if T_{max} falls below T_{cool}, the system shortens the idle time, allowing higher average power consumption and thus improving computational performance. As operationally shown in Figure 4(c), after each batch of computation is completed, the system enters an idle phase during which PEs that have finished processing are sequentially powered down. At the end of the idle period, PEs are sequentially reactivated to resume data computation. This approach successfully lowers the average power consumption of the chip during overheating events while minimizing disruption to the normal data flow.

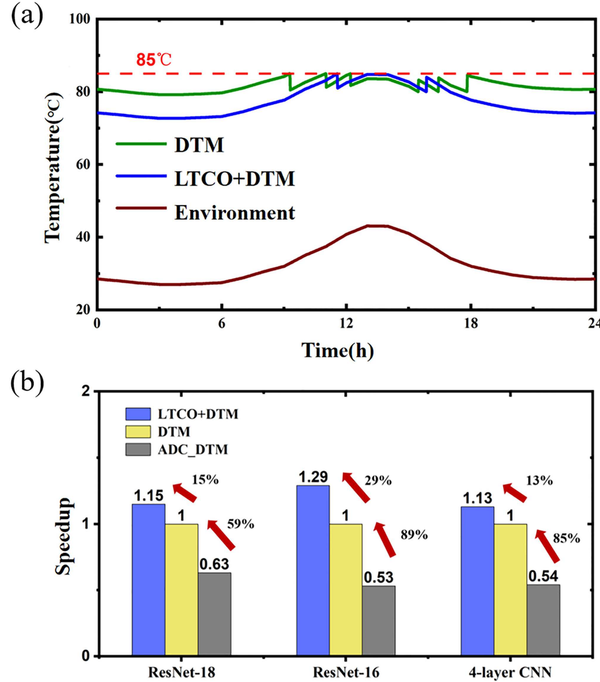

Figure 5 (Color online) Results of dynamic thermal management. (a) Maximum temperature of DTM; (b) speedup of RRAM-based CIM chips.

5.2 Analysis of test results

As shown in Figure 5(a), the maximum temperature curve of the CIM chip, using the proposed DTM for CIM, remains generally smooth. When the temperature reaches 85°C, the system responds promptly by adjusting the idle time multiple times to maintain the chip's operation near the maximum safe temperature, without causing power fluctuations. When both the LTCO and DTM for CIM chips are used, the maximum temperature is further reduced, and the duration of DTM operation is shorter compared to when LTCO is not applied.

Next, we evaluate the throughput of the CIM chip under different thermal management schemes. Three neural networks—ResNet-18, ResNet-16, and 4-Layer CNN—are selected for analysis. We compare three scenarios: using the proposed DTM for CIM chips alone, using the ADC-based DTM scheme, and combining the DTM for CIM chips with LTCO. The ADC-based DTM scheme regulates the chip temperature by adjusting the number of activated ADCs, thereby modulating both the power consumption of the ADCs and that of the associated RRAM arrays. The number of active ADCs directly determines the computational parallelism of the chip; consequently, reducing the number of active ADCs inevitably prolongs the execution time of a given task. This trade-off between thermal regulation and computational throughput characterizes the dynamic behavior of the ADC-based DTM scheme.

Given the significant architectural differences among the neural networks, the threshold temperatures for activating thermal management are set to 85°C, 55°C, and 55°C for ResNet-18, ResNet-16, and the 4-layer CNN, respectively. The corresponding thermal management periods are configured as 9:00–18:00, 10:00–17:00, and 12:00–15:00. The number of tasks completed under each scheme is then evaluated and compared. During the specified time intervals, the absence of any thermal management scheme causes the chip hotspot temperature to exceed the threshold. This results in a forced shutdown and prevents completion of the computational tasks. Therefore, only the number of tasks completed under different DTM schemes is compared. This comparison highlights the trade-off between computational throughput and thermal regulation achieved by the proposed DTM schemes. As shown in Figure 5(b), for the three neural networks—ResNet-18, ResNet-16, and 4-Layer CNN—the proposed DTM for CIM chips achieves higher throughput than the ADC-based DTM, with improvements of 59%, 89%, and 85%, respectively. Moreover, when the proposed DTM is combined with the LTCO scheme, the throughput is further improved by 15%, 29%, and 13%, respectively, compared to using the DTM for CIM chips alone.

6 Conclusion

In this work, we propose a comprehensive framework to tackle thermal challenges in RRAM-based CIM chips. The framework addresses key issues, including hotspot formation, thermal-induced performance degradation, and data-space coupling. We introduce targeted optimization strategies across three stages: design, pre-operation, and operation. At the design stage, optimized architectural strategies are employed to minimize accuracy loss caused by thermal effects. For the pre-operation phase, we present a latency-thermal co-optimization approach, combining static thermal management with a genetic algorithm to refine neural network mapping. This reduces hotspot temperatures by 6.5°C and lowers temperature standard deviation by 5.3°C without latency overhead. During operation, we develop a dynamic thermal management technique. When thermal management is triggered, the combination of LTCO and DTM achieves more than a 10% performance improvement over DTM alone.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 62495103, 92464302, 92264201).

References

1 Feng S, Yan X, Sun H, et al. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. Nat Commun, 2021, 12: 748
2 Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. ACM Trans Intell Syst Technol, 2024, 15: 1–45
3 Sandhu A K. Big data with cloud computing: discussions and challenges. Big Data Min Anal, 2021, 5: 32–40
4 Strubell E, Ganesh A, McCallum A, et al. Energy and policy considerations for modern deep learning research. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020
5 Gebregiorgis A, Du Nguyen H A, Yu J, et al. A survey on memory-centric computer architectures. J Emerg Technol Comput Syst, 2022, 18: 1–50
6 Guo X, Ipek E, Soyata T. Resistive computation. SIGARCH Comput Archit News, 2010, 38: 371–382
7 Nguyen H A D, Yu J, Lebdeh M A, et al. A classification of memory-centric computing. J Emerg Technol Comput Syst, 2020, 16: 1–26
8 Haensch W, Raghunathan A, Roy K, et al. Compute in-memory with non-volatile elements for neural networks: a review from a co-design perspective. Adv Mater, 2023, 35: 2204944
9 Horowitz M. Computing's energy problem (and what we can do about it). In: Proceedings of IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014. 10–14
10 Wong H S P, Salahuddin S. Memory leads the way to better computing. Nat Nanotech, 2015, 10: 191–194
11 Choquette J, Gandhi W. NVIDIA A100 GPU: performance & innovation for GPU computing. In: Proceedings of 2020 IEEE Hot Chips 32 Symposium (HCS), 2020. 1–43
12 Choquette J. Nvidia Hopper GPU: scaling performance. In: Proceedings of IEEE Hot Chips 34 Symposium (HCS), 2022. 1–46
13 Prezioso M, Merrikh-Bayat F, Hoskins B D, et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. Nature, 2015, 521: 61–64
14 Ambrogio S, Narayanan P, Tsai H, et al. Equivalent-accuracy accelerated neural-network training using analogue memory. Nature, 2018, 558: 60–67
15 Ielmini D, Wong H S P. In-memory computing with resistive switching devices. Nat Electron, 2018, 1: 333–343
16 Huo Q, Yang Y, Wang Y, et al. A computing-in-memory macro based on three-dimensional resistive random-access memory. Nat Electron, 2022, 5: 469–477
17 Zidan M A, Strachan J P, Lu W D. The future of electronics based on memristive systems. Nat Electron, 2018, 1: 22–29
18 Lanza M, Sebastian A, Lu W D, et al. Memristive technologies for data storage, computation, encryption, and radio-frequency communication. Science, 2022, 376: eabj9979
19 Mochida R, Kouno K, Hayata Y, et al. A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture. In: Proceedings of IEEE Symposium on VLSI Technology, 2018. 175–176
20 Khaddam-Aljameh R, Stanisavljevic M, Mas J F, et al. HERMES core—a 14 nm CMOS and PCM-based in-memory compute core using an array of 300 ps/LSB linearized CCO-based ADCs and local digital processing. In: Proceedings of Symposium on VLSI Circuits, 2021. 1–2
21 Xue C X, Chen W H, Liu J S, et al. 24.1 A 1Mb multibit ReRAM computing-in-memory macro with 14.6 ns parallel MAC computing time for CNN based AI edge processors. In: Proceedings of IEEE International Solid-State Circuits Conference-(ISSCC), 2019. 388–390
22 Ishii M, Kim S, Lewis S, et al. On-chip trainable 1.4 M 6T2R PCM synaptic array with 1.6 K stochastic LIF neurons for spiking RBM. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2019
23 Yan S Z, Cong Z R, Wang Z, et al. A monolithic 3D IGZO-RRAM-SRAM-integrated architecture for robust and efficient compute-in-memory enabling equivalent-ideal device metrics. Sci China Inf Sci, 2025, 68: 122404
24 Wan W, Kubendran R, Eryilmaz S B, et al. A 74 TMACS/W CMOS-RRAM neurosynaptic core with dynamically reconfigurable dataflow and in-situ transposable weights for probabilistic graphical models. In: Proceedings of IEEE International Solid-State Circuits Conference-(ISSCC), 2020. 498–500
25 Narayanan P, Ambrogio S, Okazaki A, et al. Fully on-chip MAC at 14 nm enabled by accurate row-wise programming of PCM-based weights and parallel vector-transport in duration-format. IEEE Trans Electron Devices, 2021, 68: 6629–6636
26 Wei S T, Gao B, Wu D, et al. Trends and challenges in the circuit and macro of RRAM-based computing-in-memory systems. Chip, 2022, 1: 100004
27 Peterson G P, Ortega A. Thermal control of electronic equipment and devices. In: Advances in Heat Transfer. Amsterdam: Elsevier, 1990. 181–314
28 Wan W, Kubendran R, Schaefer C, et al. A compute-in-memory chip based on resistive random-access memory. Nature, 2022, 608: 504–512
29 Liu Q, Gao B, Yao P, et al. A fully integrated analog ReRAM based 78.4 TOPS/W compute-in-memory chip with fully parallel MAC computing. In: Proceedings of IEEE International Solid-State Circuits Conference-(ISSCC), 2020. 500–502
30 Kaul A, Luo Y, Peng X, et al. Thermal reliability considerations of resistive synaptic devices for 3D CIM system performance. In: Proceedings of IEEE International 3D Systems Integration Conference (3DIC), 2021
31 Wang Z, Wu Y, Park Y, et al. Safe, secure and trustworthy compute-in-memory accelerators. Nat Electron, 2024, 7: 1086–1097
32 Rayapati V, Rao N, Suri M. VPU-CIM: a 130 nm, 33.98 TOPS/W RRAM based compute-in-memory vector co-processor. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2024. 1–5
33 Wei S, Yao P, Guo X, et al. A 28 nm static-power-free fully-parallel RRAM-based TD CIM macro with 1982 TOPS/W/bit for edge applications solid-state circuits letters. IEEE Solid-State Circuits Letters, 2024, 8: 21–24
34 Yang W, Zhou S, Xu H, et al. An integration and time-sampling based readout circuit with current compensation for parallel MAC operations in RRAM Arrays. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2024. 1–5

35 Ma A, Gao B, Lu Y, et al. Multi-scale thermal modeling of 3D-heterogeneous integrated processing-near-memory chip for edge large language model inference. In: Proceedings of the 9th IEEE Electron Devices Technology and Manufacturing Conference (EDTM), 2025. 2–5

36 Ma A, Gao B, Liu Y, et al. Multi-scale thermal modeling of RRAM-based 3D monolithic-integrated computing-in-memory chips. In: Proceedings of International Electron Devices Meeting (IEDM), 2022. 15–5

37 Ma A, Gao B, Mou X, et al. Thermal induced retention degradation of RRAM-based neuromorphic computing chips. In: Proceedings of IEEE International Reliability Physics Symposium (IRPS), 2023. 1–6

38 Brooks D, Martonosi M. Dynamic thermal management for high-performance microprocessors. In: Proceedings of HPCA Seventh International Symposium on High-Performance Computer Architecture, 2001. 171–182

39 Coskun A K, Rosing T S, Gross K C. Utilizing predictors for efficient thermal management in multiprocessor SoCs. IEEE Trans Comput-Aided Des Integr Circuits Syst, 2009, 28: 1503–1516

40 Beigi M V, Memik G. Thor: thermal-aware optimizations for extending reram lifetime. In: Proceedings of IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2018

41 Zhao M, Wu H, Gao B, et al. Investigation of statistical retention of filamentary analog RRAM for neuromophic computing. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2017. 39–4

42 Shim W, Meng J, Peng X, et al. Impact of multilevel retention characteristics on RRAM based DNN inference engine. In: Proceedings of IEEE International Reliability Physics Symposium (IRPS), 2021. 1–4

43 Kaul A, Peng X, Rajan S K, et al. Thermal modeling of 3D polylithic integration and implications on BEOL RRAM performance. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2020

44 Ma A, Gao B, Yao P, et al. Thermal analysis and evaluation of memristor-based compute-in-memory chips. Chips, 2025, 4: 1

45 Zhou M, Imani M, Gupta S, et al. Thermal-aware design and management for search-based in-memory acceleration. In: Proceedings of the 56th Annual Design Automation Conference 2019. 2019

46 Mead C. Neuromorphic electronic systems. Proc IEEE, 1990, 78: 1629–1636

47 Chen W H, Dou C, Li K X, et al. CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors. Nat Electron, 2019, 2: 420–428

48 Hung J M, Xue C X, Kao H Y, et al. A four-megabit compute-in-memory macro with eight-bit precision based on CMOS and resistive random-access memory for AI edge devices. Nat Electron, 2021, 4: 921–930

49 Cai F, Correll J M, Lee S H, et al. A fully integrated reprogrammable memristor-CMOS system for efficient multiply-accumulate operations. Nat Electron, 2019, 2: 290–299

50 Xue C X, Chiu Y C, Liu T W, et al. A CMOS-integrated compute-in-memory macro based on resistive random-access memory for AI edge devices. Nat Electron, 2021, 4: 81–90

51 Burr G W, Shelby R M, Sidler S, et al. Experimental demonstration and tolerancing of a large-scale neural network (165000 Synapses) using phase-change memory as the synaptic weight element. IEEE Trans Electron Devices, 2015, 62: 3498–3507

52 Joshi V, Le Gallo M, Haefeli S, et al. Accurate deep neural network inference using computational phase-change memory. Nat Commun, 2020, 11: 2473

53 Zhang W, Yao P, Gao B, et al. Edge learning using a fully integrated neuro-inspired memristor chip. Science, 2023, 381: 1205–1211

54 Yao P, Wu H, Gao B, et al. Fully hardware-implemented memristor convolutional neural network. Nature, 2020, 577: 641–646

55 Zhang W Q, Gao B, Yao P, et al. Array-level boosting method with spatial extended allocation to improve the accuracy of memristor based computing-in-memory chips. Sci China Inf Sci, 2021, 64: 160406

56 Zhao H, Liu Z, Tang J, et al. Energy-efficient high-fidelity image reconstruction with memristor arrays for medical diagnosis. Nat Commun, 2023, 14: 2276

57 Zhao H, Liu Z, Tang J, et al. Implementation of discrete Fourier transform using RRAM arrays with quasi-analog mapping for high-fidelity medical image reconstruction. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2021

58 Liu Z, Tang J, Gao B, et al. Neural signal analysis with memristor arrays towards high-efficiency brain-machine interfaces. Nat Commun, 2020, 11: 4234

59 Li X, Gao B, Lin B, et al. First demonstration of homomorphic encryption using multi-functional RRAM arrays with a novel noise-modulation scheme. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), 2022

60 Liang X, Tang J, Zhong Y, et al. Physical reservoir computing with emerging electronics. Nat Electron, 2024, 7: 193–206

61 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778