Medium-Temperature-Oxidized GeO Resistive-Switching Random-Access Memory and Its Applicability in Processing-in-Memory Computing

Kannan Udaya Mohanan
Gachon University
Seongjae Cho (felixcho@gachon.ac.kr)
Gachon University
Byung-Gook Park
Seoul National University

Research Article

Keywords: medium-temperature oxidation, germanium oxide, resistive-switching random-access memory (ReRAM), low-power hardware neural network, processing-in-memory (PIM)

Posted Date: April 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1548156/v1

Abstract

Processing-in-memory (PIM) is emerging as a new computing paradigm to replace the existing von Neumann computer architecture for data-intensive processing. For the higher end-user mobility, low-power operation capability is more increasingly required and components need to be renovated to make a way out of the conventional software-driven artificial intelligence. In this work, we investigate the hardware performances of PIM architecture that can be presumably constructed by resistive-switching random-access memory (ReRAM) synapse fabricated with a relatively larger thermal budget in the full Si processing compatibility. By introducing a medium-temperature oxidation (MTO) in which the sputtered Ge atoms are oxidized at a relatively higher temperature compared with the ReRAM devices fabricated by physical vapor deposition (PVD) at room temperature, higher device reliability has been acquired. Based on the empirically obtained device parameters, a PIM architecture has been conceived and a system-level evaluations have been performed in this work. Considerations include the cycle-to-cycle variation in the GeO_x ReRAM synapse, analog-to-digital converter (ADC) resolution, synaptic array size, and interconnect latency for the system-level evaluation with the Canadian Institute for Advance Research (CIFAR)-10 dataset. A fully Si processing compatible and robust ReRAM synapse and its applicability for PIM are demonstrated.

Introduction

Over the past several decades, the physical downscaling in process technology is approaching the limits of fundamental physics. On the other hand, the demands on higher device scalability and operation speed, and low-power consumption capability have been incessantly increased, which gets more accelerated by necessity of data-intensive processing represented by big data analytics and deep learning for making accurate decisions in recent times. Conventional von Neumann architecture suffers from the memory bottleneck in this data-intensive applications due to the physically separated central processing unit and memory domain, along with the serial communication method between them. This inevitable serial data shuttling between the processing and memory domains leads to huge amount of latency and energy expenditure, which gets worse as the data size is required to be larger. Processing-in-memory (PIM) computing architecture has been researched for a long time in the very-large-scale integration (VLSI) technology regime for higher parallelism in data processing by introducing the processing capability into the memory domain [1–5]. However, most of the technological contributions have been made for the near-memory processing (NMP) in ways that the physical distance between processing and memory domains is reduced. The rather metaphorically used expression of PIM can be more substantially literal when supported by the device-level innovations. The PIM architecture design assures highly parallel computing capabilities which stem from the localized multiplication-and-accumulation (MAC) operations preferably using nonvolatile memories woven for the crossbar array towards higher area and energy efficiencies. Resistive-switching random-access memory (ReRAM) is considered as one of the most promising candidates for the synaptic components in the PIM architecture due to its simple device structure, high scalability, and fast switching speed [6–16]. Although researches on ReRAM

devices have been focused on various aspects including device structure, electrode materials, and process integration based on wide variety of switching materials such as TiO_2, NiO_2, and TaO_x, for higher device performances and reliability [17–19], there is still room for further improving the robustness of switching materials in terms of parameter distributions. For the qualification of ReRAM for the application as PIM component, higher device reliability should be warranted to endure the highly frequent learning and inference operations.

In this work, electrically more robust and reliable ReRAM based on GeO_x as the switching material has been fabricated, characterized, and the system-level evaluations are carried out for the PIM architecture with the embedment of GeO_x ReRAM cells as the synaptic components. The switching layer of GeO_x was prepared by a medium-temperature oxidation (MTO) with a relatively higher thermal budget, in the opposite direction in which the ReRAM cells are usually fabricated by physical vapor deposition (PVD) at room temperature or low temperatures not prominently higher than that. Based on the device operation parameters extracted from the measurement results, system-level evaluations of the PIM based on GeO_x ReRAM are performed with image recognition tests by series of simulations accommodating the realistic hardware circuitry. Detailed hardware performance parameters are presumed from a system-level simulation package for 32-nm technology node [20]. Last but not the least, the effects of nonideal ReRAM operation characteristic of variation in cycle-to-cycle operations on the hardware neural network performances are closely investigated.

Results And Discussion

Device Fabrication and Characterization

There have been various candidates for the material combination to make up the metal-insulator-semiconductor (MIS) stack for ReRAM device. In this work, Ni/GeO_x/p^+ Si MIS stack was fabricated. There are two reasons for having employed the material combination; one is to equip the fabrication viability through introducing the materials with compatibility to conventional Si processing which is mostly adopted for the modern VLSI electronics and the other is to obtain more concentrated distribution of operation voltages with non-metallic switching material. Figure 1(a) shows the cross-sectional view of a fabricated ReRAM device by a high-resolution transmission microscopy (HR-TEM), by which GeO_x switching layer with 3-nm thickness is confirmed. The schematic of the fabricated ReRAM cells is shown in Fig. 1(b). Ni and p^+ Si act as the materials for the top electrode (TE) and bottom one (BE), respectively. Figure 2(a) shows the measured *I-V* curves from the fabricated ReRAM device with a diameter of 100 μm after 1, 5, 10, and 20 direct-current (DC) sweeps using a Keithley 4200A, with 0.1-mA compliance current. The distribution of set and reset voltages are confirmed to be narrow owing to the non-metallic switching dielectric material formed by a MTO and finalized by a post-deposition annealing (PDA). In order to elucidate the conduction mechanism in the Ni/GeO*x*/p^+-Si ReRAM cell, *I-V* curves in the high-resistance state (HRS) in the positive voltage HRS region and in the negative voltage LRS region are depicted in Figs. 2(b) and 2(c). As shown in Fig. 2(b) for the case of HRS, there are two linear regions with different slopes:

region I for TE voltage < 0.7 V with a slope of 1.83 and region III for TE voltage > 2.5 V with a slope of 4.78, respectively. Also, there is a nonlinear region between these two regions: region II for TE voltage is located in 0.7 V ≤ V < 2.5 V and the average slope is approximated to 2.49. Judging from the slopes in Fig. 2(b), Child's square law ($I \sim V^2$) has the dominance in the low and intermediate TE voltage regions [21–22]. Also, the behavior in the high TE voltage region demonstrating the large slope attributes to drastic increase in current by trap-controlled space-charge-limited current (SCLC) [23]. In case of LRS state in the negative TE voltage region, the slope is extracted to be 1.06 as shown in Fig. 2(c), in which the current conduction mechanism can be mainly explained by ohmic conduction. Figures 3(a) through 3(d) illustrates the construction and destruction of the conducting bridge in the Ni/GeOx/p^+-Si ReRAM device. There is no conduction filament in the pristine state (Fig. 3(a)) but Ni^{2+} ions begin to penetrate into the GeO_x switching layer as the TE voltage increases. These Ni^{2+} ions are reduced at the BE resulting in the gradual growth of conductive filaments of Ni atoms towards the TE (Fig. 3(b)). As the TE voltage increases, the filament formed by the Ni atoms touches the TE and the resistance state turns to LRS (Fig. 3(c)). As the TE voltage is reduced and goes into the negative region, the conductive filament undergoes electrochemical dissolution and gets ruptured leading to the HRS state (Fig. 3(d)). This formation and rupture of the conducting filament, or conducting bridge, are realized by the metallic species, which is more likely to be observed in the ReRAM cells employing Ni as the TE material [24, 25]. The conducting filaments repeating the construction and destruction with voltage dependence shown in Figs. 3(a) through 3(d) are randomly distributed over the ReRAM cell as illustrated in Fig. 4(a). Each filament can be described as a parallel combination of a voltage-dependence resistance and a capacitance as shown in Fig. 4(b). The series resistance (R_s) at the top of the block comes from the series combination of TE, BE, and contact resistances. Since all the filaments are connected in parallel between TE and BE, all the resistances can be lumped into an equivalent cell resistance (R_c), and likewise, all the parallel capacitances are summed into an equivalent cell capacitance (C) as demonstrated in Fig. 4(c). Although the construction and destruction of the conducting bridge are explained by the movements of the metallic atoms and the bridging mechanism can be varied according to the material combination making up the cell stack, an individual cell can be described by a variable resistor and a capacitor, and thus, the suggested equivalent electrical circuit model in Figs. 4(b) and 4(c) is allowed to have the high universality for ReRAM devices. In order to extract the passive elements in the ReRAM cell, the fabricated devices were brought to impedance analysis using an impedance analyzer by introducing the equivalent circuit model in Fig. 4(c). The Cole-Cole plots from the fabricated ReRAM device in the HRS and LRS are shown in Figs. 5(a) and 5(b), respectively. The measurement frequency was varied from 1 Hz to 200 kHz, and the x (Z') and y (Z'') axes indicate the real and imaginary parts of the impedance. As the frequency goes higher, the trajectory is plotted in the counterclockwise direction. The appearance of a single semicircle in the Cole-Cole plot is an affirmation of the fact that the charge transport mechanism in the device can be described in terms of a parallel RC circuit as described in Fig. 4(c). The square symbols in Figs. 5(a) and 5(b) show the measurement results whereas the continuous lines denote the fitted data. Table 1 shows the values of the extracted parameters from the impedance analysis of the GeO_x ReRAM device. It is revealed that the capacitance in the LRS is much smaller than that in the HRS, which attributes to the reduction in effective

area for the device capacitance taking place over the growth of a conductive filament. The higher accuracy in the impedance analysis fitting shows that the physical simplification of a realistic ReRAM cell in Fig. 4(a) and the equivalent circuits in Figs. 4(b) and 4(c) induced from the results in Fig. 4(a) have high coherence. In order to further explain the charge transport within the device, energy-band diagrams across the Ni/GeOx/p^+ Si stack in the HRS are schematically shown in Fig. 6(a). Since the workfunctions of Ni BE and p^+ Si are not significantly different, the energy-band diagram is drawn to be under nearly flat-band condition. In the HRS, there is no conduction filament bridging TE and BE, inside the switching layer dielectric as shown in Fig. 6(a), with the know material parameters. As the TE voltage gets higher, the energy bands in the GeO_x layer and p^+ Si are bent upward as demonstrated in Fig. 6(b), and further increase in the TE voltage leads to the field-enhanced migration of Ni cations from the TE into the switching layer as schematically shown in Fig. 6(c). The Ni^{2+} ions are reduced to metallic Ni atoms at the BE. A conductive filament is formed of Ni atoms which grows toward the TE. At the moment when the TE and BE are bridged by a conducting filament composed of Ni atoms, memory state transition takes places from HRS to LRS finally. Figure 6(d) presents the energy-band diagram of the Ni/GeOx/p^+Si ReRAM cell with the Ni conducting bridge in the LRS state when there is no applied TE voltage, in the steady state.

Table 1
Values of passive elements extracted from the impedance analyses.

Resistance states	Rs [Ω]	Rc [MΩ]	C [pF]	Extraction voltage [V]
HRS	197	12.1	163	2.3
LRS	216	29.2	116	0.7

Training Approach and Hardware Architecture of the PIM with GeOx ReRAM

The off-chip training capability for graphical image recognition by the GeO_x ReRAM has been evaluated using the Canadian Institute for Advanced Research (CIFAR)-10 dataset in the Visual Geometry Group (VGG)-8 neural network architecture. The architecture of the VGG-8 network comprises a total 8 layers: 6 convolutional layers and 2 fully-connected layers. The detailed schematic of the VGG-8 network architecture is shown in Fig. 7(a). The input CIFAR-10 dataset has a collection of 60,000 color (red-green-blue) images of 32 × 32 resolution. The images can be broadly classified into 10 output indexes. During the network training, the data is grouped into 50,000 train and 10,000 test images with a batch size of 200. The VGG-8 network has been trained using a stochastic gradient descent (SGD) algorithm and rectified linear unit (ReLU) activation function. The realization of hardware-sense neural network for a PIM architecture is illustrated in Fig. 7(b) [20]. The hardware design is capable of evaluating the performance of the VGG-8 network of GeO_x ReRAM synaptic devices. The system takes into account the various hardware constraints including technology node, analog-to-digital converter (ADC) precision and

the nonideal changes in the synaptic weights during the training. The system design has been hierarchically organized into chip level, processing element level, and synaptic array level elements. For the full single-chip hardware integration, peripheral circuits including ADCs, buffers, multiplexers (MUX), interconnects with the 32-nm predictive technology SPICE model parameters have been presumably used and other relevant circuitry such as digital adders and shift registers have been also considered. The accumulation circuits include the chip-level units, processing element level adders, tile-level adders, and shift adders on the edges of the ReRAM synapse array. The system level performance has been evaluated using an analog parallel read-out scheme using 64 × 64 synaptic array size and 5-bit ADC precision. The input data flows into the wordline (WL) switch matrix and the MAC operations in the crossbar array generate partial sums which are accumulated along the columns using the read-out circuits (flash ADCs). The bit-quantized ADCs are much larger in area than the synaptic array column pitch and hence they share several columns using the column MUX. The roles of adders and shift registers are to shift and accumulate partial sums by the MAC operations over repeated data cycles due to batch-wise data processing. A major concern with the batch-wise data processing lies in the large amount of intermediate data generated during the feed-forward process taking place in the computation of activations. In order to minimize the requirement for in-chip memory space, the PIM architecture can be designed to send the intermediate data to off-chip DRAM, which can be optional depending on neural network size and chip area of the GeO_x ReRAM PIM architecture.

System-Level Performance Evaluation

For evaluating the system-level performances of the PIM architecture based on GeO_x ReRAM, binary-state switching operations in the synaptic array were assumed with potentiation (write) voltage = 3 V with a pulse width = 100 μs and inference voltage = 0.7 V with a pulse width = 100 μs. Figure 8 shows the accuracy in CIFAR-10 image recognition as a function of number of epochs in comparison between software and hardware neural networks. It is observed that the PIM system with the hardware neural network of GeO_x ReRAM synapses has achieved an accuracy of 91.27%, which is comparably high with the accuracy obtained by the software neural network, 92.31%, in terms of test accuracy. A sharp jump is witnessed in both the software and hardware-based trainings at 200 epochs. This is due to the decreasing learning rate strategy employed after 200 epochs for the optimized training of the network. The inset in the Fig. 7(c) shows a subset of the CIFAR-10 dataset. The system-level parameters from the simulation of the designed PIM architecture are summarized in Table 2. Figures 9(a) through 9(c) shows the pie diagrams of portions in energy, latency, and area occupied by different hardware components in the PIM architecture. The energy distribution in Fig. 9(a) reveals that the ADCs (multi-level current sense amplifier) and interconnects consume the largest energy, followed by the accumulation circuits. The synaptic array energy consumption is extremely low as compared to other components. The latency distribution in Fig. 9(b) shows that the logic and buffer circuits along with the interconnects have the predominance in determining the overall system latency. The large size of the VGG-8 network results in the considerable amount of on-chip data transfer from the buffer memory and a large number of

synapses in the array leading to increase complexity in interconnects within the PIM chip. This is a crucial factor in limiting the overall chip latency. Finally, it is observed from Fig. 9(c) that the total chip area is largely occupied by the ADCs. Thus, the ADC area needs to be intensively optimized with regard to both energy and area efficiencies. The energy, latency, and area minimally consumed by the GeO_x ReRAM synapse array is an indication of its high applicability in the hardware PIM architecture. Further, the computational demands of the VGG-8 network on the hardware PIM design is also evaluated in order for understanding the future directions for the optimization of hardware neural network architecture. Figures 10(a) and 10(b) describes the layer-wise energy consumption and latency distribution of the VGG-8 network across the main hardware components. It is observed that the convolutional layers with additional pooling layers (previously shown in Fig. 7(a)), i.e., layers 2 and 4, demand the largest energy and time consumptions for the in-memory computations. Judging from both the histograms, it is clarified that the inference energy and latency of the synaptic array across all the layers of VGG-8 are minimized by the virtues of the fabricated GeO_x ReRAM. However, it is important to consider the variation in the device-level operations in evaluating the system-level performances. Recently, there have been several studies on the effects of nonideal variations in the synaptic devices on the PIM system performances [26, 27]. As one of the most decisive nonidealities, variation in cycle-to-cycle switching operations can be quantified as a standard deviation and can be treated as an independent variable in determining the system accuracy. Figure 10(c) shows the maximum test accuracy for the CIFAR-10 image recognition as a function of the cycle-to-cycle variation. It is explicitly shown that there is little drop in the accuracy up to the standard deviation of 0.02, which confirms the robustness of the GeO_x ReRAM synaptic devices implementing the PIM architecture. It is demonstrated in Fig. 10(d) that the inference energy monotonically increases with the standard deviation but the system preserves the robustness against the device-level variation up to standard deviation of 0.02.

Table 2
Chip-level parameters and performances computed per epoch for the GeO_x ReRAM synapse array-based PIM architecture.

PIM chip parameters	Values
Chip area	62.5 mm^2
Total energy on chip	3.35×10^{-5} J
Latency	1.33 ms
Peak energy efficiency	58.92 TOPS/W
Mean energy efficiency	36.42 TOPS/W
Inference energy in the synapse array	1.64×10^{-6} J
Other logic energy	3.55×10^{-7} J
ADC energy	1.37×10^{-5} J
Interconnect energy	1.20×10^{-5} J
Inference latency in the synapse array	2.20×10^{-5} s
Other logic latency	5.58×10^{-4} s
ADC latency	5.86×10^{-5} s
Interconnect latency	5.55×10^{-4} s

Conclusion

ReRAM cells featuring the Ni/GeOx/p^+-Si stack with a high Si processing compatiblity have been been fabricated and characterized, with a particular interest in implementing highly-scalable nonvolatile memory-based PIM architecture. The fabricated ReRAM device has been effectively described by a more accurate equivalent circuit with measured state resistances and capacitances. The circuit and performance parameters of the fabricated GeO_x ReRAM were fed into the system-level simulation with realistic peripheral circuitry to evaluate the system accuracy in image learning and the applicability of the GeO_x ReRAM technology for the future computing architecture. The CIFAR-10 image recognition accuracy and hardware parameters have been evaluated in consideration of device-level nonideality. The energy consuption, latencey, and area occupied by the synaptic array is observed to be the smallest cin comparison with other functional modules in the PIM architecture. The computational demands of the pooling layer in the VGG-8 network on the overall chip energy consumption has been revealed by layer-wise neural network evaluation. In conclusion, a high image recognition accuracy above 90%, high energy

efficienty, low latency, and minimal area requirement warrant that the GeO_x ReRAM can be a plausible candidate for realizing the chip-packaged PIM architecture.

Methods

Device Fabrication

The proposed ReRAM devices was fabricated in the class-controlled nanofabrication facility. After preparing and initial cleaning of 6” *p*-type (100) Si wafers, ion implantation was performed with a dose of B 5×10^{15} cm^{-2} at an acceleration energy of 40 keV. Dopant activation was carried out in the furnace at 900 ºC for 20 min. and Ge of 3-nm thickness was deposited by a thermal evaporator at 96-A source current and 40×10^{-6} torr vacuum pressure. Then, dry oxidation of Ge was performed by a medium-temperature oxidation (MTO) at 550 ºC with an O_2 flow of 7,250 sccm for 10 min. and the wafers were sent to a thermal tube for an additional annealing at 600 ºC with an N_2 flow of 5,000 sccm for 20 min. Lithography was performed using a mask aligner for circular patterns. 200-nm-thick Ni was deposited on the GeO_x switching layer, and then, acetone and isopropyl alcohol (IPA) were put in the cyclic uses for lift-off and residual removal processes. Finally, the wafers were rinsed in the de-ionized (DI) water and completely dried for finishing the device fabrication.

Electrical Measurement

Electric switching characteristics of the fabricated GeO_x RRAM device were obtained at room temperature using a Keithley 4200A emiconductor parameter analyzer inside a probe station. The impedance analyses were carried out using a Hioki IM3590 impedance analyzer in the air ambient.

System-Level PIM Evaluation Environments

The system level simulations were carried in a high-end workstation employing a 32-core AMD Ryzen 9 Processor as the central processing unit (CPU) and an NVIDIA RTX 3090 as the graphic processing unit (GPU). For the neural network training, a stochastic gradient descent (SGD) algorithm was adopted with rectified linear unit (ReLU) activation function. During the network training, a batch size of 200 and a learning rate of 1 were used. The weight and the gradient were considered to have 5-bit precisions whereas the activation and the error were computed with 8-bit precision.

Declarations

Ethics Approval and Consent to Participate

Not applicable.

Consent for Publication

All authors consent to the publication of this manuscript.

Availability of Data and Materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT of Korea (MSIT) under the Grant Number of 2021M3F3A2A01037927.

Authors' Contributions

K.U.M. characterized the fabricated devices, wrote a part of the main manuscript text, prepared a part of artworks, and conducted the system-level simulation tasks. S.C. developed the base material and process integration, fabricated the devices, schemed the measurement approaches, and prepared a part of the main manuscript text and artworks. B.-G. P. attained the government project for this research, gave the overall research direction and insights into the relation among material, device operation, and system-level performances, and verified the data.

Acknowledgements

The researchers and engineers in the Inter-university Semiconductor Research Center (ISRC) at Seoul National University is acknowledged for help with highly responsible maintenance and calibration of the nanofabrication tools.

References

1. Ielmini, D.; Wong, H. S. P. In-memory computing with resistive switching devices. Nat. Electron. 2018, *1*, 333–343.
2. Sun, X.; Khwa, W. S.; Chen, Y. S.; Lee, C. H.; Lee, H. Y.; Yu, S. M.; Naous, R.; Wu, J. Y.; Chen, T. C.; Bao, X.; Chang, M. F.; Diaz, C. H.; Wong, H. S. P.; Akarvardar K. PCM-Based Analog Compute-In-Memory: Impact of Device Non-Idealities on Inference Accuracy. IEEE Trans. Electron Devices 2021, *68*, 5585–5591.
3. Sebastian, A.; Gallo, M. L.; Khaddam-Aljameh, R.; Eleftheriou, E. Memory devices and applications for in-memory computing. Nat. Nanotechnol. 2020, *15*, 529–544.
4. Ziegler, T.; Waser, R.; Wouters, D. J.; Menzel, S. In-Memory Binary Vector–Matrix Multiplication Based on Complementary Resistive Switches. Adv. Intell. Syst. 2020, *2*, 2000134.
5. Meng, J.; Wang, T.; He, Z.; Li, Q.; Zhu, H.; Ji, L.; Chen, L.; Sun, Q.; Zhang, D. W. A high-speed 2D optoelectronic in-memory computing device with 6-bit storage and pattern recognition capabilities.

Nano Res. 2022, *15*, 2472–2478.

6. Sung, S. H.; Kim, T. J.; Shin, H.; Namkung, H.; Im, T. H.; Wang, H. S.; Lee, K. J. Memory-centric neuromorphic computing for unstructured data processing. Nano Res. 2021, *14*, 3126–3142.
7. Choi, Y. J.; Kim, M. H.; Bang, S.; Kim, T. H.; Lee, D. K.; Hong, K.; Kim, C. S.; Kim, S.; Cho, S.; Park, B. G. Insertion of Ag Layer in TiN/SiN_x/TiN RRAM and Its Effect on Filament Formation Modeled by Monte Carlo Simulation. IEEE Access 2020, *8*, 228720–228730.
8. Bang, S.; Kim, M. H.; Kim, T. H.; Lee, D. K.; Kim, S.; Cho, S.; Park, B.G. Gradual switching and self-rectifying characteristics of Cu/α-IGZO/p+-Si RRAM for synaptic device application. Solid-State Electron. 2018, *150*, 60–65.
9. Kim, M. H.; Cho, S.; Park, B.G. Nanoscale wedge resistive-switching synaptic device and experimental verification of vector-matrix multiplication for hardware neuromorphic application. Jpn. J. Appl. Phys. 2021, *60*, 050905.
10. Kim, T. H.; Kim, M. H.; Bang, S.; Lee, D. K.; Kim, S.; Cho, S.; Park, B.G. Fabrication and characterization of TiO_x memristor for synaptic device application. IEEE Trans. Nanotechnol. 2020, *19*, 475–480.
11. Rasheed, U.; Ryu, H.; Mahata, C.; Khalil, R. M. A.; Imran, M.; Rana, A. M.; Kousar, F.; Kim, B.; Kim, Y.; Cho, S.; Hussain, F. Resistive switching characteristics and theoretical simulation of a Pt/a-Ta_2O_5/TiN synaptic device for neuromorphic applications. J. Alloys Compd. 2021, *877*, 160204.
12. Ryu, J. H.; Kim, B.; Hussain, F.; Ismail, M.; Mahata, C.; Oh, T.; Imran, M.; Min, K. K.; Kim, T. H.; Yang, B. D.; Cho, S. Zinc tin oxide synaptic device for neuromorphic engineering. IEEE Access 2020, *8*, 130678–130686.
13. Lee, D. K.; Kim, M. H.; Kim, T. H.; Bang, S.; Choi, Y. J.; Kim, S.; Cho, S.; Park, B. G. Synaptic behaviors of HfO_2 ReRAM by pulse frequency modulation. Solid-State Electron. 2019, *154*, 31–35.
14. Lee, J. Y.; Kim, Y.; Kim, M. H.; Go, S.; Ryu, S. W.; Lee, J. Y.; Ha, T. J.; Kim, S. G.; Cho, S.; Park, B. G. 2019. Ni/GeO_x/p^+ Si resistive-switching random-access memory with full Si processing compatibility and its characterization and modeling. *Vacuum* **2019**, *161*, 63–70.
15. Ansari, M. H. R.; Kannan, U. M.; Cho, S. Core-Shell Dual-Gate Nanowire Charge-Trap Memory for Synaptic Operations for Neuromorphic Applications. *Nanomaterials* 2021, *11*, 1773.
16. Ielmini, D. Brain-inspired computing with resistive switching memory (RRAM): Devices, synapses and neural networks. Microelectron. Eng. 2018, *190*, 44–53.
17. Gale, E. TiO_2-based memristors and ReRAM: materials, mechanisms and models (a review). Semicond. Sci. Technol. 2014, *29*, 104004.
18. Ishihara, T.; Ohkubo, I.; Tsubouchi, K.; Kumigashira, H.; Joshi, U. S.; Matsumoto, Y.; Koinuma, H.; Oshima, M. Electrode dependence and film resistivity effect in the electric-field-induced resistance-switching phenomena in epitaxial NiO films. Mater. Sci. Eng. B 2008, *148*, 40–42.
19. Sung, C.; Padovani, A.; Beltrando, B.; Lee, D.; Kwak, M.; Lim, S.; Larcher, L.; Marca, V. D.; Hwang, H. Investigation of I–V linearity in TaO_x-Based RRAM devices for neuromorphic applications. IEEE J. Electron Devices Soc. 2019, *7*, 404–408.

20. Peng, X.; Huang, S.; Jiang, H.; Lu, A.; Yu, S.; DNN + NeuroSim V2.0: An end-to-end benchmarking framework for compute-in-memory accelerators for on-chip training. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 2021, *40*, 2306–2319.
21. Liu, Q.; Guan, W.; Long, S.; Jia, R.; Liu, M.; Chen, J. Resistive switching memory effect of ZrO_2 films with Zr^+ implanted. Appl. Phys. Lett. 2008, *92*, 012117.
22. Prakash, A.; Jana, D.; Samanta, S.; Maikap, S. Self-compliance-improved resistive switching using Ir/TaO_x/W cross-point memory. Nanoscale Res. Lett. 2013, *8*, 1–6.
23. Parmenter, R. H.; Ruppel, W. Two-Carrier Space-Charge-Limited Current in a Trap-Free Insulator. J. Appl. Phys. 1959, *30*, 1548–1558.
24. Wu, X.; Cha, D.; Bosman, M.; Raghavan, N.; Migas, D. B.; Borisenko, V. E.; Zhang, X. X.; Li, K.; Pey, K. L.; Intrinsic nanofilamentation in resistive switching. J. Appl. Phys. 2013, *113*, 114503.
25. Sun, J.; Liu, Q.; Xie, H.; Wu, X.; Xu, F.; Xu, T.; Long, S.; Lv, H.; Li, Y.; Sun, L.; Liu, M. In situ observation of nickel as an oxidizable electrode material for the solid-electrolyte-based resistive random access memory. Appl. Phys. Lett. 2013, *102*, 053502.
26. Pedretti, G.; Ielmini, D. In-memory computing with resistive memory circuits: status and outlook. Electronics 2021, *10*, 1063.
27. Milo, V.; Glukhov, A.; Perez, E.; Zambelli, C.; Lepri, N.; Mahadevaiah, M. K.; Quesada, E. P. B.; Olivo, P.; Wenger, C.; Ielmini, D. Accurate program/verify schemes of resistive switching memory (rram) for in-memory neural network circuits. IEEE Trans. Electron Devices 2021, *68*, 3832–3837.

Figures

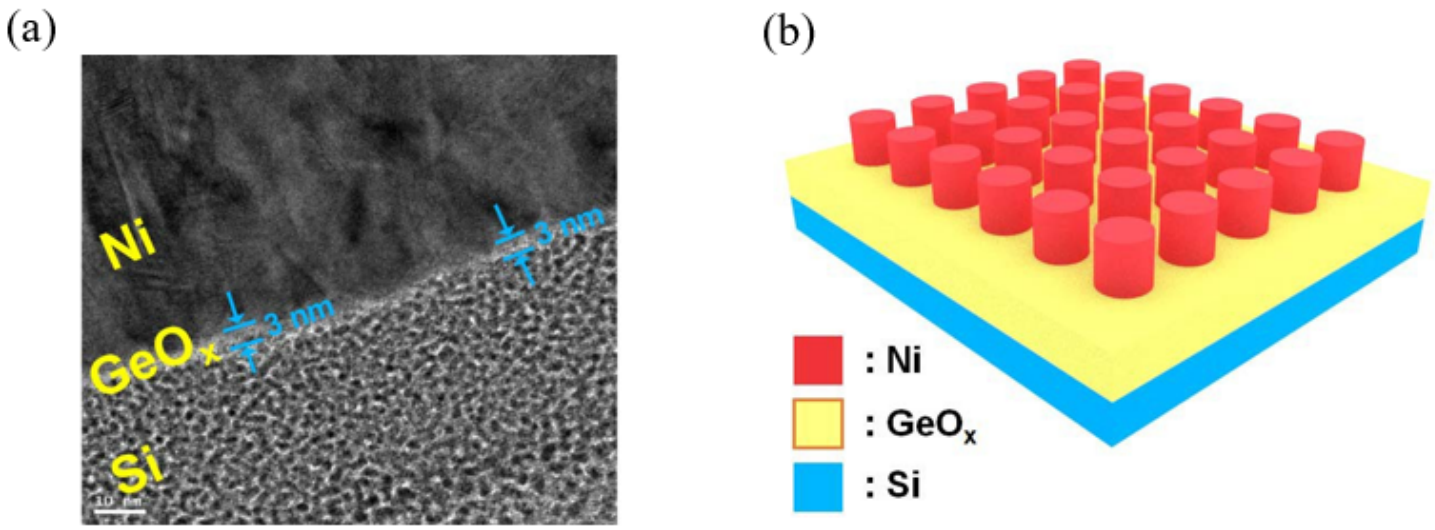

Figure 1

Fabricated GeO_x ReRAM cell. (a) Cross-sectional image by high-resolution transmission electron microscopy (HR- TEM). (b) Schematic of the ReRAM array in the Ni/GeOx/p^+ Si metal-insulator-semiconductor (MIS) stack.

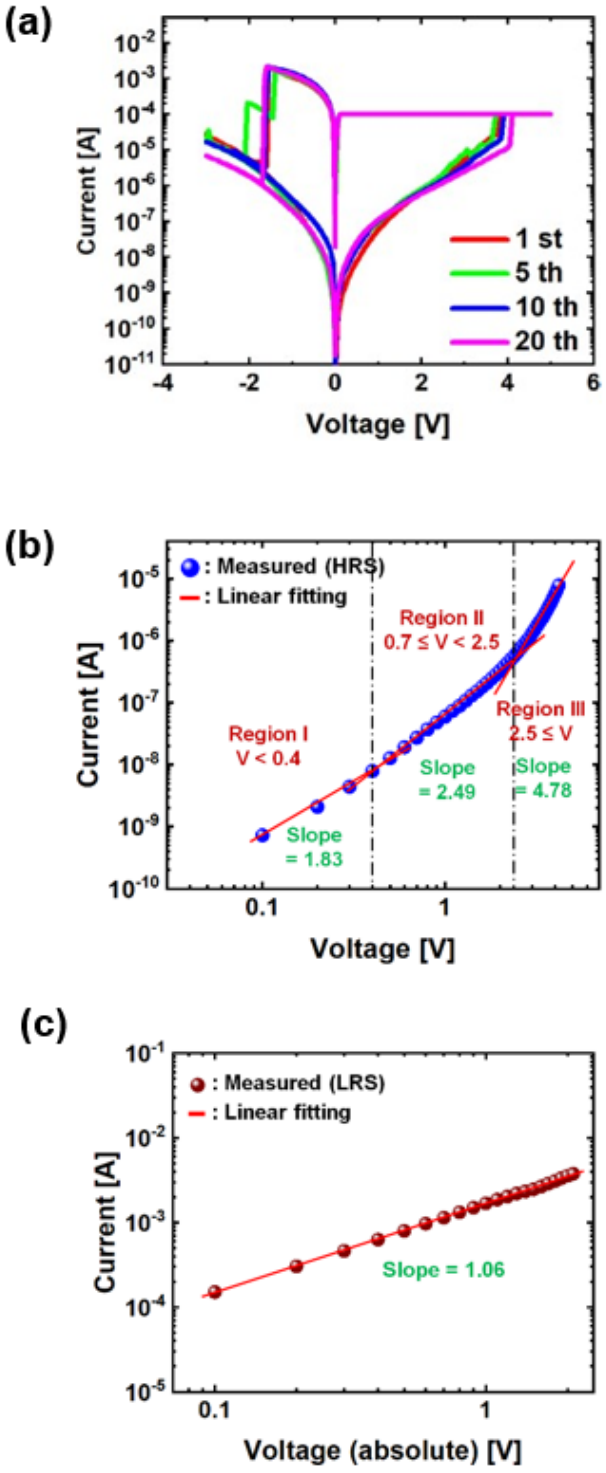

Figure 2

Measurement and fitting results from the fabricated GeO^x ReRAM cell. (a) *I-V* curves at 1, 5, 10, and 20 sweeps. *I-V* characteristics (b) in the positive-voltage HRS region and (c) in the negative-voltage LRS region.

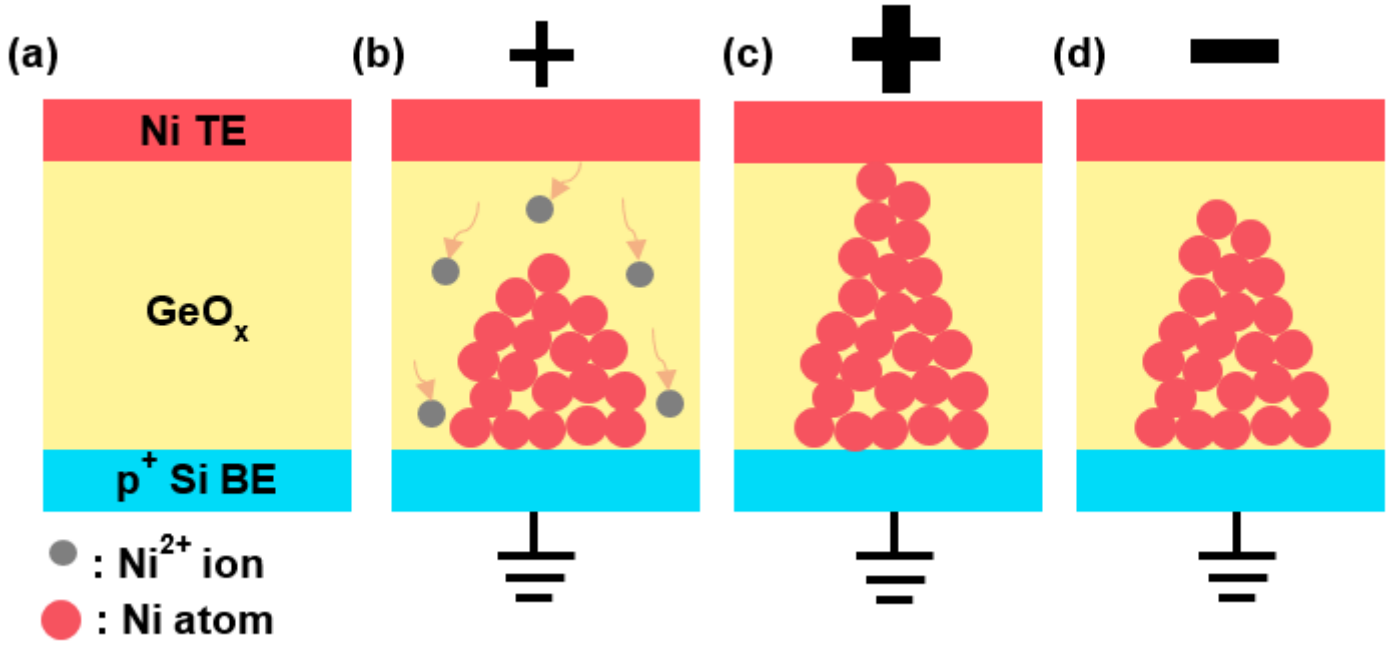

Figure 3

Schematic of the switching process of the Ni/GeO_x/p^+-Si ReRAM device (larger sign denotes larger bias). (a) Pristine state. (b) Ni^{2+} ions (grey circles) move towards the p^+ Si bottom electrode and undergoes reduction to form Ni atoms (red circles), by which a conductive filament is formed and grows toward the Ni top electrode. (c) Conductive filament touches the top electrode (LRS). (d) Rupture of the conductive filament due to the application of opposite polarity bias voltage (HRS).

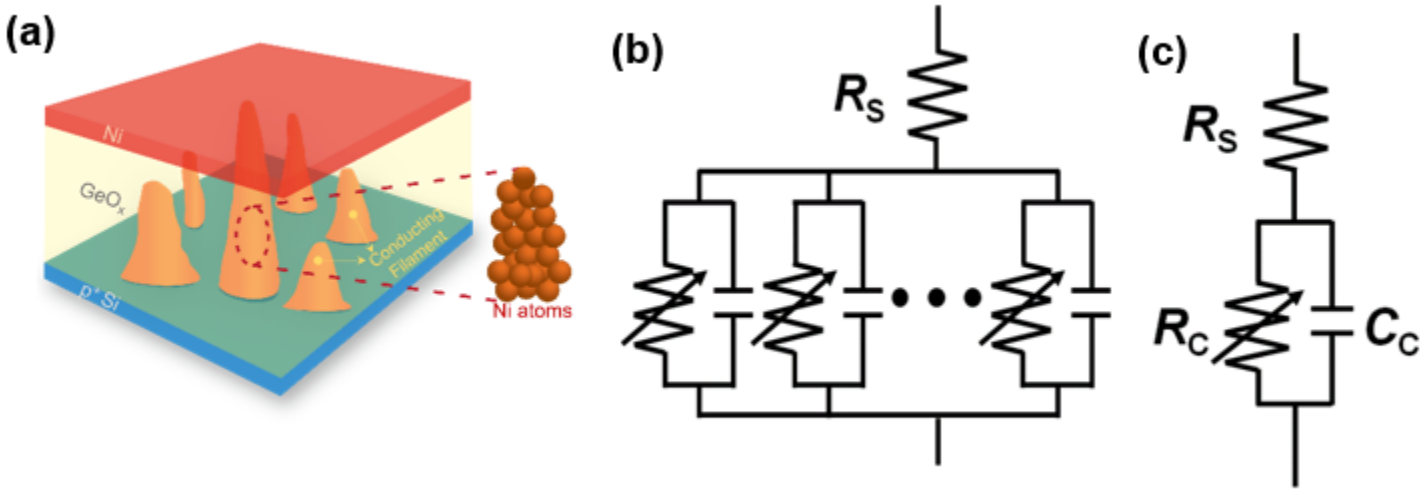

Figure 4

Physical and electrical representations of a GeO_x ReRAM cell. (a) Three-dimensional schematic of an ReRAM cell and the conducting filaments. (b) Electrical circuit model considering the multiple growths of conducting filaments. (c) Simplest ReRAM equivalent circuit.

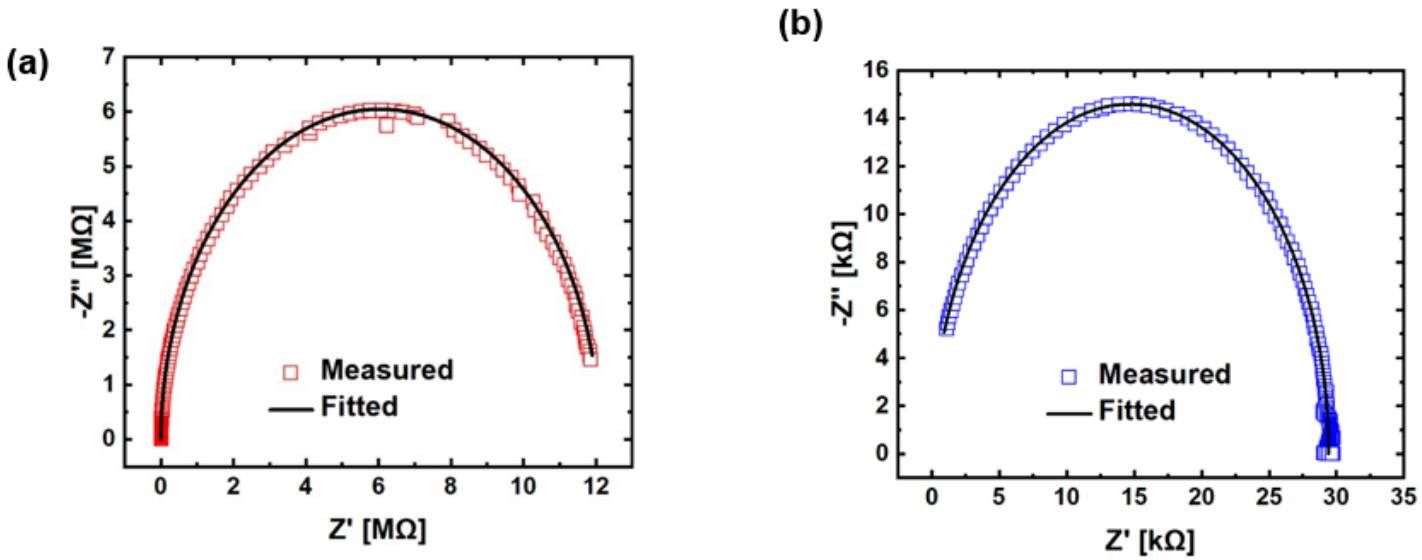

Figure 5

Plot of impedance of the fabricated GeO_x ReRAM in the complex plane (Cole-Cole plot). Impedances in the (a) HRS at TE voltage = 2.3 V and (b) LRS at TE voltage = 0.7 V.

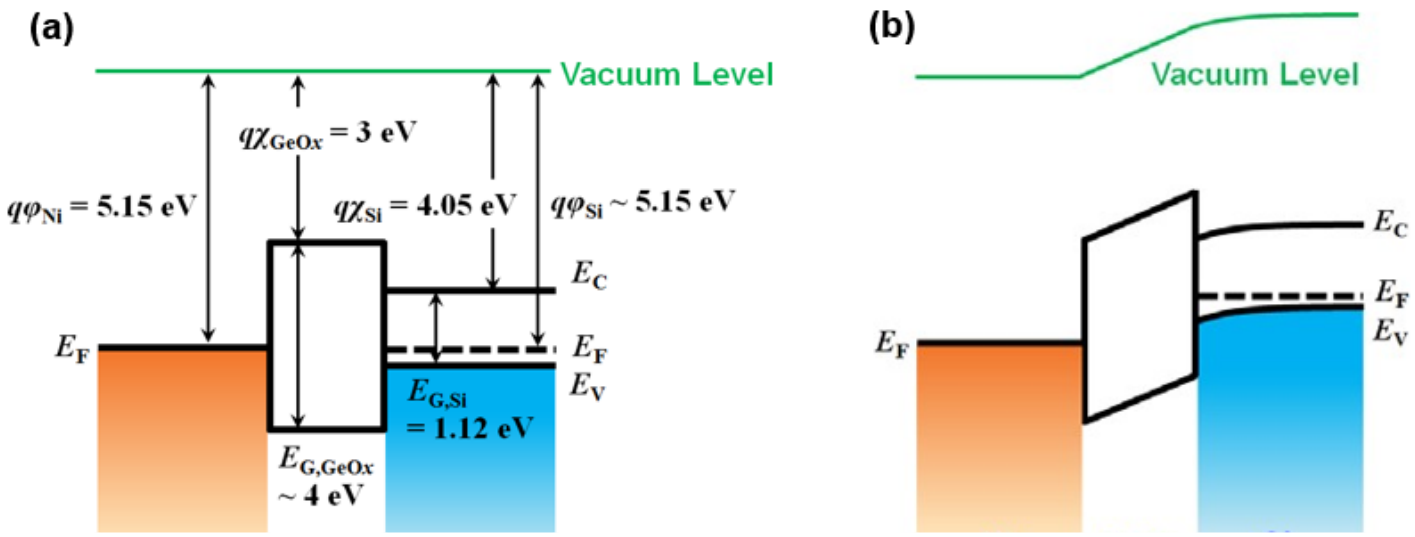

Figure 6

Energy-band diagrams of the Ni/GeO_x/p^+ Si ReRAM cell under different bias conditions. (a) Energy-band diagram in the HRS with the full material information. (b) Under positive bias in the HRS. (c) Migration of Ni ions into the GeO_x switching dielectric layer. (d) LRS has been established by the Ni conducting bridge between TE and BE.

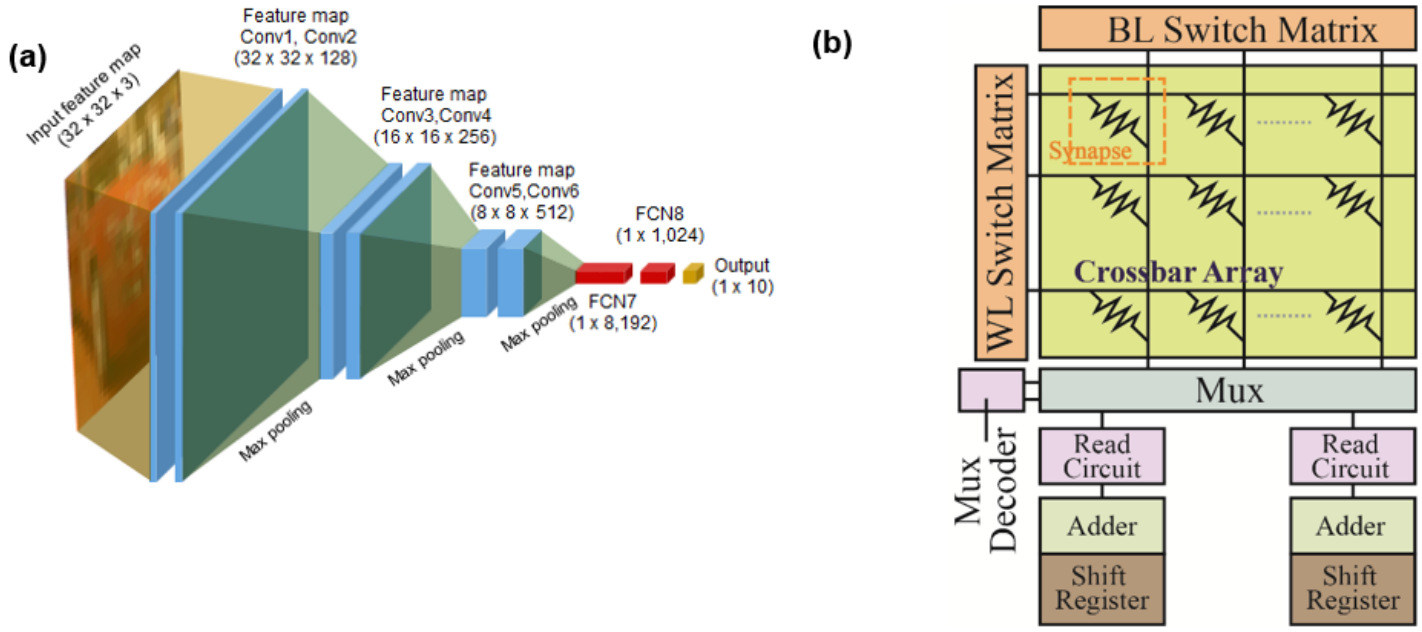

Figure 7

Schematic of neural network and PIM architecture. (a) VGG-8 neural network for CIFAR-10 image recognition. The convolutional layers are marked as Conv1 to Conv6, and the corresponding feature maps are indicated. The fully-connected (FCN) layers are identified as FCN7 and FCN8 with their dimension of weights in the brackets. (b) Design of a tile of PIM architecture embedding the GeO_x ReRAM synapse array.

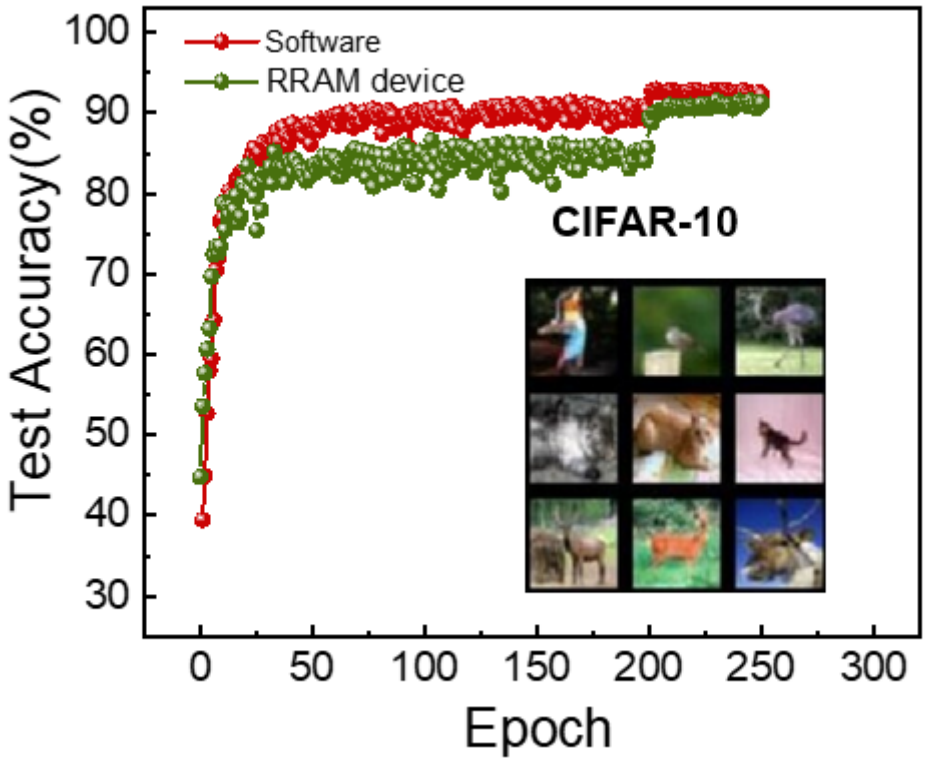

Figure 8

Accuracy in image test as a function of the number of training epochs for the GeO_x synapse array in comparison with the software neural network. The inset shows a subset of the CIFAR-10 input data.

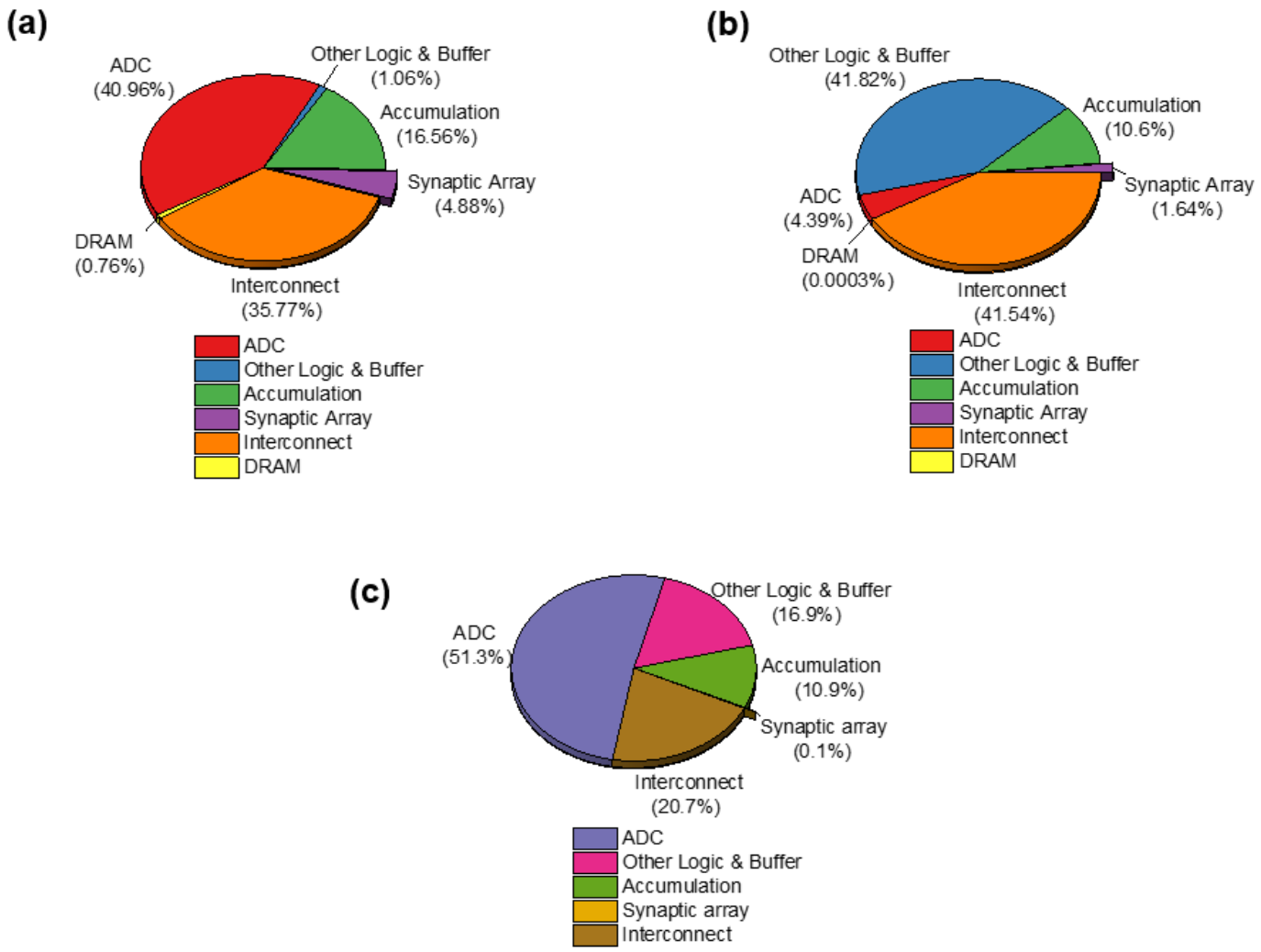

Figure 9

Pie charts illustrating the distributions of major metrics. (a) Energy consumption, (b) latency, and (c) area occupied by various hardware elements in the chip-level PIM architecture. Except the off-chip DRAM memory domain used for the storage of intermediately generated data from the PIM chip (can be optional depending on architecture design), all the other components can be presumably implemented on chip.

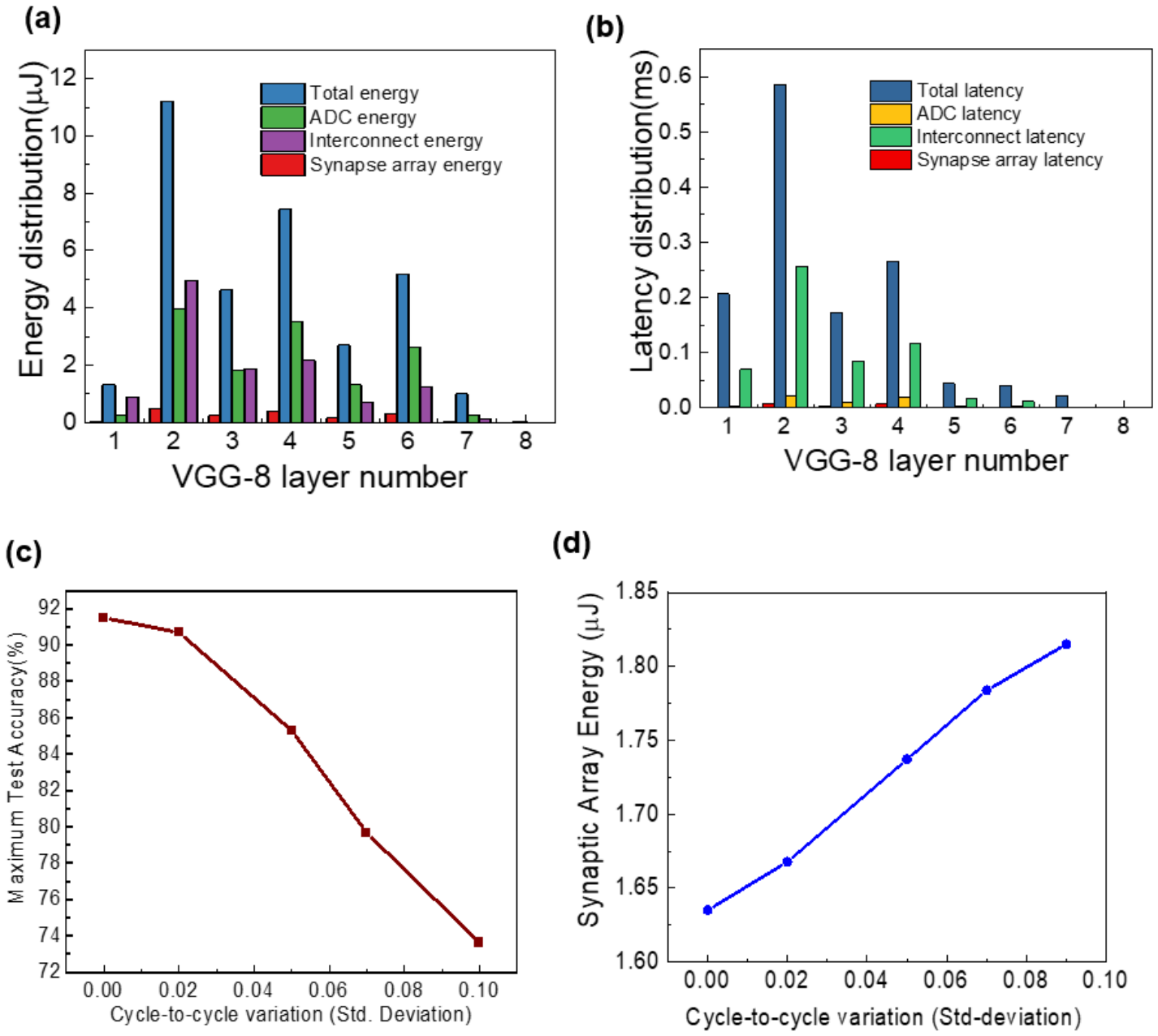

Figure 10

Histograms demonstrating (a) energy and (b) latency distribution across the different computational layers in the VGG-8 network. (c) Maximum test accuracy and (d) inference energy in the synapse array as a function of variation in the cycle-to-cycle GeO_x ReRAM operations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GraphicalAbstract.png