scientific reports

Check for updates

OPEN

Charge-trap synaptic device with polycrystalline silicon channel for low power in-memory computing

Min-Kyu Park[1], Joon Hwang[1], Soomin Kim[2], Wonjun Shin[1], Wonbo Shim[3], Jong-Ho Bae[4], Jong-Ho Lee[1] & Seongjae Cho[2]✉

Processing-in-memory (PIM) is gaining tremendous research and commercial interest because of its potential to replace the von Neumann bottleneck in current computing architectures. In this study, we implemented a PIM hardware architecture (circuit) based on the charge-trap flash (CTF) as a synaptic device. The PIM circuit with a CT memory performed exceedingly well by reducing the inference energy in the synapse array. To evaluate the image recognition accuracy, a Visual Geometry Group (VGG)-8 neural network was used for training, using the Canadian Institute for Advanced Research (CIFAR)-10 dataset for off-chip learning applications. In addition to the system accuracy for neuromorphic applications, the energy efficiency, computing efficiency, and latency were closely investigated in the presumably integrated PIM architecture. Simulations that were performed incorporated cycle-to-cycle device variations, synaptic array size, and technology node scaling, along with other hardware-sense considerations.

Keywords Processing-in-memory (PIM), Memory wall, Charge-trap flash, poly-Si channel, Inference, Hardware neural network, Off-chip learning

Modern computing workloads are dominated by data-intensive applications, such as deep learning, which demand accelerated computing capabilities at a lower energy budget. The remarkable success of the latest deep neural network (DNN) applications in diverse fields, such as computer vision, language translation, and audio processing, has further increased the demand for efficient hardware support for newer and larger DNN models. Huge language models, such as WuDao 2.0 (1.75 trillion parameters), GPT-3 (175 billion parameters), and Bloom (176 billion parameters), are already stretching the limits of modern computing capabilities[1]. Although hardware architectural innovations such as tensor processing units (TPU) and graphical processing units (GPU) have constructively supplemented state-of-the-art DNN computations, the inherent serialism in data transfer between memory and processing units has severely constrained further progress. This architectural design flaw is widely known as the Von Neumann bottleneck or *memory wall*. An alternative computing paradigm that relies on merging the processing and memory domains is processing-in-memory (PIM) computing[2–4]. The PIM was developed in the 1970s with the aim of reducing data shuttling latency and thereby shifting computations within a storage unit. In recent years, PIM research has focused on accelerating multiply-and-accumulate (MAC) operations, which are the core mathematical operations in DNN algorithms. A plethora of novel nonvolatile memories, such as resistive random-access memories (RRAMs), phase change memories (PCM), magnetic RAMs (MRAMs), and ferroelectric RAMs (FRAMs), have emerged as synaptic devices owing to their fast and low-power operation capabilities[5–15]. Flash memory devices have also received considerable attention because of their CMOS compatibility, mature technology, excellent reliability, multilevel operation, and scalability[16–18]. NAND and NOR flash memories have become the mainstay of commercial memory technologies with ultrahigh-density storage capabilities. The recent development of three-dimensional (3D) NAND flash memories has further increased research interest in these technologies as prospective devices for PIM architecture realization.

The fundamental element behind these architectures is the charge trap (CT) memory, which works on the basis of threshold voltage (V_T)modulation based on the trapping and detrapping of electrons during program and erase operations. By varying the channel conductance through V_T modulation, the charge trap flash (CTF) device can be used effectively as a synaptic device to realize the PIM architecture. Several groups have reported

[1]Department of Electrical and Computer Engineering with Inter-university Semiconductor Research Center (ISRC), Seoul National University, Seoul 08826, Republic of Korea. [2]Division of Electronic and Semiconductor Engineering, Ewha Womans University, Seoul 03760, Republic of Korea. [3]Department of Electrical and Information Engineering, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea. [4]School of Electrical Engineering, Kookmin University, Seoul 02707, Republic of Korea. ✉email: felixcho@ewha.ac.kr

CTF-based arrays for both convolutional and spiking neural networks[19,20]. However, these early results clearly lack more detailed system-level evaluations and benchmarks.

In this study, we fabricated a low-power CTF with a poly-Si channel and evaluated its performance as a synapse element in PIM-based system-level simulations. The experimentally measured device characteristics were used for system-level simulations with the necessary peripheral circuitry on a 22-nanometer CMOS technology node[21]. A more complex image-recognition test using the Canadian Institute for Advanced Research (CIFAR)-10 dataset was evaluated with the PIM circuit architecture comprising all peripheral circuitry. Further, efforts have been dedicated to more practical and substantial evaluations of the system in terms of energy efficiency, latency in data transfer, and area occupancy by the system-level simulation tasks. Additionally, the effect of nonideal behaviors in the CTF are considered for evaluating a presumably designed PIM architecture. Finally, the CTF device system-level performance was evaluated at lower technology nodes to examine the scaling trends of the device.

Results and discussion

Device characteristics & operation

A schematic of the CTF device and transmission electron microscopy (TEM) image of the fabricated CTF device are shown in Fig. 1a[22]. Note that a detailed explanation of device fabrication steps is explained in the "Methods" in Section. The fabricated device is a thin-film transistor (TFT) type. The device utilizes poly-silicon as the channel material, which allows for relatively low current flow, suitable as a PIM device. As depicted in the TEM image in Fig. 1a, the source and drain are electrically isolated, and the device features a *p*-type doped body. By providing holes, this *p*-body plays a crucial role in memory operations, particularly during the erase (ERS) operation. Note that the height of the source and drain electrodes is lower than that of the *p*-body, resulting in a portion of the poly-Si channel near the source and drain being curved. This structural characteristic causes the electric field to concentrate in the curved channel, enhancing gate controllability and providing a significant advantage is resisting short channel effects. Figure 1b shows the measured output characteristics of the fabricated CTF synaptic device, which exhibited saturation characteristics. The channel length and width were both set to 1 µm. The synaptic device can withstand a higher tolerance to unexpected drain voltage (V_{DS}) variations when the device is in saturation mode. Figure 1c shows the transfer characteristics of the fabricated CTF device. The transfer characteristics revealed that increasing the gate voltage (V_{GS}) increases the drain current (I_D), which is typical of an *n*-type MOS device. During device programming, as the positive gate voltage is applied with both the body and source grounded, electrons tunnel from the channel to the CT layer through the Fowler-Nordheim (FN) tunneling process, as schematically shown in Fig. 2a. The trapped electrons in the CT layer screen the effective gate electric field, thereby increasing V_T. Similarly, charges stored in the CT layer can be detrapped and tunneled back into the channel during the erase state by applying a negative gate voltage (Fig. 2b). The amount of charge stored in the CT layer can be modulated based on V_{GS}, which can be used to adjust the conductance or synaptic weight of the device.

As an increasing number of electrons are detrapped from the CT layer, the effective gate voltage increases, and hence, V_T increases during the erase process. The change in V_T and its effect on drain current I_D and synaptic conductance G can be explained as follows:

$$I_D = \alpha(V_{GS} - V_T)V_{DS}, \tag{1}$$

$$I_D = GV_{DS}, \tag{2}$$

From Eqs. (1) and (2),

$$G = \alpha(V_{GS} - V_T) \tag{3}$$

Here, α is the proportionality constant[23]. During the erase process, the change in threshold voltage (ΔV_T) is less than zero; hence, from Eq. (3), the change in device conductance (ΔG) is greater than zero. Therefore, the device undergoes a long-term potentiation (LTP) during the application of consecutive erase pulses (7 V/ 1 ms), as shown in Fig. 3a. However, during programming, $\Delta V_T > 0$; hence, $\Delta G < 0$. The CTF device undergoes long-term depression (LTD) with continuous program pulses (8 V/100 µs). In addition, the CTF device has a high on/off ratio of 1.19×10^3 and a low on-state conductance of 3.73×10^{-8} S. Such a high on/off ratio and low on-state conductance are highly suited for low-power synaptic array operations. The synaptic device energy is calculated as follows:

$$Energy_{CTF} = GV_{\mathrm{Read}}{}^2 T_{\mathrm{Read}}, \tag{4}$$

where V_{Read} is the drain read voltage (2 V), and T_{Read} is the read pulse width (100 ns). From Eq. (4), the maximum energy consumption for a read operation using the fabricated CTF device is ~ 14.94 fJ per synaptic read operation, which is much more energy efficient than previously reported devices and is similar to the energy consumption observed in biological systems (~ 10 fJ)[24]. The synaptic weights extracted from the CTF device were further used to evaluate their applicability in deep learning applications using a PIM architecture. Figure 3(b) shows the memory window (ΔV_T) between the fully erased and fully programmed states of the CTF device, which is a measure of the maximum range of conductance or weight values available for the device, as shown in Eq. (3). The memory window is 0.625 V, as shown in Fig. 3(b).

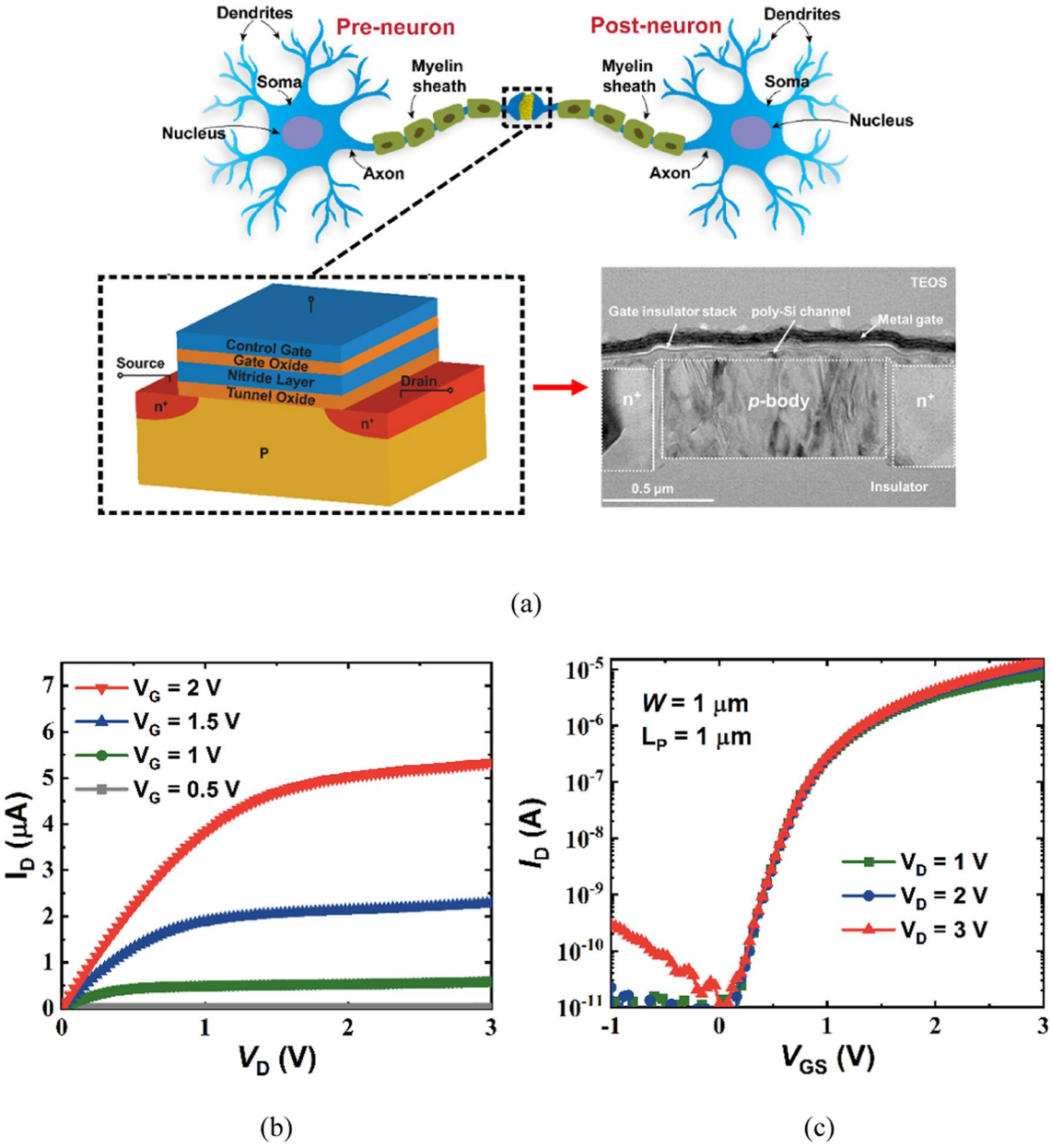


Fig. 1. (**a**) Schematic of biological synapse and the charge trap flash (CTF) device structure, and TEM image of the fabricated device. (**b**) output characteristics of the CTF device at various gate voltages (V_{GS}), and (**c**) transfer characteristics of the CTF device at various drain voltages (V_{DS}).

Neural network design and PIM architecture

The synaptic weights extracted from the LTP/LTD characteristics were further used for off-chip training of a hardware sense neural network using a PIM circuit design. A Visual Geometry Group (VGG)-8 neural network architecture was envisaged for this purpose using the Canadian Institute for Advanced Research (CIFAR)-10 dataset as network input. The neural network architecture is schematically shown in Fig. 4a. The VGG-8 neural network has six convolutional layers and two fully-connected layers. The CIFAR-10 dataset is a widely used three-channel (RGB) color image database with 50,000 training and 10,000 test images with a 32×32 resolution.

The images belong to one of the 10 output classes: airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks. Network training uses the rectified linear unit (ReLU) activation function during forward propagation. The loss is computed using the cross-entropy loss function, whereas the weight update is performed based on the stochastic gradient descent (SGD) algorithm. A batch size of 200 and an initial learning rate of 1.0 were used in the simulations. A decreasing learning rate strategy (*lrDecay*) is used during the VGG-8 training process[25]. *lrDecay* is a popular technique that is widely used in deep learning applications to improve the convergence of the network training process. This technique involves a large initial learning rate, which is decayed by a certain factor at predefined epochs. The advantage of this approach is that the initially large

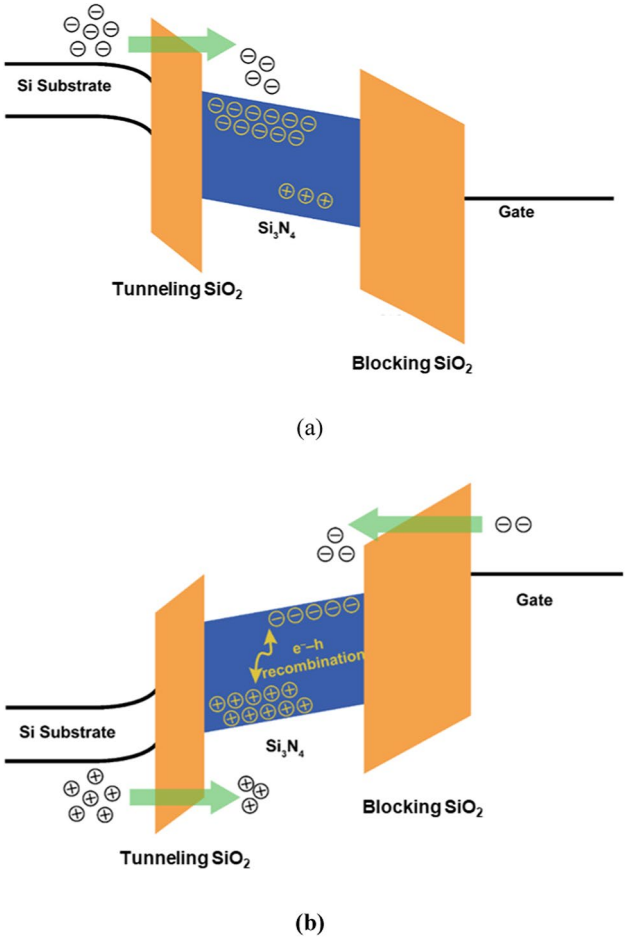
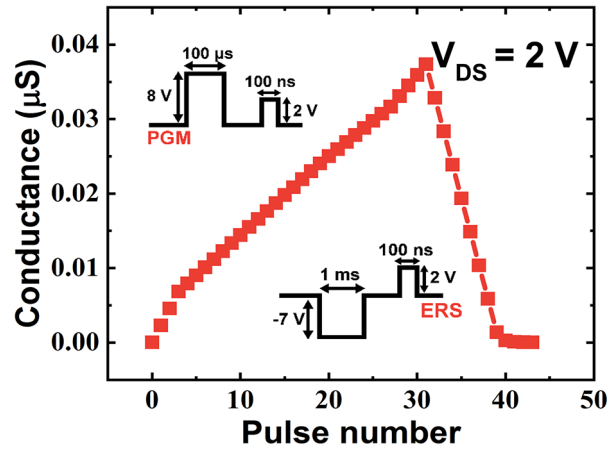


Fig. 2. Schematic of the energy band diagram of a CTF flash device during (**a**) the program state and (**b**) erase state.

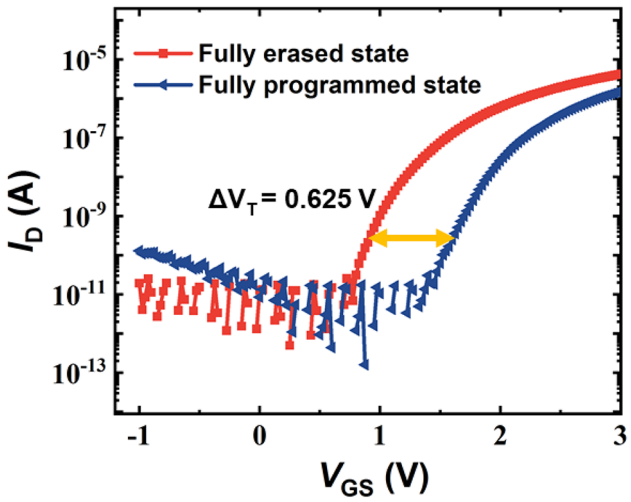
learning rate escapes accidental local minima at the beginning epochs during the SGD optimization. However, a high learning rate can overshoot the global minimum if it is maintained throughout the training process. Hence, adjusting to a lower learning rate can help the SGD optimization to effectively converge to the global minima[26]. The VGG-8 training process started with an initial learning rate of 1.0, which subsequently decreased to 1/8 after 200 epochs of training. In addition, Weights, Activations, Gradients, and Errors (WAGE) method is employed to quantize the input, weight, and output values[27]. WAGE applies a uniform quantization strategy where the first step is the shifting of each element by a constant value to ensure the data falls within the desired integer limits. Then, a linear constriction process is performed where the values are scaled and rounded to fit within the previously chosen bit precision. Note that for the simulation, 5-bit precision is used for the weights, and 8-bit precision is used for the input and the output values.

The MAC operation between the input vector and vectors in the weight matrix essentially constitute both convolutional layers and linear layers[28]. For each layer in the VGG8 model, the weights are mapped as the CTF cell conductance, where each junction between a row and a column crossing in a subarray represents a CTF device, as shown in Fig. 4b. During the inference operation, the input vector or the activation output from the previous layer is passed as parallel row-wise voltages. At each cross point of the sub-array, the input voltage is multiplied by the CTF device conductance, which forms the multiply operation in MAC calculation. Finally, the current at the output of each column is the total current summed up using Kirchhoff's current law. This summation constitutes the accumulation in the MAC operation. The sub-array thus implements a one-shot implementation of the MAC operations at each column endings. Figure 4b shows the schematic of a convolutional layer operation and its weight kernels being mapped to a sample 3×3 subarray in the PIM simulation.

The PIM chip-level architecture is illustrated in Fig. 5[21]. The chip-level architecture has been hierarchically structured into tiles, which are further divided into processing elements (PEs) that constitute the final synaptic subarrays. To represent negative weights, two PEs in a pair represent positive and negative weights. The outputs of PEs are hierarchically transferred to digital modules through global buffers and then subtracted to get partial sums considering the sign of the weights. The partial sums are digitally accumulated to acquire the final MAC results. For neural network training applications, we require access to individual analog states of each memory cell. In order to have such precise control of the different analog states of the CTF device, we need to employ a



(a)



(b)

Fig. 3. Experimentally measured (**a**) LTP and long-term depression characteristics of a CTF device. (**b**) I_D–V_{GS} curve depicting the memory window of the CTF device.

selector transistor-based pseudo-crossbar array architecture. The main advantage of the selector transistor is that any arbitrary CT flash device can be programmed without interference and power leakage to half-selected cells[29]. The chip architecture considers various device non-idealities, peripheral circuit limitations, such as analog-to-digital converter (ADC) resolution, and technology scaling parameters. For technology scaling, the following parameters were added based on the predictive SPICE technology model: interconnect metal wire thickness, DC supply voltage (V_{DD}), logic transistor threshold voltage (V_T), transistor gate length, fringe capacitance, on (and off) current of the NMOS and PMOS transistors and the NMOS-to-PMOS transistor size ratio. For the proper functioning of a PIM circuit, peripheral circuits, including multiplexers (MUX), ADCs, buffers, interconnects (copper for the 22-nm technology node), adders, and shift-registers, are included. Parallel read-out schemes are regularly used, in which flash ADCs are responsible for receiving partial sums along the columns owing to MAC operations. ADCs have high precision and, hence, occupy a relatively large area. The adders and shift registers were used for the shift-and-sum of the input data over multiple input cycles. To reduce the on-chip memory requirements for large amounts of batch-wise data generated during the feedforward process, the PIM architecture relies on off-chip DRAM.

The energy consumption has similar components at each of the hierarchical structures of tiles, PEs, and subarrays. At the subarray level, the energy is the sum of energy consumed by the CTF synaptic subarrays and the energy consumed by each of the sub-circuit modules. For each of these subcircuits, the main energy component is the dynamic energy consumption of the constituent logic gates. For a logic gate with capacitance (C) and bias voltage (V_{DD}), the energy can be taken as CV_{DD}^2. The total subcircuit energy can then be calculated by summing up the gate energy consumption at each node multiplied by the number of times the subcircuit module has been accessed. For the synaptic array energy calculation, $Energy_{CTF}$ from (4) is multiplied by the number of CTF devices in the subarray. A similar procedure is then repeated for the PE level and the tile level

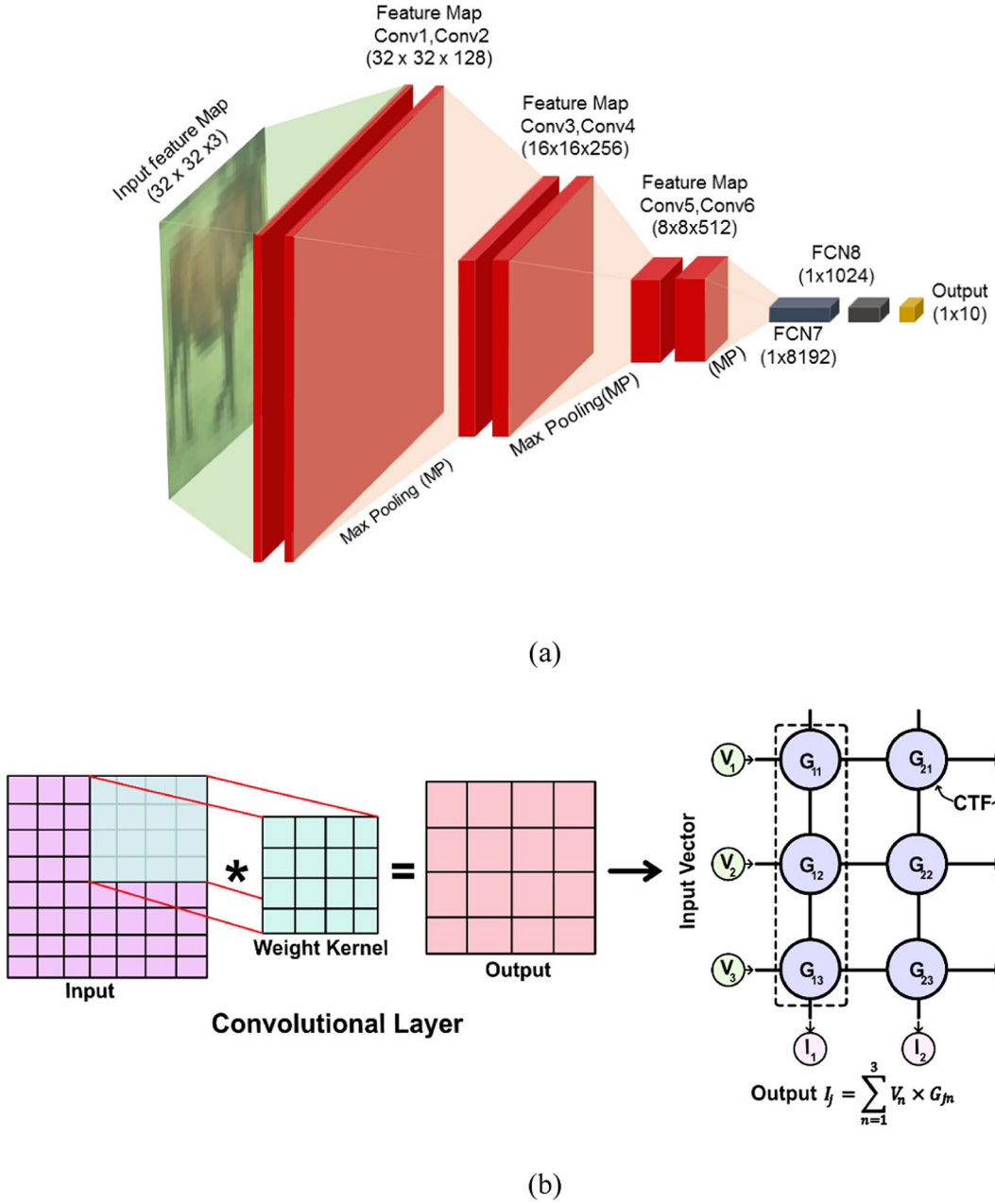


Fig. 4. Schematic of the (**a**) VGG-8 neural network architecture. (**b**) convolutional layer operation and corresponding MAC operation executed through VMM operations inside a simple 3 × 3 PIM subarray.

energy calculation to obtain the total chip energy. In addition, to calculate the compute efficiency, the total chip area is first calculated based on the CTF subarray area (based on 1T1R) and the logic gate dimensions as per the technology node used. Then, the number of MAC operations is calculated per second, which is computed in terms of Tera operations per second (TOPS). TOPS divided by the total chip area gives the area normalized compute efficiency. The energy efficiency is also similarly calculated by dividing the TOPS by the total chip energy computed, as explained above.

System-level simulations

For the PIM system performance evaluation, DNN + NeuroSim V2.0 package was adopted[21]. The energy calculation methodology adopted has been previously compared and validated with SPICE-based simulations[30]. The synaptic sub-array size was assumed to be 64 × 64, and a 5-bit analog-to-digital circuit (ADC) was used, which acts as the neuron circuit along the end of each synapse sub-array column. A column-sharing approach is used for the ADC, such that eight subarray columns share a single ADC. This is because the peripheral ADC size is too large, compared to the synapse subarray column pitch. Figure 6a shows the pattern recognition accuracy of the VGG-8 neural network based on the CTF PIM. The accuracies were reported as the average of five consecutive

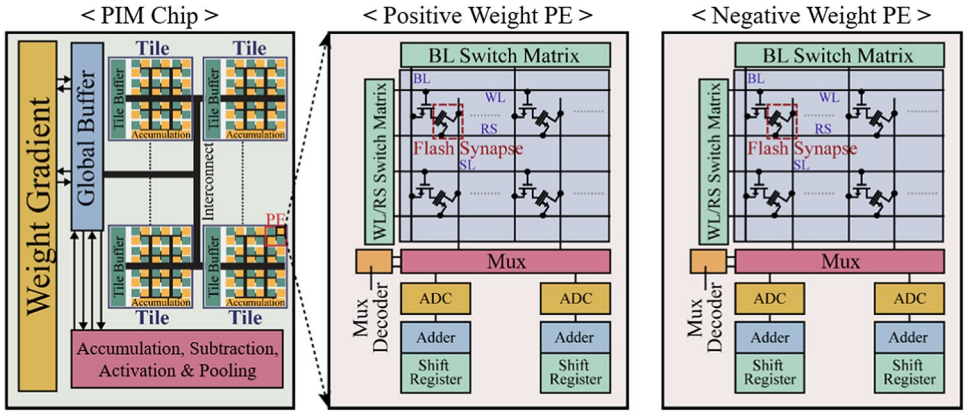
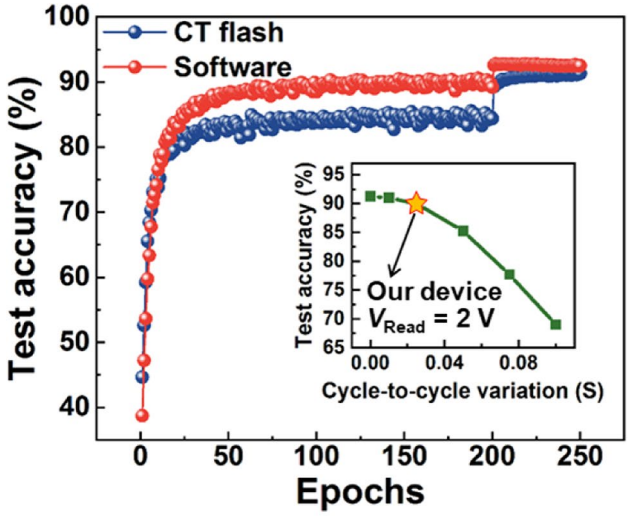
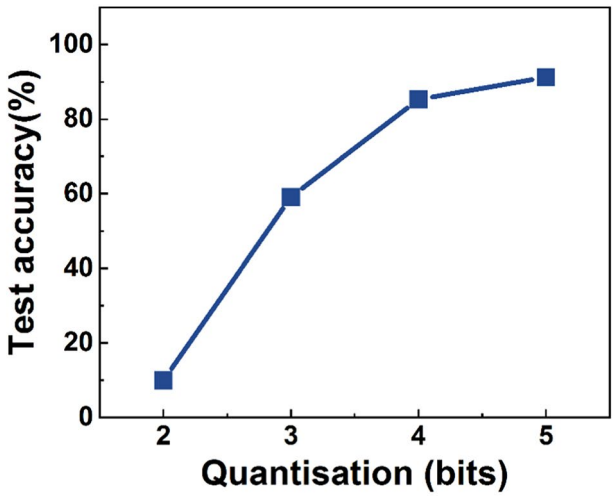


Fig. 5. Schematic of the PIM chip design with peripheral circuits.

runs with different seed values. A high-test accuracy of 91.3% was obtained for the CTF device, which is a small accuracy drop compared with that obtained from purely software-based training (92.5%). An abrupt jump was observed at 200 epochs in both the device- and software-based training processes owing to the learning-rate decay strategy used. At 200 epochs, the sudden learning rate decay helps the SGD algorithm find the global minima and, hence, attain a better test accuracy for image recognition. The effect of device nonidealities on PIM performance was of critical interest in several recent research reports[30]. The inset of Fig. 6a shows the effect of the test accuracy on the CTF cycle-to-cycle variations (C2C) in the conductance. The C2C variations were computed in terms of the standard deviation of each conductance value from the original value. The test accuracy remained above 90% until a C2C variation of 0.02, which reveals a high robustness of the fabricated CTF device for PIM implementation. Especially, the proposed charge trap synaptic device exhibits sufficiently low C2C variation to achieve high accuracy. Neural network quantization has also been extensively studied owing to its applicability to the development of compact neural networks. Although several techniques, such as pruning and singular value decomposition, already exist for neural network size reduction, quantization of weights and gradients has been widely adopted owing to their universality and exceptional performance retention[31]. Figure 6b shows the test accuracy variation of the CT flash device with bit quantization of weights and gradients during the inference process. It is shown that the device performs reasonably well above the 3-bit quantization. Performance may be further enhanced by employing techniques such as quantization-aware training. The distributions of energy and area consumption by various hardware units are depicted in Fig. 7a–b, respectively. The energy distribution was calculated by computing the dynamic energy consumed at every node by the total capacitance across a logic gate and the chip drive voltage, which was then summed to calculate the total energy consumed by a peripheral unit in a single operation. This value was then multiplied several times depending on the number of operations. As shown in Fig. 7a, the ADCs, interconnects, and accumulation circuits consume the largest share of the energy consumed by the PIM chip. In comparison, a synaptic array consumes minimal energy. Similarly, the bulk of the chip area is occupied by ADCs and interconnects, whereas the synaptic array occupies only a very small area. This also indicates that the ADCs and the interconnect architecture should be further optimized for developing energy- and area-efficient PIM systems. We also benchmarked our system-level simulations with previously reported results using different types of synaptic devices, as summarized in Table 1[30–39]. As shown in the table, the fabricated CTF device performs exceedingly well in terms of energy efficiency and latency, compared to other devices. The high energy efficiency may be due to the low read energy for a synaptic event of the CTF device, which is approximately 14.94 fJ, whereas the low latency may be attributed to the narrow-read pulse width of 100 ns used for each read operation. We also observed that the computing efficiency, which is an important parameter for evaluating the area efficiency, is better for the CTF device than for the RRAM and SRAM devices. The effect of the synaptic subarray size variation on the computational efficiency and energy efficiency is shown in Fig. 8a. As the synapse subarray size decreased from 256×256 to 16×16, the number of tiles in the PIM chip correspondingly increased from 15 tiles to 646 tiles. Hence, we observed that the reduction in synapse array size does not have a beneficial effect on the computation and energy efficiency because of the increase in tiles needed to store the small-sized synapse arrays. This can also be verified from the chip area and energy, which increase with the synapse-array size reduction, as shown in Fig. 8b. Similarly, the energy and computing efficiency also suffer from a reduced synapse array size owing to multiple additional ADCs (see Fig. 8c) required for the small-sized synapse sub-arrays. Technology scaling is an integral part of evaluating the PIM performance for current and future technological requirements. We evaluated the performance of the CTF device with technology nodes ranging from 45 to 10 nm. As previously discussed, the technology nodes were adapted into the PIM simulations using scaled logic and interconnect specifications from the predictive SPICE models. As shown in Fig. 9a, both the computation and energy efficiency increase with the technology node scaling to 10 nm. This may be attributed to the reduced area and energy consumption of the peripheral circuits owing to the low-power logic transistors in scaled technology nodes. This is also supported by the overall reduced energy consumption of the PIM chip with technology scaling, as shown in Fig. 9b. Although technology scaling does not directly affect the synapse subarrays, the energy consumption is observed to decrease owing to the effect of the synapse array switch-matrix energy reduction with technology node scaling.



(a)



(b)

Fig. 6. VGG-8 neural network simulation. (**a**) Test accuracy (%) variation with the number of training epochs. Inset shows the change in the test accuracy (%) with the cycle-to-cycle variations. (**b**) Test accuracy (%) variation with bit quantization of weight and gradients.

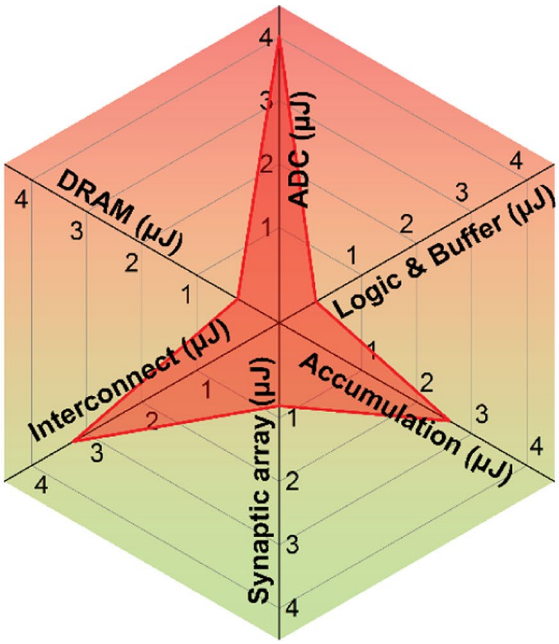
Conclusion

In this study, a more detailed and realistic system-level evaluation was performed for a presumed PIM architecture based on poly-Si CTF memory synapses fabricated for ultralow-power operation. A VGG-8 neural network structure was used for the hardware-sense simulation of the designed PIM architecture incorporating nonideality in synaptic weight modulation. It is evident from the system-level simulation that the CTF synapse array consumes extremely low energy and space, compared to other circuital elements that make up the neuromorphic PIM chip. The study of weight quantization reveals important information regarding the compressibility of neural networks. The synaptic subarray and technology node variations had a significant effect on the overall functioning of the PIM circuit. Judging from the high CIFAR-10 image-recognition accuracy of 91.31%, high-speed, and ultra-low-power operation capability, the poly-Si CTF memory is a plausible candidate for the hardware-oriented PIM architecture for chip implementation.

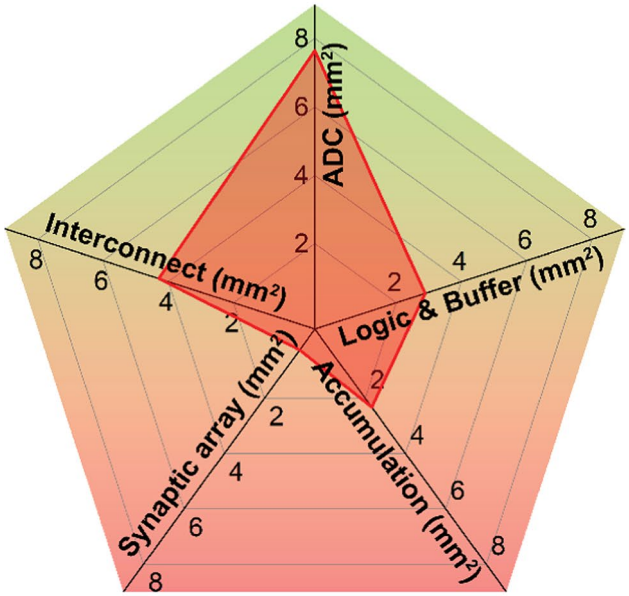
Methods

Device fabrication

The CTF synaptic device was fabricated through 0.5-μm Si CMOS processing. The channel material was polycrystalline Si deposited through low-pressure chemical vapor deposition (LPCVD), by which the inference energy can be drastically lowered. The tunneling oxide (3.6 nm), CT nitride layer (5.3 nm), and Al_2O_3 as the



(a)



(b)

Fig. 7. Radar plot showing the (**a**) energy distribution and (**b**) area distribution of various peripheral circuits within the PIM chip.

blocking oxide (6.2 nm) were deposited sequentially by thermal dry oxidation, LPCVD, and atomic-layer deposition (ALD), respectively.

Electrical characterization

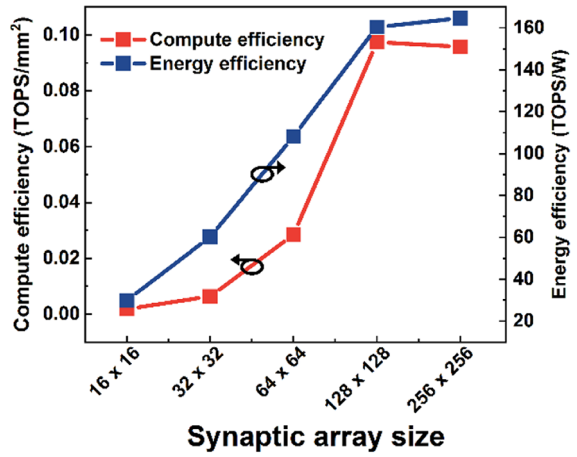
All the electrical characteristics of the CTF device were obtained at room temperature and inside an electrically shielded probe station. The device characteristics were measured using a Keithley 4200 A-SCS semiconductor parameter analyzer.

Parameter	81-SRAM[30]	KKAM[31]	FcFET[31]	FET[33]	PCM[34]	RRA.M[35]	RRAM[36]	MRAM[37]	PCM[38]	cDRAM[39]	This study
Technology node [mm]	65	22	22	60	14	40	65	28	14	28	*22 mrn
Chip area [mm^2]	0.055	45.55	25.52	16	-	1.55	0.38	9.72	104	32.23	–
Bitcell area IF^2I	– 700	8	16	–	– 64	-6	–	–	– 6	– 6	16
Accuracy (%) "	96.37[a]	_[b]	_[b]	93.2[a]	98.3[a], 85.6[b]	–		91.90[b]	98.3[a]	98.5[a]	89.5[b]
ON/OFF ratio	–	150	100	–	–	–	47.5	–	–	–	17.4
ADC precision	1–5-bit	5-bit	5-bit	–	12-bit	4-bit		8-bit	**Current Controlled Oscillator**	–	5-bit
Energy efficiency [TOPS/\V]	15.8–490	40.89	89.14	143.9	10.5	118.44	2.07	709.3	10.5	234.6	108.2
Latency [ms]	–	1.46	1.43	–	0.13x10[a] 4	–	–	–	–	–	2.49
Leakage Power [mW]	–	0.8	0.5	–	–	–	–	–	–	–	0.074
Compute efficiency [TOPS/mm^2]	112	0.019	0.033		1.587			3.9	1.59		0.028

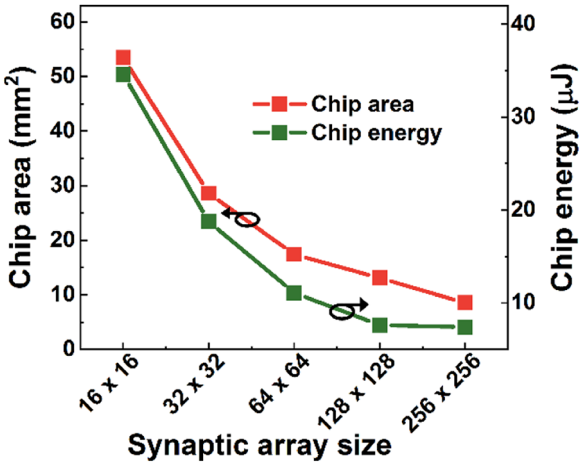
Table 1. Comparison of the system-level performances between the CTF-based neuromorphic systems in the existing reports. *Technology node only considering periphery circuits. [a]MNIST [b]CIFAR-10

Simulation environment

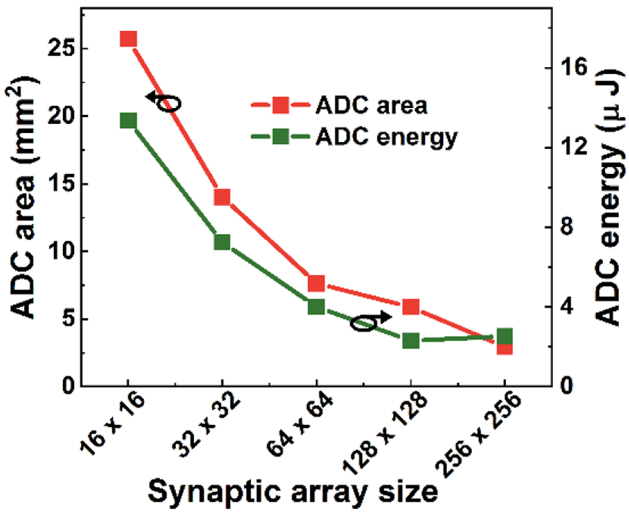
System-level simulations were carried out in a 32-core workstation with one of the most recent processor products, AMD Ryzen 9 Processor, 64-GB DRAM, and a high-capacity graphic processing unit (GPU), NVIDIA RTX 3090.



(a)

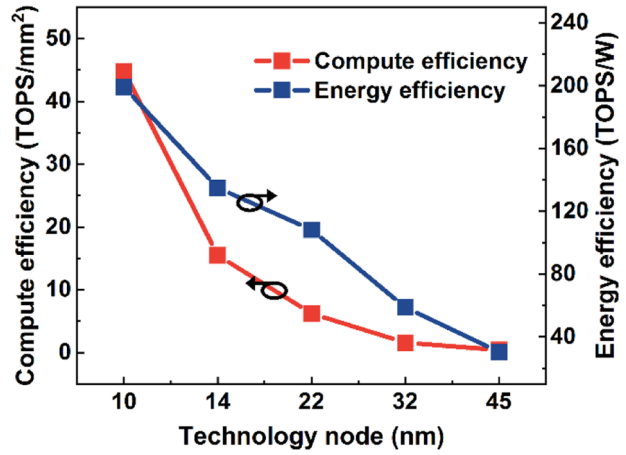


(b)

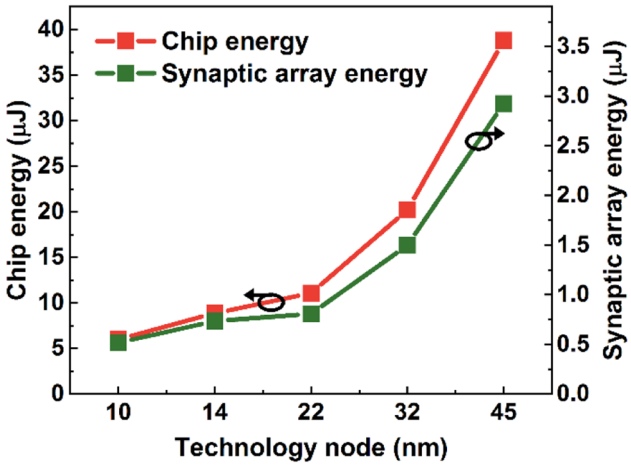


(c)

Fig. 8. Effect of the synaptic array size variation on (**a**) computed efficiency (TOPS/mm^2) and energy efficiency (TOPS/W). (**b**) Unit chip area (mm^2) and chip energy (µJ). (**c**) Unit ADC area (mm^2) and energy (µJ).



(a)



(b)

Fig. 9. Effect of technology node variation on (**a**) computed efficiency (TOPS/mm^2) and energy efficiency (TOPS/W). (**b**) Chip energy (μJ) and synaptic array energy (μJ).

Data availability

The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 2 October 2023; Accepted: 18 November 2024
Published online: 23 November 2024

References

1. Jing, K. & Xu, J. A Survey on Neural Network Language Models. *arXiv* (2019).
2. Ielmini, D. & Wong, H. S. P. In-memory computing with resistive switching devices. *Nat. Electron.* **1**, 333–343 (2018).
3. Sun, X. et al. PCM-Based Analog Compute-In-Memory: impact of device non-idealities on Inference Accuracy. *IEEE Trans. Electron. Devices*. **68**, 5585–5591 (2021).
4. Sebastian, A., Gallo, M. L., Khaddam-Aljameh, R. & Eleftheriou, E. Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* **15**, 529–544 (2020).
5. Cho, S. Volatile and Nonvolatile Memory Devices for Neuromorphic and Processing-in-memory applications. *J. Semicond. Technol. Sci.* **22**, 30–46 (2022).
6. Han, R. et al. Flash memory array for efficient implementation of deep neural networks. *Adv. Intell. Syst.* **3**, 2000161–2000161 (2021).
7. Choi, Y. J. et al. Insertion of Ag layer in TiN/SiN_x/TiN RRAM and its effect on filament formation modeled by Monte Carlo simulation. *IEEE Access.* **8**, 228720–228730 (2020).
8. Bang, S. et al. Gradual switching and self-rectifying characteristics of Cu/α-IGZO/p^+-Si RRAM for synaptic device application. *Solid-State Electron.* **150**, 60–65 (2018).
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.* **60**, 050905 (2021).

10. Kim, T. H. et al. Fabrication and characterization of TiO_x memristor for synaptic device application. *IEEE Trans. Nanotechnol.* **19**, 475–480 (2020).
11. Rasheed, U. et al. Resistive switching characteristics and theoretical simulation of a Pt/a-Ta_2O_5/TiN synaptic device for neuromorphic applications. *J. Alloys Compd.* **877**, 160204 (2021).
12. Ryu, J. H. et al. Zinc tin oxide synaptic device for neuromorphic engineering. *IEEE Access.* **8**, 130678–130686 (2020).
13. Lee, D. K. et al. Synaptic behaviors of HfO_2 ReRAM by pulse frequency modulation. *Solid-State Electron.* **154**, 31–35 (2019).
14. Meng, J. et al. A high-speed 2D optoelectronic in-memory computing device with 6-bit storage and pattern recognition capabilities. *Nano Res.* **15** (3), 2472–2478 (2022).
15. Zhao, J. et al. Charge trap-based carbon nanotube transistor for synaptic function mimicking. *Nano Res.* **14** (11), 4258–4263 (2021).
16. Lee, J., Cha, M. & Kwon, M. W. Charge trap flash structure with feedback field effect transistor for processing in memory. *J. Semiconduct Technol. Sci.* **23** (5), 295–302 (2023).
17. Lee, S. T. et al. High-density and highly-reliable binary neural networks using NAND flash memory cells as synaptic devices. In *2019 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, USA, (2020).
18. Park, S. et al. Highly-reliable cell characteristics with 128-layer single-stack 3D-NAND flash memory. In 2021 Symposium on VLSI Technology, Kyoto, Japan, (2021).
19. Choi, H. S., Kim, H., Lee, J. H., Park, B. G. & Kim, Y. AND flash array based on charge trap flash for implementation of convolutional neural networks. *IEEE Electron Device Lett.* **41** (11), 1653–1656 (2020).
20. Lee, S., Kim, H., Lee, S. T., Park, B. G. & Lee, J. H. SiO_2 fin-based flash synaptic cells in and array architecture for binary neural networks. *IEEE Electron. Device Lett.* **43** (1), 142–145 (2021).
21. Peng, X. et al. 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.* **40**, 2306–2319 (2021).
22. Park, M. K. et al. Cointegration of the TFT-Type AND flash synaptic array and CMOS circuits for a hardware-based neural network. *IEEE Trans. Electron. Devices.* **70**, 1, 93–98 (2022).
23. Taur, Y. & Ning, T. H. *Fundamentals of Modern VLSI Devices* (Cambridge University Press, 2013).
24. Kuzum, D., Yu, S. & Wong, H. P. Synaptic electronics: materials, devices and applications. *Nanotechnology* **24** (38), 382001 (2013).
25. You, K., Long, M., Wang, J. & Jordan, M. I. How Does Learning Rate Decay Help Modern Neural Networks? *arXiv* (2019).
26. Kleinberg, B., Li, Y. & Yuan, Y. An alternative view: When does SGD escape local minima? In International Conference on Machine Learning (PMLR), (2018).
27. Luo, Y., Wang, P. & Yu, S. Accelerating On-Chip Training with Ferroelectric-Based Hybrid Precision Synapse, *ACM J. Emerg. Technol. Comput. Syst.*, vol. 18, no. 2, pp. 1–20, Apr. (2022).
28. Yu, S. *Semiconductor Memory Devices and Circuits*, First edition. USA: CRC Press, (2022).
29. Guo, X. et al. Dec., Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology, in 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA: IEEE, 6. 5.1–6.5.4. doi: https://doi.org/10.1109/IEDM.2017.8268341. (2017).
30. Lu, A., Peng, X., Li, W., Jiang, H. & Yu, S. NeuroSim simulator for compute-in-memory hardware accelerator: validation and benchmark. *Front. Artif. Intell.* 70. (2021).
31. Qin, H. et al. Binary neural networks: a survey. *Pattern Recognit.* **105**, 1072817 (2020).
32. Yu, C., Yoo, T., Chai, K., Kim, T. & Kim, B. A 65-nm 8T SRAM compute-in-memory macro with column ADCs for processing neural networks. *IEEE J. Solid-State Circ.* 57. (2022).
33. Baba, H. et al. *Novel Analog In-Memory Compute with > 1 nA Current/Cell and 143.9 TOPS/W Enabled by Monolithic Normally-off Zn-rich CAAC-IGZO FET-on-Si CMOS Technology, in 2021 IEEE International Electron Devices Meeting (IEDM)*pp. 21 (IEEE, Dec. 2021). https://doi.org/10.1109/IEDM19574.2021.9721312
34. Khaddam-Aljamesh, R. et al. Jun. HERMES Core—A 14nm CMOS and PCM-based In-Memory Compute Core using an array of 300ps/LSB Linearized CCO-based ADCs and local digital processing, in *2021 Symposium on VLSI Technology*, pp. 1–2. (2021).
35. Yoon, J. H. et al. A 40-nm 118.44-TOPS/W voltage-sensing compute-in-memory RRAM macro with write verification and multi-bit encoding, *IEEE J. Solid-State Circ.* 57 (3), 845–857, Mar. (2022).
36. Zhang, F. et al. A 65-nm RRAM compute-in-memory macro for genome processing. *IEEE J. Solid-State Circ.* **59** (7), 2093–2104 (2024). Jul.
37. Xie, W. et al. A 709.3 TOPS/W Event-Driven Smart Vision SoC with High-Linearity and Reconfigurable MRAM PIM, *2023 IEEE Symposium on VLSI Technology and Circuits*, Jun. (2023).
38. Khaddam-Alijamesh, R. et al. HERMES-Core—A 1.59-TOPS/mm2 PCM on 14-nm CMOS In-Memory Compute Core using 300-ps/LSB linearized CCO-Based ADCs. *IEEE J. Solid-State Circ.* 57. (2022).
39. Zhan, Y., Yu, W. H., Un, K-F. & Martins, R. A 28-nm 18.7 TOPS/mm2 89.4-to-234.6 TOPS/W 8b single-finger eDRAM compute-in-memory macro with bit-wise sparsity aware and kernel-wise weight update/refresh. *IEEE J. Solid-State Circ.* (2024).

Acknowledgements

This work was supported by the National Research Foundation (NRF) Grants funded by the Ministry of Science and ICT of Korea (MSIT) (Grant No. 2021M3H4A6A01048300 and RS-2023-00258527).

Author contributions

M.-K. P. performed the device fabrication and revised the manuscript. J. H. took part in the manuscript revision and validated the measurement results. S. K. performed the draft writing and a part of system-level simulation. W. Shin provided the measurement results from the fabricated devices. W. Shim performed the system-level simulation with optimization of the codes. J.-H. B. rendered the process integration for fabricating the synaptic devices. J.-H. L. schemed the overall process integration and provided the fabrication and measurement environments. S. C. clarified the motivation and direction of the research, advised for device fabrication, schemed the simulation approaches and tools, and wrote the manuscript. All the authors have reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024