Yu *et al. Nano Convergence* (2026)13:42
https://doi.org/10.1186/s40580-026-00575-9

Nano Convergence

FULL PAPER Open Access

Capacitive in-memory vector–matrix multiplication computing with charge-trap memcapacitors

Junsu Yu[1,2], Hwiho Hwang[3], Hyungjin Kim[3*] and Woo Young Choi[1*]

Abstract

Energy-efficient neural computing is increasingly limited not by computational throughput but by memory access and data movement. In-memory computing architectures offer a promising solution by collocating storage and computation, yet their practical realization remains constrained by static power dissipation, thermal challenges, and limited integration density in existing memory devices. Here, we demonstrate a vertically stackable charge-trap-flash (CTF)-based memcapacitor array that combines the high integration density of 3D NAND technology with the charge-domain computation. A continuous in-situ-doped N^+ bottom readout electrode is introduced beneath the lightly doped active layer, decoupling lateral readout resistance from depletion-based capacitance modulation and alleviating the resistance–depletion trade-off of the previous architecture. The fabricated 24×48 memcapacitor array operates through transient charge displacement and demonstrates highly uniform array-level characteristics ($\sigma/\mu < 0.37\%$), reliable 16-state closed-loop weight programming, and linear charge-domain VMM with an error below 0.227%. When scaled toward modern 3D NAND dimensions, a verification–inference electrostatic mismatch introduces systematic VMM error, which is reduced by up to 84.6% through an intercell trapped-charge scheme in TCAD simulations. Geometric scaling based on the CV^2 relation projects femtojoule-level intrinsic cell read energy at scaled dimensions. A hybrid hardware–software spiking neural network evaluation, in which only the final 24×10 layer is mapped to measured arrays, achieves 88.01% CIFAR-10 accuracy compared with 88.17% in software. These results establish vertically stackable memcapacitors as a scalable charge-domain computing platform combining reliable array-level operation with high-density vertical integration.

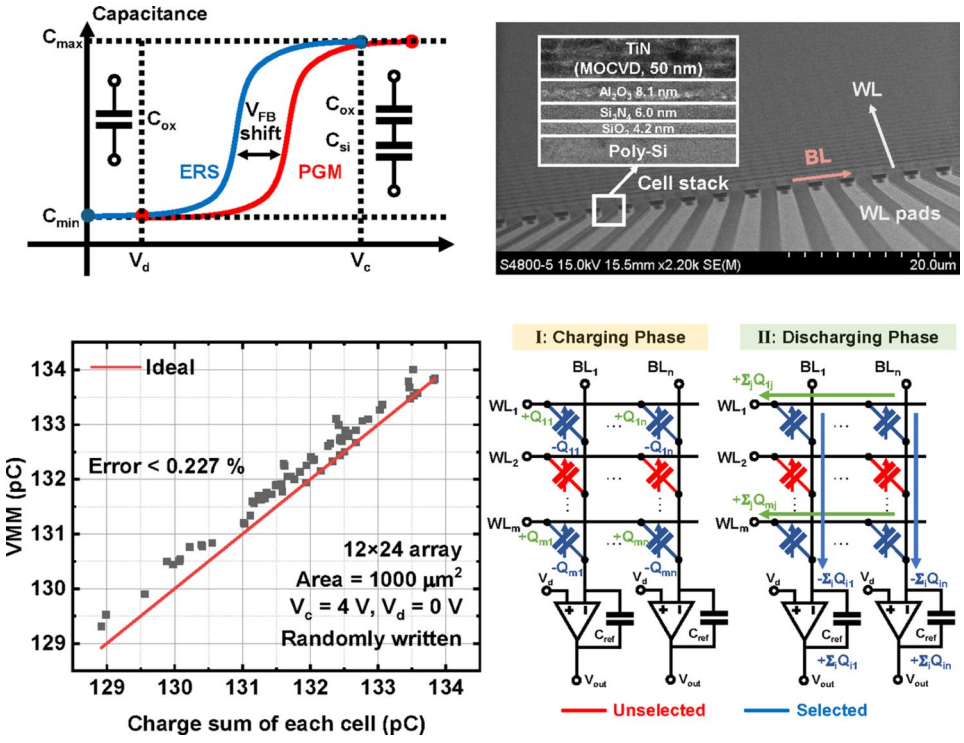
Keywords Charge trap flash (CTF), 3D NAND, In-memory computing (IMC), Memcapacitors, Artificial neural networks

*Correspondence:
Hyungjin Kim
hkim12@hanyang.ac.kr
Woo Young Choi
wooyoung@snu.ac.kr

Full list of author information is available at the end of the article

© The Author(s) 2026. **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/.

Graphical abstract



1 Introduction

The rapid advancement of artificial intelligence (AI) and the emergence of large language models (LLMs) have dramatically increased computational and memory demands [1–3]. As model parameters scale into the hundreds of billions, feed-forward network (FFN) layers become the dominant contributors to computational complexity, since the vector–matrix multiplication (VMM) workload grows significantly with model size. This scaling amplifies not only arithmetic operations but also data transfer across memory hierarchies. Despite the remarkable arithmetic throughput of modern GPUs and TPUs, their overall performance remains constrained by the limited capacity of high-bandwidth memory (HBM). Consequently, model weights must be repeatedly transferred between HBM, DRAM, and secondary storage, leading to substantial latency and energy overhead. This imbalance between computation and data movement exposes the fundamental inefficiency of von Neumann architectures and underscores the urgent need for memory-centric computing architectures that tightly integrate storage and computation to sustain LLM-scale workloads [4, 5]. As model size and context length continue to increase, the demand for higher memory capacity further necessitates highly integrated memory architectures based on three-dimensional (3D) integration.

Neuromorphic and in-memory computing (IMC) have emerged as promising paradigms for realizing such architectures. By performing analog VMM operations directly within memory arrays, IMC systems can substantially reduce data-transfer overhead [6–13]. Nevertheless, their widespread adoption has been impeded by device-level limitations. Memristor-based devices, such as resistive switching memory and phase change memory, offer analog programmability and enable conductance-based VMM through current summation during read pulses, but their scalability toward highly integrated three-dimensional arrays remains challenging [14–24]. Ferroelectric devices provide low-voltage switching but often exhibit narrow analog tuning ranges, endurance degradation, or read-disturb effects [25–35]. Meanwhile, 3D NAND flash technology already provides ultra-high vertical integration density, which is sufficient for emulating large-scale neural networks; however, its serially connected cell architecture prevents parallel analog operations required for efficient VMM [36–43]. These constraints highlight a fundamental trade-off between integration density and analog parallelism in existing memory technologies. To overcome this limitation, a promising direction is to enable capacitive readout within 3D NAND-compatible structures, allowing parallel VMM operations across vertically stacked cells while preserving the inherent density advantage of NAND

flash [44–50]. In contrast to conductance-weighted current summation, capacitive VMM encodes weights in capacitance and performs computation through transient displacement-charge accumulation. In this context, leveraging the existing charge-trap-flash (CTF) stack of 3D NAND as a memcapacitor cell provides a practical and technologically mature pathway toward vertically stackable analog computing arrays, effectively bridging the gap between ultra-high density and parallel analog computing.

Building upon this concept, this work experimentally demonstrates a vertically stackable CTF-based memcapacitor array that integrates the density advantage of 3D NAND with the energy efficiency of capacitive readout. Compared with our previous CTF memcapacitor architecture [44], the present device incorporates a continuous in-situ-doped N^+ bottom readout electrode beneath the lightly doped active layer, decoupling lateral readout resistance from depletion-based capacitance modulation. This structural modification overcomes the resistance–depletion trade-off of the previous architecture and enables highly uniform operation across the 24 × 48 array. The fabricated devices exhibit excellent electrical performance, including a large dynamic range (MW_Q), stable multi-bit weight modulation, and long-term retention exceeding ten years, validating the reliability of the capacitive readout mechanism. Furthermore, geometric scaling based on the CV^2 relation projects intrinsic cell-level read energy in the femtojoule range, although practical energy depends on parasitic capacitance and peripheral circuitry. However, as device dimensions approach those of commercial 3D NAND and array density increases, electrostatic interactions between neighboring cells, as well as inconsistencies between verification and inference conditions, can introduce discrepancies in the effective charge accumulation region, leading to systematic nonidealities in analog VMM accuracy. TCAD simulations at scaled cell dimensions comparable to commercial NAND flash reveal the emergence of VMM errors induced by intercell charge interference, which are effectively suppressed through electrostatic stabilization of the intercell region via intercell trapped charge [36]. Finally, a hybrid hardware–software SNN evaluation maps only the final 24 × 10 layer to the measured arrays and achieves software-comparable accuracy. The improved VMM directly translates to enhanced inference accuracy, highlighting the importance of co-optimizing electrostatics and neural network computation in highly integrated IMC systems. These results establish vertically stackable memcapacitors as a promising charge-domain computing platform combining reliable array-level operation with the potential for high-density vertical integration.

2 Device concept and operating principles

Figure 1a compares the operating principles of three major categories of synaptic devices used in neuromorphic hardware: FET-based devices, two-terminal resistive devices, and memcapacitors. In conventional field-effect-transistor synapses and two-terminal resistive devices, such as resistive switching memory and phase-change memory, vector–matrix multiplication is performed by modulating device conductance. Because the output current follows the relation $I = GV$, conductance-based VMM accumulates current through the programmed memory elements during finite read pulses. In highly integrated three-dimensional arrays, such current-based operation can introduce additional thermal-management considerations. In contrast, capacitor-based memory devices, or memcapacitors, encode synaptic weights through capacitance modulation rather than conductance modulation. Their operation follows the relation $Q = CV$, where the output signal is a displacement charge generated only during voltage transitions. Accordingly, memcapacitor-based VMM performs computation through transient displacement-charge accumulation rather than conductance-weighted current summation. This charge-domain operation, combined with the ability to continuously tune capacitance via trapped charge, makes memcapacitors a promising platform for high-density three-dimensional analog computing arrays.

Building on the conceptual distinction discussed above, the structural implementation of the proposed charge-trap-flash-based memcapacitor is illustrated in Fig. 1b. The device adopts a planar crossbar configuration composed of TiN word lines and a charge-trap stack consisting of $Al_2O_3/Si_3N_4/SiO_2$ (A/N/O) over poly-Si channel layers. To ensure stable and uniform capacitive readout across the array, two body configurations, n-type and p-type, were evaluated as shown in Fig. S1, confirming that the n-type body exhibits markedly improved uniformity in both C_{max} and C_{min} across the array. This behavior originates from the higher electron mobility compared to holes, which reduces series resistance along the channel and suppresses voltage degradation caused by IR drop, even for cells located farther from the driving terminals. In contrast, the p-type body shows a gradual reduction in measured capacitance with increasing word-line (WL) index, indicating resistance-induced degradation of the capacitive readout signal. Owing to its superior readout uniformity and reduced resistive loss, the n-body configuration is adopted for all subsequent demonstrations. This memcapacitor structure can naturally extend into a vertical configuration by stacking multiple WLs along a cylindrical channel, as shown in Fig. 1c. Each WL forms an independent gate-all-around (GAA) capacitive cell, enabling multiple memory elements to be integrated along a single bit-line (BL) string. This architecture

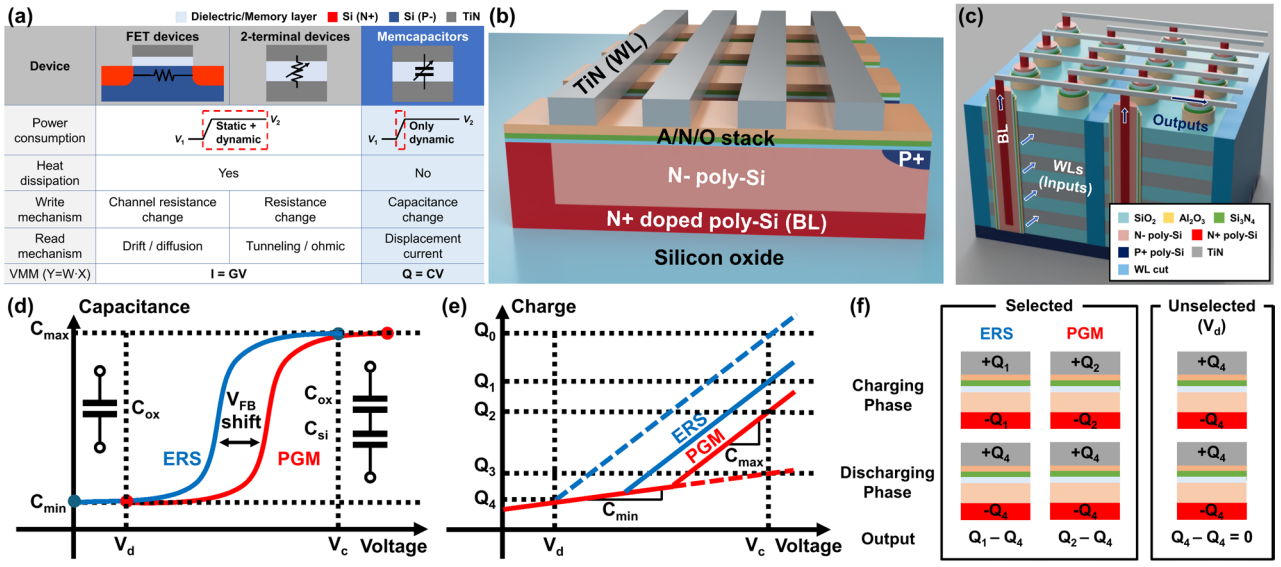
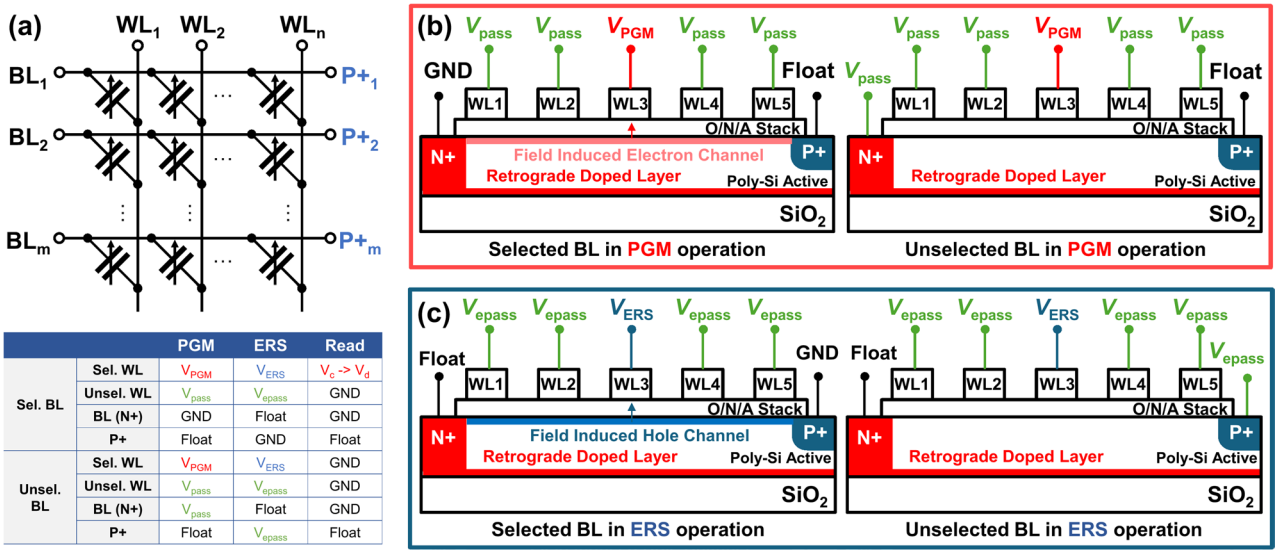


Fig. 1 Operating principle and device architecture of the CTF-based memcapacitor. **a** Comparison of operating principles among three representative synaptic device classes used in neuromorphic hardware: FET-based synapses, two-terminal resistive devices, and memcapacitors. While conductance-domain devices perform VMM through current flow during finite read pulses, memcapacitors encode synaptic weights via capacitance modulation and generate output signals only through transient displacement charge. The power-consumption comparison in panel (**a**) refers to the intrinsic memory-cell readout; practical arrays additionally include peripheral-circuit power. **b** Schematic illustration of the planar CTF–based memcapacitor adopting a TANOS (TiN/ Al_2O_3/Si_3N_4/SiO_2/Si) gate stack with a retrograde-doped poly-Si channel, forming a crossbar-compatible cell. **c** Extension of the planar memcapacitor into a vertically stacked configuration, where multiple WLs are integrated along a cylindrical channel to form GAA capacitive cells compatible with 3D NAND architectures. **d***C–V* characteristics of the charge-trap stack, showing a flat-band voltage shift between ERS and PGM states due to trapped charge in the Si_3N_4 layer. **e** Illustration of displacement-charge-based readout, where the output charge follows $Q = CV$ and varies according to the programmed capacitance state, enabling analog weight representation. **f** Summary of induced charge behavior during read operations, highlighting that selected cells generate net displacement charge during voltage transitions, while unselected cells remain electrically inactive

preserves the compact $4F^2$ footprint of NAND flash technology while enabling massive vertical scalability, allowing high-density weight storage and parallel vector–matrix multiplication within the same string.

Figure 1d presents the capacitance–voltage (C–V) characteristics of the charge-trap stack, showing a clear flat-band voltage shift between the erased (ERS) and programmed (PGM) states. Charge trapped in the Si_3N_4 layer modulates the surface potential of the channel, producing a measurable change in capacitance under identical bias conditions. Cells in the ERS state, which contain fewer trapped electrons, enter accumulation at lower gate voltages and therefore exhibit larger capacitance. In contrast, PGM cells shift the C–V curve toward higher voltages, reducing capacitance in the operating region. This electrostatic modulation forms the basis of analog weight storage in the proposed device. The state-dependent C–V modulation directly translates into differences in displacement charge, as illustrated in Fig. 1e. During readout, the output charge follows the relation $Q = CV$ and is therefore determined by the PGM capacitance of each cell. A fully ERS cell exhibits the largest effective capacitance and yields the maximum displacement charge ($Q_0 - Q_4$), corresponding to the highest synaptic weight, whereas a fully programmed cell produces the minimum displacement charge ($Q_3 - Q_4$). Intermediate states generate proportionally graded charge levels, enabling multi-bit weight representation. The usable dynamic range is defined as $MW_Q = Q_0 - Q_4 = (C_{max} - C_{min})(V_c - V_d)$. Figure 1f summarizes the induced charge behavior under different cell states and operation modes. Selected cells undergo both charging and discharging phases, producing a net displacement charge determined by their programmed capacitance. In contrast, unselected cells maintain both terminals at a discharge voltage V_d throughout the read cycle and therefore generate zero output.

The biasing scheme for PGM, ERS, and read operations is designed to ensure that each operation is confined to the selected WL–BL intersection, as shown in Fig. 2a, summarizing the voltage conditions applied to selected and unselected WLs and BLs. During PGM and ERS operations, only the selected WL receives the full PGM or ERS voltage, while all unselected WLs and BLs are biased at pass voltage (V_{pass}) that maintains the local electric field below the threshold required for carrier generation. This biasing strategy effectively suppresses unintended channel formation and prevents disturbance of unselected cells. When a positive V_{PGM} is applied to the selected WL, a field-induced electron channel forms above the N^+ retrograde-doped layer, as shown in Fig. 2b. This channel provides a conduction path for electron injection



		PGM	ERS	Read
Sel. BL	Sel. WL	V_{PGM}	V_{ERS}	V_c -> V_d
	Unsel. WL	V_{pass}	V_{epass}	GND
	BL (N+)	GND	Float	GND
	P+	Float	GND	Float
Unsel. BL	Sel. WL	V_{PGM}	V_{ERS}	GND
	Unsel. WL	V_{pass}	V_{epass}	GND
	BL (N+)	V_{pass}	Float	GND
	P+	Float	V_{epass}	Float

Fig. 2 Biasing scheme and selective programming/erasing mechanism of the memcapacitor array. **a** Biasing conditions applied to selected and unselected WLs and BLs during PGM, ERS, and read operations. Only the selected WL receives the full PGM or ERS voltage, while unselected WLs and BLs are biased at pass voltages that keep the local electric field below the threshold for carrier generation, ensuring cell-level selectivity. **b** Programming operation, where a positive V_{PGM} applied to the selected WL induces an electron channel above the N^+ retrograde-doped region, enabling electron injection into the charge-trap layer exclusively at the selected WL–BL intersection. **c** Erasing operation, where a negative erase voltage V_{ERS} applied to the selected WL induces a hole channel, facilitating hole-assisted detrapping of stored charge from the charge-trap layer. Unselected BLs remain biased at erase-pass voltages V_{epass}, suppressing channel formation and preventing unintended ERS events

into the charge-trap layer, driving the device toward the PGM state. Because only the selected BL experiences the appropriate vertical electric field, electron channel formation and charge injection occur exclusively at the target cell. Similarly, during the erase operation, applying a negative V_{ERS} to the selected WL induces a hole channel, as shown in Fig. 2c, enabling charge removal from the trap layer. Unselected BLs remain biased at erase-pass voltage (V_{epass}), which suppresses hole channel formation and confines hole-assisted detrapping to the selected cell. This selective formation of electron or hole channels during PGM and ERS operations ensures robust cell-level selectivity, enabling reliable weight updates without disturbing neighboring cells.

3 Results and discussion

3.1 Fabrication and measurement methodology

Figure 3a outlines the overall fabrication process flow of the proposed memcapacitor array, starting from the formation of the buried oxide layer. The gate stack consists of a retrograde poly-Si channel layer, a $SiO_2/Si_3N_4/Al_2O_3$ dielectric stack, and a TiN WL electrode, forming a TANOS ($TiN/Al_2O_3/Si_3N_4/SiO_2/Si$) charge-trap memory structure. Following deposition of the dielectric layers, TiN is deposited by metal–organic chemical vapor deposition (MOCVD) and subsequently patterned to define the WLs that form each column of the array. The fabricated memcapacitor arrays were implemented in two layouts, 12 × 24 and 24 × 48. A continuous highly doped N^+ region was formed beneath the lightly doped active poly-Si layer to provide a low-resistance lateral readout path while preserving the lightly doped region required for depletion-based capacitance modulation. BL regions were formed in doped poly-Si through in-situ dopant incorporation followed by activation annealing. This separation of the readout path from the active depletion layer alleviates the resistance–depletion trade-off present in the previous architecture [44]. This dopant gas modulation technique enables the formation of a sharp retrograde doping profile, which is difficult to achieve using conventional ion implantation and thermal diffusion. A comparison between the two approaches, including secondary ion mass spectrometry (SIMS) results, is provided in Fig. S2. A schematic illustration of the key fabrication steps is shown in Fig. 3b, highlighting the sequential construction of alternating WL and BL structures to realize the crossbar array. Poly-Si mesas are first defined, followed by conformal deposition of the O/N/A dielectric stack and the TiN WL electrode. Photolithography and wet etching are then used to pattern the WLs, while BLs are accessed through heavily doped poly-Si contact regions. To establish appropriate carrier reservoirs and ensure reliable operation, N^+ regions are formed at the bottom of the channel by phosphorus implantation, serving as electron sources during PGM, while P^+ regions are introduced via boron implantation to supply minority carriers during ERS operations.

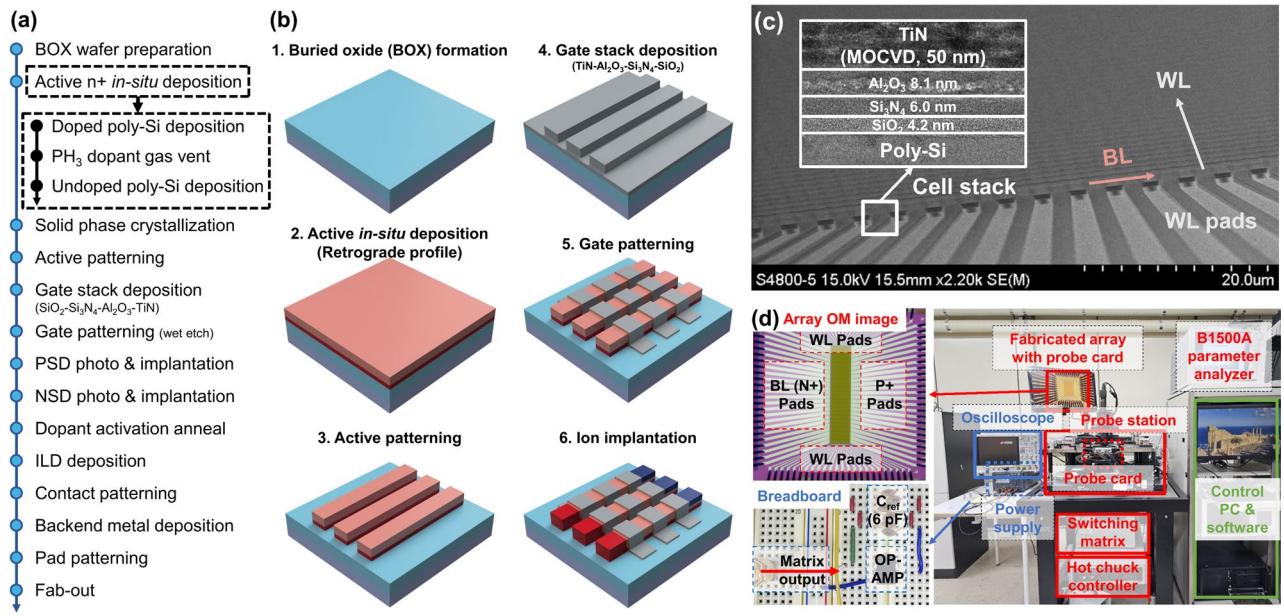


Fig. 3 Fabrication process and characterization setup of the CTF-based memcapacitor array. **a** Overall fabrication process flow of the CTF–based memcapacitor array, starting from buried oxide formation, followed by retrograde-doped poly-Si channel formation, deposition of the A/N/O charge-trap dielectric stack, and TiN WL electrode patterning to define the crossbar structure. **b** Schematic illustration of the key fabrication steps, highlighting the sequential construction of alternating WL and BL structures to realize the planar memcapacitor array, including poly-Si mesa definition, dielectric and TiN deposition, and contact formation. **c** Scanning electron microscopy (SEM) image of the completed 24 × 48 memcapacitor array, showing the periodic arrangement of WLs and BLs with uniform line widths and excellent layer alignment. The inset presents a magnified view of a single memcapacitor gate stack, confirming the designed layer thicknesses. **d** Electrical characterization setup used for device- and circuit-level measurements. A multi-pin probe card interfaces with all WL and BL pads, while a switching matrix selectively connects the array to a semiconductor parameter analyzer for device characterization or to an external sensing circuit for displacement-charge measurements

Figure 3c presents a scanning electron microscopy (SEM) image of the completed 24 × 48 memcapacitor array. The periodic arrangement of WLs and BLs is clearly observed, demonstrating uniform line widths, a gate spacing of 500 nm, and excellent layer alignment across the array. The inset shows a magnified view of a single memcapacitor gate stack, confirming the designed layer thicknesses of TiN (50 nm), Al_2O_3 (8.1 nm), Si_3N_4 (6.0 nm), SiO_2 (4.2 nm), and poly-Si. Electrical characterization was performed using the measurement setup shown in Fig. 3d. A multi-pin probe card simultaneously interfaces with all WL and BL pads, while a switching matrix selectively routes signals to a semiconductor parameter analyzer (Keysight B1500A) for device-level characterization. For circuit-level measurements, the array output is connected to an external breadboard circuit comprising an operational amplifier, a reference capacitor, and an oscilloscope. This configuration enables direct measurement of displacement charge generated during capacitive readout, allowing accurate evaluation of the circuit output signal.

3.2 Electrical characteristics

The *C*–*V* characteristics measured from all pristine cells, shown in Fig. 4a, exhibit a typical n-type MOSCAP behavior, with maximum capacitance in accumulation followed by a gradual reduction toward depletion. The extremely small variation across the entire array ($\sigma/\mu < 0.37\%$) confirms excellent device uniformity and demonstrates the effectiveness of the N^+ bottom readout structure in suppressing lateral voltage degradation while preserving the depletion-based capacitance window of the lightly doped active region. Analog weight modulation was also demonstrated using incremental step-pulse programming (ISPP), as shown in Fig. 4b. The application of 41 programming pulses (100 µs, 8–12 V with 0.1 V increments) induces a gradual and monotonic shift in the *C*–*V* curve, indicating stable and well-controlled potentiation behavior. The corresponding flat-band voltage shifts are summarized in Fig. 4c, where the combined use of ISPP and incremental step-pulse erase (ISPE) enables symmetric and continuous bidirectional tuning through electron trapping and hole-assisted detrapping. In addition, inhibit operations were verified by monitoring the output displacement charge Q during ISPP and ISPE. As demonstrated in Fig. 4d, only the selected cell exhibits a change in Q, while all neighboring cells remain unchanged, confirming that the pass-bias scheme effectively suppresses channel formation in unselected bit lines and prevents unintended programming or erase events.

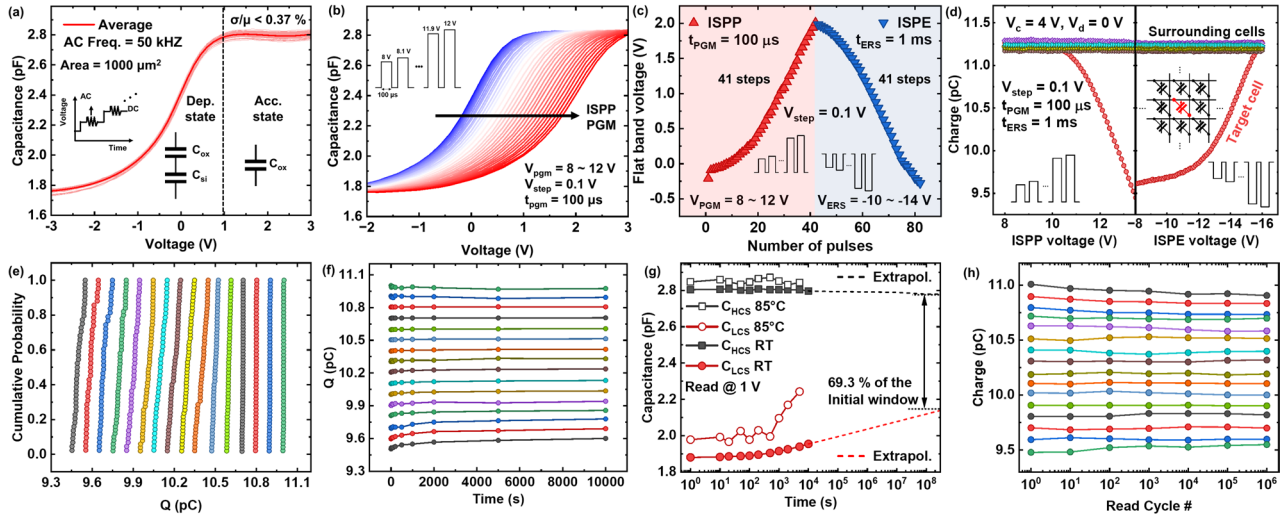


Fig. 4 Analog weight modulation, uniformity, and reliability of the memcapacitor array. **a***C–V* characteristics measured from all pristine cells in the array, exhibiting a typical n-type MOS capacitor behavior with minimal device-to-device variation with σ/μ < 0.37%, confirming excellent array-level uniformity. **b** ISPP–induced evolution of the *C–V* characteristics, showing continuous and monotonic modulation of capacitance with increasing programming pulses. **c** Corresponding flat-band voltage shifts under ISPP and ISPE, demonstrating symmetric and bidirectional analog weight tuning enabled by electron trapping and hole-assisted detrapping. **d** Verification of inhibit operation during ISPP/ISPE, where only the selected cell shows a change in output displacement charge *Q*, while neighboring cells remain unaffected, confirming effective suppression of unintended programming and erasing events. **e** Cumulative distributions of 16 discrete analog weight states (4-bit resolution) generated using closed-loop programming, showing well-separated and highly uniform charge levels across the array. **f** Retention characteristics of the programmed analog states, demonstrating stable separation of all 16 levels over time with minimal drift. **g** Retention behavior of the HCS and LCS measured at room temperature and 85 °C, confirming stable memory window preservation under elevated temperature conditions. **h** Read endurance characteristics evaluated through repeated transient *I–t* measurements, showing no measurable degradation even after 10^6 read cycles, indicating negligible electrical stress during capacitive readout

Fine-grained analog weight generation was achieved by tuning 16 discrete levels (4-bit resolution) from 9.5 pC to 11.0 pC in 0.1 pC increments using a closed-loop programming scheme. The cumulative distributions shown in Fig. 4e demonstrate that all 16 states are well separated and highly uniform across the array. The closed-loop programming required an average of 45 program/verify iterations to reach the target state, with a maximum of 80 iterations. The final programming error exhibited a mean absolute error of 0.0178 pC and a maximum absolute error of 0.0500 pC, with 95% of the absolute errors below 0.0465 pC, as further summarized in Fig. S3. Retention characteristics of the programmed states are presented in Fig. 4f, where all levels remain clearly distinguishable over 10^4 s, with only minor drift attributed to gradual electron or hole detrapping. Additional retention tests for the highest-capacitance (HCS) and lowest-capacitance (LCS) states, conducted at room temperature and 85 °C, further confirm stable preservation of the memory window with predictable temperature-dependent degradation, as shown in Fig. 4g. The robustness of the capacitive read operation was evaluated through repeated transient *I–t* measurements, as shown in Fig. 4h. It confirms that no measurable drift is observed even after 10^6 transient read cycles, verifying that the capacitive readout mechanism introduces negligible electrical stress on the dielectric stack. In addition, it is important to note that the presented planar devices represent a conservative case in terms of charge modulation capability. In vertically integrated structures, the A/N/O stack is formed prior to silicon deposition, which effectively suppresses dopant diffusion within the silicon channel during subsequent processing. As shown in the SIMS profile in Fig. S2c, this enables better preservation of the engineered retrograde silicon distribution. The impact of this profile on device operation is further reflected in the TCAD-simulated *C-V* characteristics and the extracted MW_Q in Fig. S2d and e, indicating that improved profile control can lead to enhanced charge trapping efficiency and a larger dynamic range in vertically stacked implementations.

The functional capability of the fabricated memcapacitor array was validated through array-level demonstrations, hardware-based readout measurements, and energy benchmarking, as summarized in Fig. 5. To evaluate random-access and multi-level write operations at the array scale, a 4-bit grayscale image was mapped onto the 24 × 48 array, with each pixel intensity encoded into one of 16 analog weight states. As shown in Fig. 5a, the image is accurately programmed in the fabricated array, confirming uniform write–verify operations across the array as well as reliable bit-alterable programming without inducing disturbances in neighboring cells. To further verify VMM functionality, weighted-sum operations were performed using the programmed weights. The

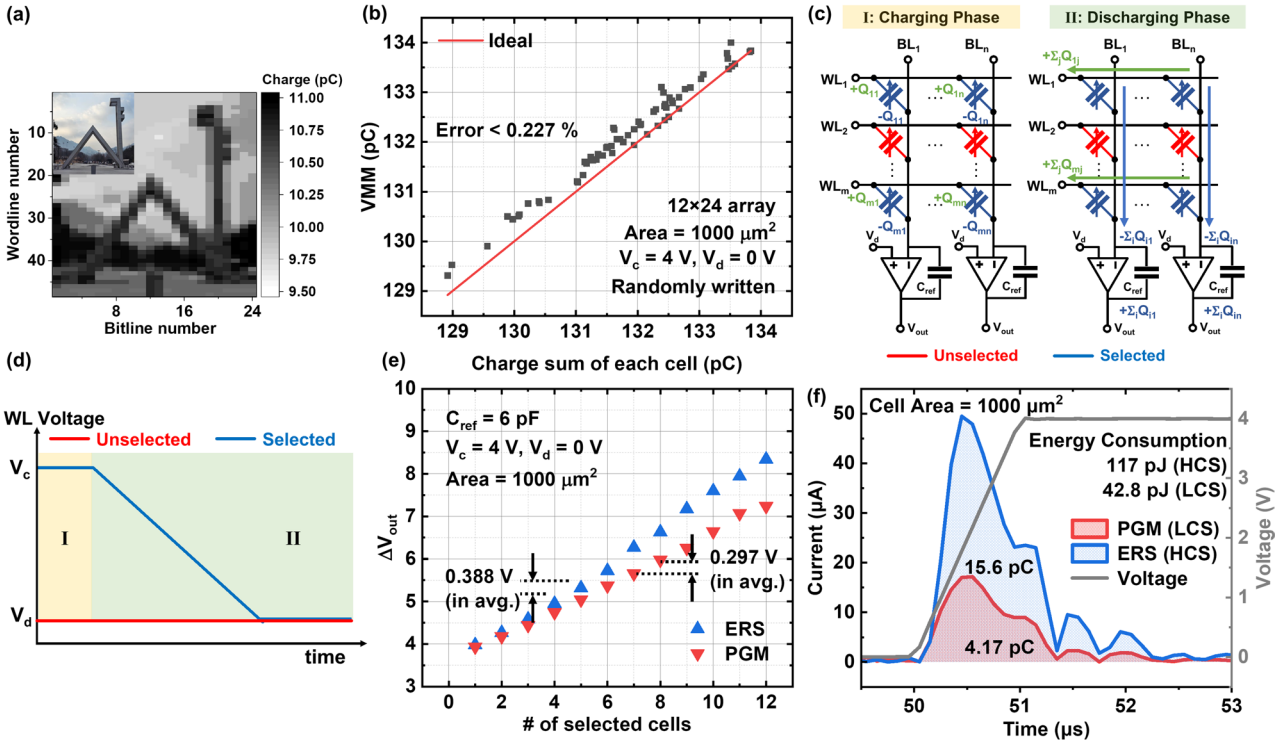


Fig. 5 Array-level vector–matrix multiplication and energy-efficient capacitive readout. **a** Array-level demonstration of random-access, multi-level programming using a 4-bit grayscale image mapped onto a 24 × 48 memcapacitor array, confirming uniform write–verify convergence and disturbance-free bit-alterable operation across the array. **b** Verification of VMM operation through charge accumulation, showing that the simultaneously measured output charge closely follows the ideal linear sum of individually measured cell charges. **c** Schematic of the capacitive sensing circuit used for array readout, where displacement charge generated during voltage transitions is transferred to a reference capacitor (C_{ref}) and converted into an output voltage by an operational amplifier. **d** Applied voltage pulse waveforms for selected and unselected cells during capacitive readout, illustrating selective charge generation only in the addressed cells. **e** Measured output voltage as a function of the number of simultaneously activated cells for the HCS (ERS) and LCS (PGM), confirming linear scaling of the sensing circuit response. **f** Extracted read energy per operation obtained by integrating the transient current over the voltage-transition window, showing distinct energy levels for HCS and LCS states and enabling energy scaling analysis. **g** Comparison of projected read energy with previously reported synaptic devices

displacement charge of each cell was first measured individually, and the total output charge was then measured when multiple selected cells were read simultaneously. As shown in Fig. 5b, the measured total charge closely follows the ideal linear sum of the individually measured values, demonstrating accurate VMM operation and confirming that charge accumulation scales linearly with the number of contributing cells. The operation of the sensing circuit used for these measurements is illustrated in Fig. 5c. During the charging phase, the selected cell experiences a potential difference across its electrodes and stores charge according to its programmed capacitance state, while unselected cells are symmetrically biased to suppress displacement current generation. In the subsequent discharging phase, both electrodes of the selected cell transition to the discharge voltage V_d, forcing the stored charge to transfer into the reference capacitor (C_{ref}) connected to an operational amplifier (OP-AMP). The applied pulse waveforms for selected and unselected cells are shown in Fig. 5d, illustrating the voltage transitions required for selective operation.

As the number of simultaneously activated cells increases, the output voltage of the sensing circuit increases proportionally. Figure 5e confirms this linear relationship for both the HCS (ERS) and LCS (PGM). The steeper slope observed for ERS cells originates from their larger stored charge, validating the circuit function to resolve multi-level analog weights with high sensitivity. The energy consumption during read operations was extracted as shown in Fig. 5f. By integrating the transient current over the 1-μs voltage-transition window, read energies of approximately 117 pJ for the HCS and 42.8 pJ for the LCS were obtained for the fabricated devices. Geometric scaling of the measured 1000 μm^2 cells to a 0.01 μm^2 footprint gives an intrinsic cell-level read-energy projection of 0.428–1.17 fJ. This projection represents a lower-bound estimate of the read energy based on the intrinsic cell capacitance and the CV^2 relation, excluding fringe-field and interconnect capacitances, as well as energy consumed by the driver, sensing circuitry, switches, and control circuitry. Figure S4 presents the sensitivity of the projected energy to parasitic

capacitance, where $\alpha = C_{par}/C_{cell}$ is varied from 0 to 10 to account for increasing capacitive loading. The external OP-AMP circuit was used for functional readout validation rather than as an optimized integrated peripheral. In addition, the present device operation employs 8–12 V programming pulses and a 4 V read swing, which require appropriate high-voltage driving circuitry and can introduce additional switching-energy, area, and latency overheads. These peripheral contributions are not included in the intrinsic cell-level read-energy projection discussed above.

3.3 TCAD verification of vertically stacked devices and scaling-induced verification error suppression

The memcapacitor structure demonstrated in planar form can be naturally extended into a vertically stacked configuration to achieve ultrahigh integration density by leveraging the fabrication principles of commercial 3D NAND flash technology. As illustrated in Fig. S5, the vertical device follows a 3D NAND-derived process flow that includes alternating oxide–nitride mold deposition, channel-hole etching, sidewall formation of the A/N/O charge-trap stack, and in-situ polysilicon deposition to form the vertical channel body and N^+ bit-line core. Subsequent WL metal filling and patterning complete the three-dimensional structure. This process flow is fully compatible with existing high-volume 3D NAND manufacturing and enables memcapacitors to be stacked over hundreds of layers while preserving the compact 4 F^2 footprint. To evaluate the feasibility of PGM, ERS, and read operations in scaled vertical geometries, full three-dimensional TCAD simulations were performed using the parameters shown in Fig. 6a. The simulated device structure reproduces the cylindrical channel geometry and multilayer WL configuration characteristic of 3D NAND-like arrays. The C–V characteristics extracted from multiple WL positions exhibit negligible layer-to-layer variation, as shown in Fig. 6b, indicating that the electrostatic behavior remains uniform across vertically stacked cells. Such uniformity is essential for realizing large-capacity synaptic memory with consistent analog behavior. Carrier transport during PGM and ERS operations remains strictly confined to the selected WL plane, as shown in Fig. 6c. Electron trapping and hole-assisted detrapping occur exclusively at the targeted WL, while unselected layers show no appreciable charge accumulation. Figure 6d further confirms that the trapped charge is tightly localized to the selected cell, with adjacent cells exhibiting no measurable charge trapping. These results demonstrate that the proposed memcapacitor retains robust cell-level selectivity and electrostatic isolation even in deeply scaled, vertically stacked architectures, consistent with the behavior observed in planar devices.

While PGM and ERS operations scale effectively, VMM in a scaled vertical array introduces an additional challenge. As the gate length (L_{gate}) and intercell spacing (L_{ITC}) are reduced to dimensions comparable to modern 3D NAND technology, electrostatic coupling between neighboring cells becomes increasingly significant. During verification, only a single WL is biased for readout, whereas during inference all WLs simultaneously receive the input voltage. This difference in biasing conditions leads to a mismatch in the electrostatic potential of the intercell silicon region between verification and inference, referred to here as a verification error. As the number of simultaneously activated cells increases, the resulting change in intercell potential can modify the fraction of the channel that enters accumulation beneath the selected WL. This effect causes a discrepancy between the individually verified cell weight and the effective weight during parallel VMM accumulation. As shown in Fig. 7a, although the overall VMM response remains approximately linear, the deviation from the ideal linear relation increases with the number of activated cells, confirming the emergence of inference errors associated with verification error. The underlying mechanism is illustrated in Fig. 7b. During verification, the selected cell experiences a slightly expanded effective capacitance area, whereas during inference the accumulation region is more uniformly distributed along the channel string. Consequently, verification overestimates the effective capacitance per cell and thus the summed displacement charge.

To suppress this scaling-induced error, an intercell trapped-charge (ITC) method can be introduced, which stabilizes the electrostatic potential of the intercell silicon region and enforces depletion during both verification and inference. As shown in Fig. 7c, this method employs a single global initialization programming pulse applied to all WLs prior to weight programming, uniformly trapping electrons in the intercell regions and fixing their electrostatic state. ITC initialization is performed before weight programming and therefore adds no per-inference read-energy overhead. However, long-term drift beyond the measured time window may require occasional reinitialization followed by weight remapping. Subsequent ERS operations modify only the selected WL without disturbing the intercell trapped charge, as confirmed in Figs. 7d and e. Because the intercell region remains depleted irrespective of variations in the intercell potential V_0 between verification and inference, the effective capacitance area becomes identical in both phases, thereby reducing the source of verification error. This behavior is further visualized through electron density analysis in Fig. S6. Quantitatively, the ITC method reduces the VMM inference error by 84.6% at L_{ITC} = 30 nm, as summarized in Fig. 7f. Under these conditions,

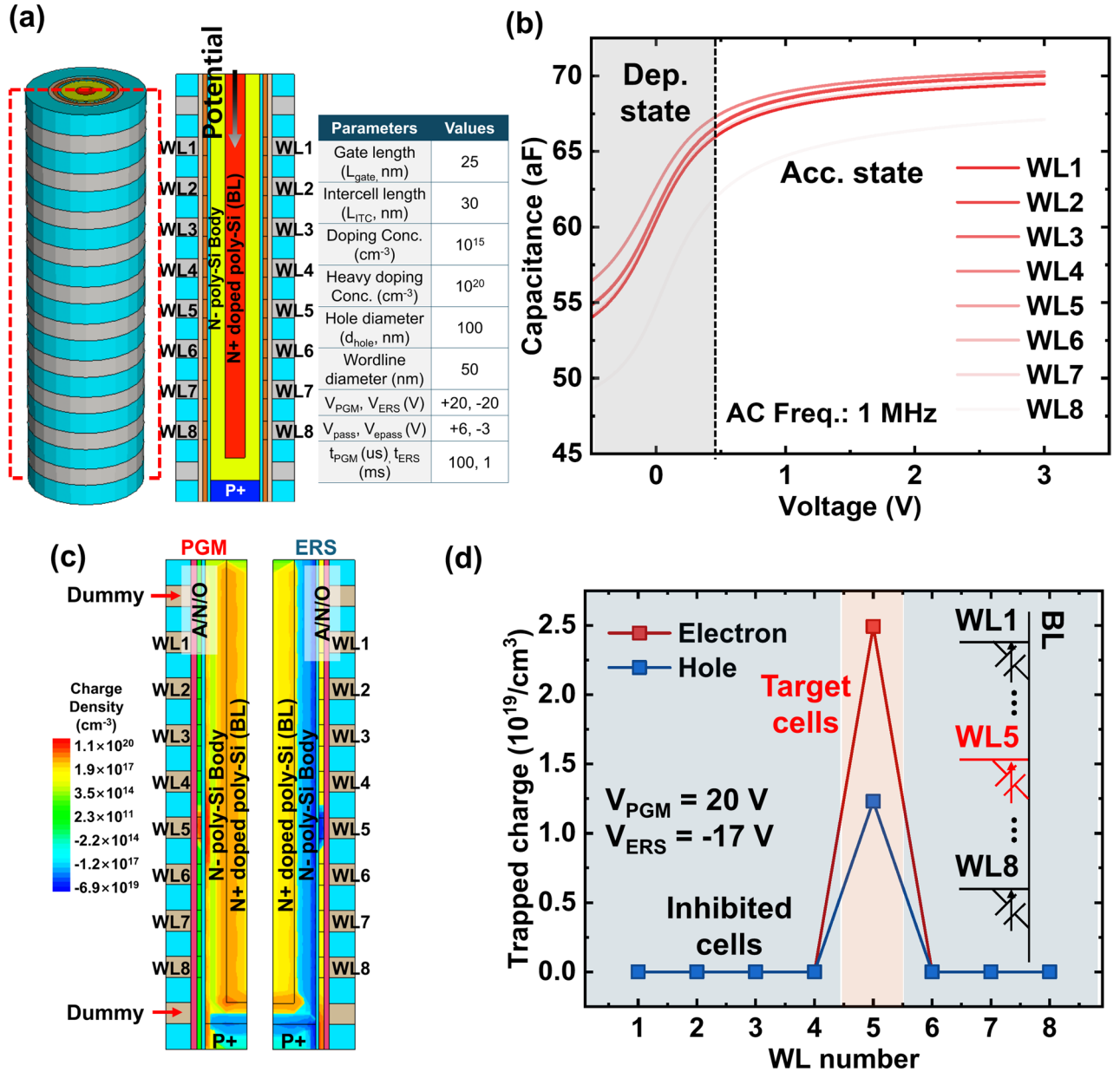


Fig. 6 Vertically stacked memcapacitor structure and cell-level electrostatic behavior. **a** Three-dimensional TCAD simulation model of the vertically stacked memcapacitor array, showing a cylindrical poly-Si channel surrounded by multiple WLs and a charge-trap dielectric stack compatible with 3D NAND architectures. **b** Simulated *C–V* characteristics extracted from multiple WL layers along the vertical stack, demonstrating negligible layer-to-layer variation and uniform electrostatic behavior across vertically integrated cells. **c** Spatial distribution of trapped charge during PGM and ERS operations, confirming that carrier trapping and detrapping are confined to the selected WL plane. **d** Enlarged view of the trapped-charge profile around the selected cell, showing strong localization of charge and the absence of disturbance in adjacent WLs, consistent with robust cell-level selectivity in the vertically stacked configuration

the VMM error decreases from 10.58% for the conventional scheme to 1.63% with ITC, restoring near-ideal VMM accuracy while maintaining compatibility with scaled 3D NAND process dimensions.

Although the trapped-charge density in the intercell nitride region was not directly monitored, the functional stability of ITC-assisted error suppression was experimentally evaluated using the fabricated structure with L_{ITC} = 500 nm. As shown in Fig. S7, the area-normalized RMSE of the ITC-initialized scheme remained between 0.00263 and 0.00279 fC/μm^2 over 1-10^4 s, whereas the conventional RMSE remained approximately 0.012 fC/μm^2. The corresponding RMSE reduction remained between 76.7% and 78.2%. The 78.2% reduction measured at 1 s agrees well with the approximately 78.1% suppression observed for the 500-nm experimental point

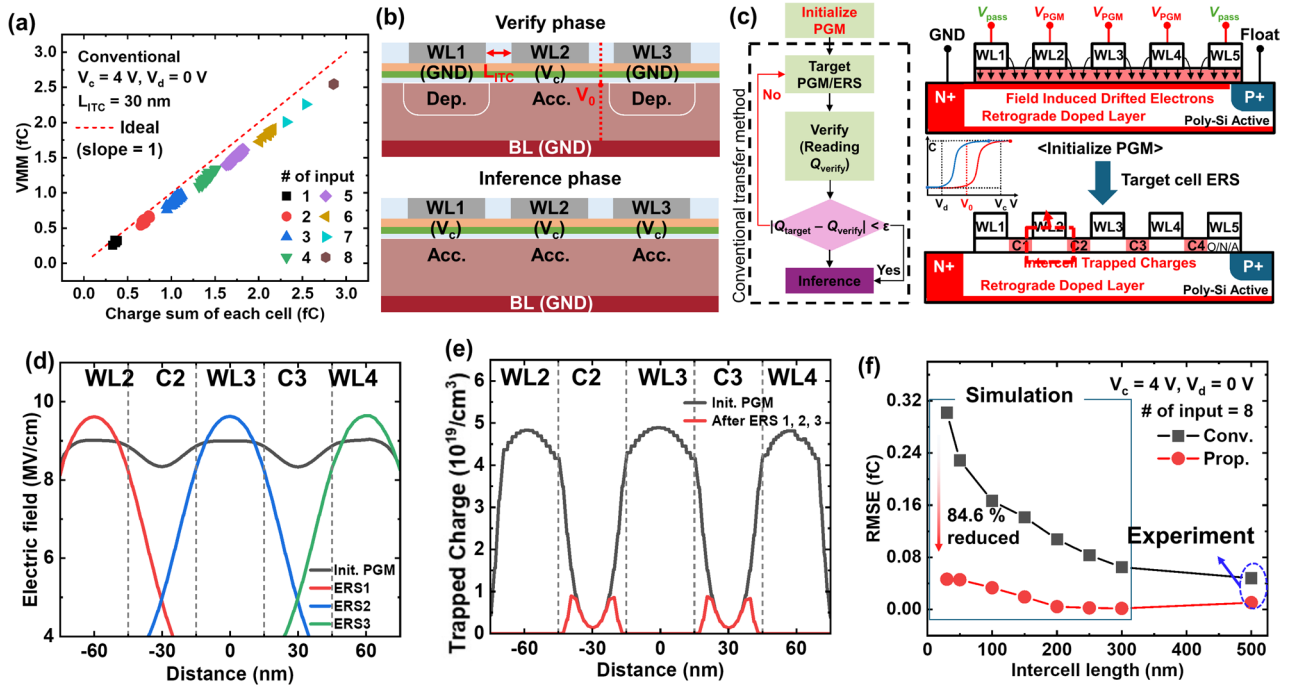


Fig. 7 Verification error in scaled vertical arrays and suppression via ITC method. **a** Simulated VMM characteristics in a scaled vertical array, showing increasing deviation from ideal linear accumulation as the number of simultaneously activated cells increases due to verification error. **b** Schematic illustration of the verification error mechanism, highlighting differences in accumulation area and effective capacitance between verification and inference phases caused by variations in intercell electrostatic potential. **c** Conceptual diagram of the ITC method, in which a global initialization pulse traps electrons in intercell regions to stabilize the electrostatic potential. **d** Simulated trapped-charge distribution after ITC initialization, showing uniform electron trapping in intercell regions while preserving programmability of the selected WL. **e** Trapped-charge profiles after subsequent erase operations, confirming that intercell trapped charge remains intact and undisturbed during weight updates. **f** Quantitative comparison of VMM inference error with and without the ITC method as a function of L_{ITC}, demonstrating an 84.6% reduction in inference error at L_{ITC} = 30 nm and near-ideal VMM accuracy under scaled 3D NAND-compatible dimensions

in Fig. 7f. No appreciable degradation of the ITC-assisted error suppression was observed within the measured 10^4-s window. Because local intercell fields depend on the surrounding weight pattern, pattern-dependent stability beyond the tested condition remains to be characterized. Cycling endurance specific to repeated ITC initialization was not directly evaluated. Because the ITC state is established by electron trapping and is not intentionally erased during normal weight updates, the intercell region is expected to undergo less frequent cycling stress than active cells subjected to repeated hole-assisted erase operations. Nevertheless, high-field electron injection can also generate oxide or interface traps, and direct cycling measurements are required for quantitative lifetime assessment. If substantial long-term drift occurs beyond the measured window, ITC reinitialization followed by weight remapping could be employed to restore the initialized state.

3.4 Hybrid hardware-software spiking neural network demonstration

The inference evaluation used a hybrid hardware-software approach in which the convolutional and preceding fully connected layers were executed in software, whereas only the final 24 × 10 fully connected layer was mapped to the measured memcapacitor arrays. Accordingly, the present implementation should be regarded as a final-layer hardware-mapped proof of concept rather than an end-to-end hardware SNN demonstration. Beyond device compatibility, SNNs are inherently well aligned with energy-efficient computing, as sparse spike-based communication naturally reduces switching activity and energy consumption. Because the memcapacitor generates displacement current only during voltage transitions, its transient charge-based operation is naturally matched to the event-driven dynamics of SNNs. In contrast to conventional artificial neural networks (ANNs), where inputs are encoded in amplitude and the limited read window of the memcapacitor can constrain the effective dynamic range, SNNs accumulate output charge over multiple spikes, enabling temporal integration of displacement charge and finer spike-count resolution while leaving the physical cell-weight window unchanged.

To generate hardware-compatible weights, a standard ANN-to-SNN conversion was applied, as shown in Fig. 8a, in which each layer's weights were normalized

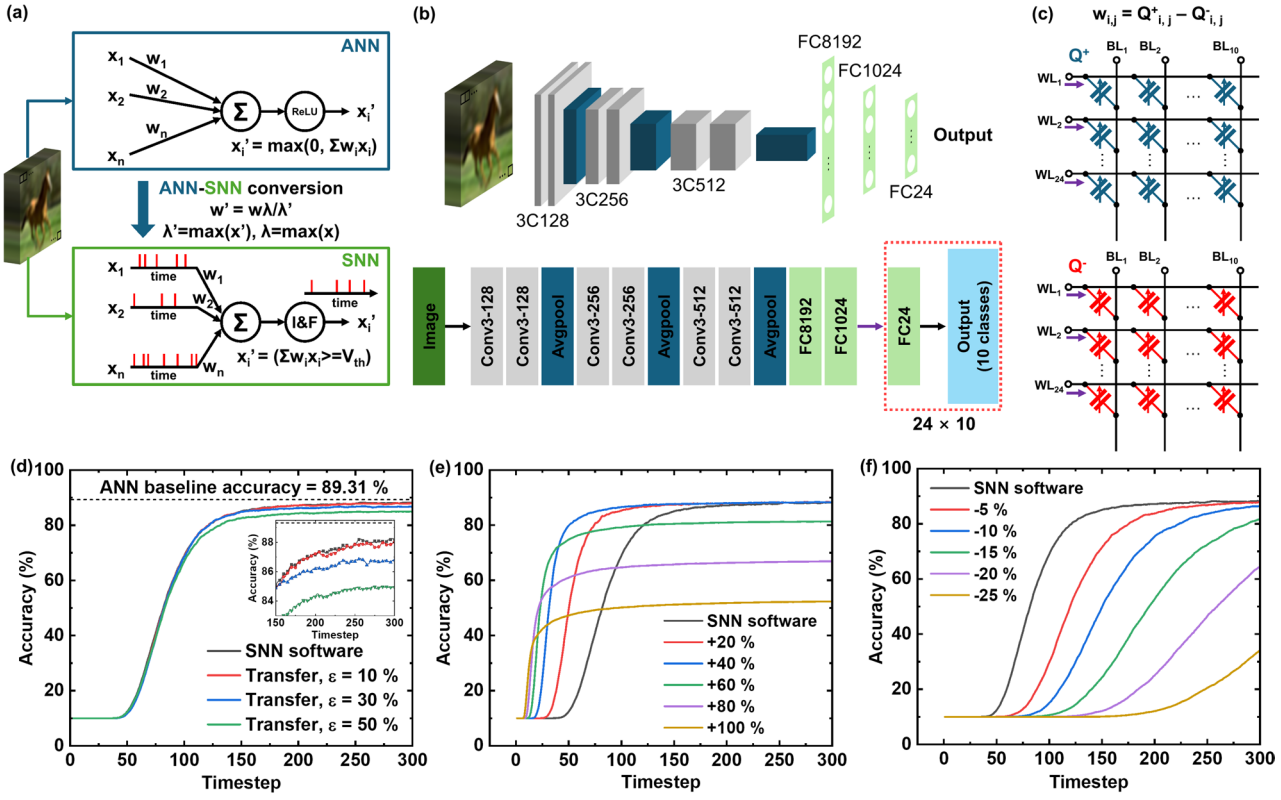


Fig. 8 Hybrid hardware–software spiking neural network inference using the memcapacitor array. **a** Schematic of the ANN–to–SNN conversion process used to generate hardware-compatible weights, where layer-wise normalization enables mapping of ANN-trained weights to spike-based inference. **b** Modified VGG-8 network architecture employed for CIFAR-10 classification, used as the baseline model for the hybrid evaluation. **c** Hardware implementation of signed synaptic weights, where each weight w_{ij} is encoded by the differential displacement charge $Q^{+}_{i,j} - Q^{-}_{i,j}$ using two 24×10 memcapacitor arrays storing the positive and negative weight components. **d** Classification accuracy of the hardware-mapped final-layer SNN as a function of the allowable weight-transfer error margin, showing software-comparable performance at a 10% margin and graceful degradation under looser constraints. **e** Impact of positive VMM errors on SNN dynamics, where over-accumulated charge leads to premature neuron activation and faster but erroneous convergence. **f** Impact of negative VMM errors on SNN dynamics, showing delayed spike generation and suppressed neuron activity, which can be partially compensated by extended inference time or weight renormalization

by the maximum ANN activation [51–54]. The network was trained on the CIFAR-10 dataset using a modified VGG-8 architecture, as shown in Fig. 8b, achieving a classification accuracy of 89.31%, as shown in Fig. S8. After conversion, a 24 × 10 weight matrix corresponding to the final fully connected layer was extracted, normalized, and transferred to the hardware within a 10% error margin, as shown in Fig. S9. Closed-loop ISPP/ISPE programming was used to implement these weights in the memcapacitor array. As illustrated in Fig. 8c, the signed synaptic weight w_{ij} is encoded by the charge difference $Q^{+}_{i,j} - Q^{-}_{i,j}$. Two 24 × 10 arrays were employed to store the positive (Q^{+}) and negative (Q^{-}) weight components, while an alternative single-array differential implementation using opposite dV/dt slopes is presented in Fig. S10.

Figure 8d summarizes the hybrid hardware–software inference results under different allowable weight-transfer error margins. With a 10% margin, the hardware-mapped SNN achieves an accuracy of 88.01%, which is 0.16% points below the software SNN accuracy of 88.17% and 1.30% points below the ANN baseline of 89.31%. Even with relaxed constraints of 30% and 50%, accuracies of 86.79% and 84.92% are maintained, demonstrating strong robustness against write-induced weight variations. The gradual degradation observed at larger error margins reflects a trade-off between programming precision, write cycles, and inference accuracy. To further investigate the inference impact of device-level VMM inaccuracies, artificial weight perturbations were injected with either positive or negative polarity. As shown in Fig. 8e, positive VMM errors (+20 to +100%) lead to over-accumulated charge, causing premature neuron activation and faster but erroneous convergence. In contrast, as shown in Fig. 8f, negative VMM errors reduce the integrated charge, delaying spike generation and suppressing neuron activity. While this behavior lowers accuracy at early timesteps, it can be mitigated by extending the inference window or renormalizing weights. Notably, the VMM error originating from vertical-structure scaling, discussed in Sect. 3.3, inherently

results in a negative-direction error, indicating that the temporal dynamics of SNN inference can partially compensate for its impact. For T timesteps, the array-access latency increases approximately in proportion to T, while the dynamic cell energy scales with the total number of active spike-driven read operations and therefore depends on both spike activity and timestep count. Accordingly, 300 timesteps represents an accuracy-oriented operating point rather than one optimized for minimum latency or energy. These results demonstrate that the measured array enables accurate hardware mapping of the final SNN layer, whereas the effects of vertical scaling are assessed separately through TCAD simulations. Although this work focuses on SNN inference as a proof of concept, the demonstrated capacitive VMM provides a basis for future extension to larger feed-forward layers, where dense weight storage and efficient accumulation are important.

4 Conclusions

In this work, we have demonstrated a vertically stackable CTF–based memcapacitor array that combines the ultra-high integration density of 3D NAND technology with the intrinsic energy efficiency of capacitive readout. By encoding synaptic weights in capacitance rather than conductance, the proposed device avoids a sustained conductive read path through the memory element, while practical array energy still includes transient charging and peripheral overhead. The fabricated memcapacitor array exhibits uniform analog characteristics, stable multibit weight modulation, long-term retention, and robust read endurance, validating the reliability of the capacitive readout mechanism at the array level. Accurate vector–matrix multiplication is achieved through linear charge accumulation, and energy analysis based on the fundamental CV^2 relation projects intrinsic cell-level read energy in the femtojoule range at scaled dimensions, while practical array energy depends on parasitic capacitance and peripheral circuitry. Importantly, potential scaling-induced inference errors arising from intercell electrostatic coupling in vertically stacked architectures are effectively suppressed using the proposed ITC method, enabling near-ideal VMM accuracy under scaled geometries. The functional error suppression provided by ITC remains stable over 10^4 s in the measured 500-nm-spacing structure, while cycling endurance specific to repeated ITC initialization remains to be established. A hybrid hardware–software SNN evaluation, in which only the final 24×10 layer is mapped to measured arrays, achieves software-comparable accuracy. The transient charge-based operation is compatible with spike accumulation, although longer inference windows increase latency and access energy. We believe these results establish vertically stackable memcapacitors as a scalable and energy-efficient synaptic device platform for memory-centric computing. By utilizing the manufacturing maturity of 3D NAND and the favorable energy scaling of capacitive operation, the proposed architecture offers a practical route toward high-density, low-energy neural computing hardware.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40580-026-00575-9.

Supplementary Material 1.

Acknowledgements
Not applicable.

Author contributions
All authors read and approved of the final manuscript.

Funding
This work was supported by the NRF grant funded by Korean government (RS-2023-00260527, 10%), in part by the IITP grant (RS-2021-II212052, 10%, RS-2024-00506767, 20%), in part by KIAT grant funded by the Korean government (RS-2024-00410066, HRD Program for Industrial Innovation, 10%), and in part by the MSIT under grant RS-2026-25536380 (Intelligent Semiconductor Technology Development Program, 50%). The EDA tool was supported by the IC Design Education Center (IDEC), Korea.

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

Declarations

Competing interests
The authors declare no competing financial interests.

Author details
[1]Department of Electrical and Computer Engineering and Inter-University Semiconductor Research Center (ISRC), Seoul National University, Seoul 08826, Republic of Korea
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[3]Division of Materials Science and Engineering, Department of Semiconductor Engineering, Hanyang University, Seoul 04763, Republic of Korea

Received: 31 May 2026 / Accepted: 29 August 2026

References

1. A. Gholami, Z. Yao, S. Kim, C. Hooper, M.W. Mahoney, K. Keutzer, AI and memory wall. IEEE Micro. **44**(3), 33–39 (2024). https://doi.org/10.1109/MM.2024.3373763
2. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature. **521**(7553), 436–444 (2015)
3. T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal et al., Language models are few-shot learners. Adv. Neural Inf. Process. Syst. **33**, 1877–1901 (2020)
4. J. Backus, Can programming be liberated from the von Neumann style? a functional style and its algebra of programs. Commun. ACM. **21**(8), 613–641 (1978). https://doi.org/10.1145/359576.359579
5. N. Talati, R. Ben-Hur, N. Wald, A. Haj-Ali, J. Reuben, S. Kvatinsky, mMPU—a real processing-in-memory architecture to combat the von Neumann Bottleneck.

In: Applications of Emerging Memory Technology: Beyond Storage, ed. by M. Suri (Springer Singapore, Singapore, 2020), pp. 191–213
6. K. Kim, M.S. Song, H. Hwang, S. Hwang, H. Kim, A comprehensive review of advanced trends: from artificial synapses to neuromorphic systems with consideration of non-ideal effects. Front. Neurosci. (2024). https://doi.org/10.3389/fnins.2024.1279708
7. S. Yu, Neuro-inspired computing with emerging nonvolatile memorys. Proc. IEEE. 2018;106(2):260–85. https://doi.org/10.1109/JPROC.2018.2790840
8. S.B. Eryilmaz, D. Kuzum, S. Yu, H.S.P. Wong, Device and system level design considerations for analog-non-volatile-memory based neuromorphic architectures. In: Proc. IEEE International Electron Devices Meeting (IEDM). 2015. p. 4.1.-4.1.4
9. S. Youn, J. Lee, S. Kim, J. Park, K. Kim, H. Kim, Programmable threshold logic implementations in a memristor crossbar array. Nano Lett. **24**(12), 3581–3589 (2024)
10. G. Kim, S. Youn, J. Park, H. Kim, Configurable kernel map implementation in memristor crossbar for convolution neural network. Adv. Intell. Syst. **7**(11), 2500032 (2025). https://doi.org/10.1002/aisy.202500032
11. D. Ielmini, S. Ambrogio, Emerging neuromorphic devices. Nanotechnology. **31**(9), 092001 (2020). https://doi.org/10.1088/1361-6528/ab554b
12. A. Kumar, M. Krishnaiah, J. Park, D. Mishra, B. Dash, H.-B. Jo et al., Multibit, Lead-Free Cs2SnI6 resistive random access memory with self-compliance for improved accuracy in binary neural network application. Adv. Funct. Mater. **34**(16), 2310780 (2024). https://doi.org/10.1002/adfm.202310780
13. J. Bae, J. Won, W. Shim, The rise of memtransistors for neuromorphic hardware and in-memory computing. Nano Energy. **126**, 109646 (2024). https://doi.org/10.1016/j.nanoen.2024.109646
14. D.S. Jeong, K.M. Kim, S. Kim, B.J. Choi, C.S. Hwang, Memristors for energy-efficient new computing paradigms. Adv. Electron. Mater. **2**(9), 1600090 (2016). https://doi.org/10.1002/aelm.201600090
15. S. Youn, K. Kim, J. Park, H. Kim, Energy-efficient Ising solver implementations in forming-free memristor crossbar arrays for combinatorial optimization problems. Nano Energy. **148**, 111633 (2026). https://doi.org/10.1016/j.nanoen.2025.111633
16. J. Park, H. Kim, Physical unclonable function with 3D stacked memristor crossbar array using self-differential pair. ACS Nano. **19**(31), 28135–28145 (2025)
17. F. Aguirre, A. Sebastian, Le M. Gallo, W. Song, T. Wang, J.J. Yang et al., Hardware implementation of memristor-based artificial neural networks. Nat. Commun. **15**(1), 1974 (2024). https://doi.org/10.1038/s41467-024-45670-9
18. T.-H. Kim, S. Kim, J. Park, S. Youn, H. Kim, Memristor crossbar array with enhanced device yield for in-memory vector–matrix multiplication. ACS Appl. Electron. Mater. **6**(6), 4099–4107 (2024)
19. M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa et al., Phase change memory as synapse for ultra-dense neuromorphic systems: application to complex visual pattern extraction. In: Proc. International Electron Devices Meeting (IEDM). (2011), p. 4.1–4
20. Y. Cheng, Z. Li, Y. Lin, Z. Wang, X. Shan, Y. Tao et al., Color recognition achieved in multiwavelength controlled plasmonic optoelectronic memristor for neuromorphic visual system. Adv. Funct. Mater. **35**(5), 2414404 (2025)
21. S. Youn, Y. Hwang, T.-H. Kim, S. Kim, H. Hwang, J. Park, H. Kim, Threshold learning algorithm for memristive neural network with binary switching behavior. Neural Netw. **176**, 106355 (2024)
22. Y.S. An, D. Kim, Y.R. Park, J.S. Eo, M. Kim, D. Kim et al., Implementation of monolithic 3D integrated TiOx memristor-based neural network for high-performance in-memory computing. Nano Energy. **139**, 110999 (2025). https://doi.org/10.1016/j.nanoen.2025.110999
23. J.W. Lee, J. Han, B. Kang, Y.J. Hong, S. Lee, I. Jeon, Strategic development of memristors for neuromorphic systems: low-power and reconfigurable operation. Adv. Mater. **37**(19), 2413916 (2025). https://doi.org/10.1002/adma.202413916
24. J. Park, S. Kim, M.S. Song, S. Youn, K. Kim, T.-H. Kim, H. Kim, Implementation of convolutional neural networks in memristor crossbar arrays with binary activation and weight quantization. ACS Appl. Mater. Interfaces. **16**(1), 1054–1065 (2024)
25. S. Boyn, J. Grollier, G. Lecerf, B. Xu, N. Locatelli, S. Fusil et al., Learning through ferroelectric domain dynamics in solid-state synapses. Nat. Commun. **8**(1), 14736 (2017). https://doi.org/10.1038/ncomms14736
26. S. Oh, T. Kim, M. Kwak, J. Song, J. Woo, S. Jeon et al., HfZrOx-based ferroelectric synapse device with 32 levels of conductance states for neuromorphic applications. IEEE Electron. Device Lett. **38**(6), 732–735 (2017). https://doi.org/10.1109/LED.2017.2698083
27. J. Yu, K.K. Min, Y. Kim, S. Kim, S. Hwang, T.-H. Kim et al., A novel physical unclonable function (PUF) using 16 × 16 pure-HfOx ferroelectric tunnel junction array for security applications. Nanotechnology. **32**(48), 485202 (2021). https://doi.org/10.1088/1361-6528/ac1dd5
28. S. Youn, H. Hwang, H. Kim, ZrO_2 seed layer effects on switching characteristics and low-frequency noise in HZO ferroelectric tunnel junction arrays. J. Alloys Compd. **1050**, 185701 (2026). https://doi.org/10.1016/j.jallcom.2025.185701
29. Z. Dang, F. Guo, Z. Wang, W. Jie, K. Jin, Y. Chai, J. Hao, Object motion detection enabled by reconfigurable neuromorphic vision sensor under ferroelectric modulation. ACS Nano. **18**(40), 27727–27737 (2024)
30. H. Hwang, S. Youn, H. Kim, Recent advances in ferroelectric materials, devices, and in-memory computing applications. Nano Conv. **12**(1), 55 (2025). https://doi.org/10.1186/s40580-025-00520-2
31. R.-H. Koo, W. Shin, J. Kim, J. Yim, J. Ko, G. Jung et al., Polarization pruning: reliability enhancement of hafnia-based ferroelectric devices for memory and neuromorphic computing. Adv. Sci. **11**(43), 2407729 (2024). https://doi.org/10.1002/advs.202407729
32. S.W. Kim, W. Shin, R.-H. Koo, J. Kim, J. Im, D. Koh et al., A new back-end-of-line ferroelectric field-effect transistor platform via laser processing. Small. **21**(15), 2406376 (2025). https://doi.org/10.1002/smll.202406376
33. J. Du, B. Sun, C. Yang, Z. Cao, G. Zhou, H. Wang, Y. Chen, Ferroelectric memristor and its neuromorphic computing applications. Mater. Today Phys. **50**, 101607 (2025)
34. S. Youn, H. Hwang, H. Kim, Ferroelectric tunnel junction memristor crossbar array with annealing optimization for in-memory computing. Adv. Intell. Syst. **8**(3), e202500817 (2026). https://doi.org/10.1002/aisy.202500817
35. Y. Park, E. Lim, S. Lee, V. Georgiev, S. Kim, S. Kim, Ferroelectric memristor crossbar arrays for highly integrated neuromorphic computing system. Nano Energy. **141**, 111137 (2025). https://doi.org/10.1016/j.nanoen.2025.111137
36. J. Yu, D. Ryu, W.Y. Choi, A novel verification method of minimizing verification error in AND-type flash arrays. IEEE Electron. Device Lett. **45**(8), 1472–1475 (2024). https://doi.org/10.1109/LED.2024.3416185
37. J. Shi, Z. Liu, J. Wei, A. Azam, C.-H. Lin, Y. Liu, S. Li, Revolutionizing nonvolatile memory: advances and future prospects of 2D floating-gate technology. ACS Nano. **19**(24), 21905–21926 (2025). https://doi.org/10.1021/acsnano.5c02740
38. R. Micheloni, L. Crippa, A. Marelli, Inside NAND flash memories (Springer Science & Business Media, 2010)
39. K.T. Park, S. Nam, D. Kim, P. Kwak, D. Lee, Y.H. Choi et al., Three-dimensional 128 Gb MLC vertical nand flash memory with 24-WL stacked layers and 50 MB/s high-speed programming. IEEE J. Solid-State Circuit. **50**(1), 204–213 (2015). https://doi.org/10.1109/JSSC.2014.2352293
40. W. Jung, H. Kim, D.B. Kim, T.H. Kim, N. Lee, D. Shin et al., 13.3 A 280-Layer 1 Tb 4b/cell 3D-NAND flash memory with a 28.5Gb/mm2 areal density and a 3.2GB/s high-speed IO rate. In: Proc. IEEE International Solid-State Circuits Conference (ISSCC). (2024), pp. 236-7
41. P. Wang, F. Xu, B. Wang, B. Gao, H. Wu, H. Qian, S. Yu, Three-dimensional nand flash for vector–matrix multiplication. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **27**(4), 988–991 (2019). https://doi.org/10.1109/TVLSI.2018.2882194
42. S.T. Lee, S. Lim, N. Choi, J.H. Bae, C.H. Kim, S. Lee et al., Neuromorphic technology based on charge storage memory devices. In: Proc. IEEE Symposium on VLSI Technology. (2018), pp. 169 – 70
43. G.H. Lee, S. Hwang, J. Yu, H. Kim, Architecture and process integration overview of 3D NAND flash technologies. Appl. Sci. (2021). https://doi.org/10.3390/app11156703
44. S. Hwang, J. Yu, M.S. Song, H. Hwang, H. Kim, Memcapacitor crossbar array with charge trap NAND flash structure for neuromorphic computing. Adv. Sci. **10**(32), 2303817 (2023). https://doi.org/10.1002/advs.202303817
45. T.H. Kim, O. Phadke, Y.C. Luo, H. Mulaosmanovic, J. Mueller, S. Duenkel et al., Tunable non-volatile gate-to-source/drain capacitance of FeFET for capacitive synapse. IEEE Electron. Device Lett. **44**(10), 1628–1631 (2023). https://doi.org/10.1109/LED.2023.3311344
46. J. Hur, Y.-C. Luo, A. Lu, T.-H. Wang, S. Li, A.I. Khan, S. Yu, Nonvolatile capacitive crossbar array for in-memory computing. Adv. Intell. Syst. **4**(8), 2100258 (2022). https://doi.org/10.1002/aisy.202100258
47. K.-U. Demasius, A. Kirschen, S. Parkin, Energy-efficient memcapacitor devices for neuromorphic computing. Nat. Electron. **4**(10), 748–756 (2021)
48. S. Ahn, J. Yu, H. Hwang, M.S. Song, D. Yu, S. Hwang et al., Coupling-free readout scheme for memcapacitors With NAND flash structure. IEEE Trans. Electron. Devices. **71**(8), 4670–4676 (2024). https://doi.org/10.1109/TED.2024.3418289

49. C.-Y. Kao, J.-G. Hwu, Inversion-sensing SiO2-based MOS capacitive synapse for neuromorphic computing. Appl. Phys. Lett. **126**(21), 213504 (2025)
50. H. Hwang, J. Yu, S. Youn, W.Y. Choi, H. Kim, 2C-Ternary content addressable memory in memcapacitor crossbar array with NAND flash structure. Small. **21**(9), 2408618 (2025). https://doi.org/10.1002/smll.202408618
51. S. Hwang, H. Kim, J. Park, M.W. Kwon, M.H. Baek, J.J. Lee, B.G. Park, System-level simulation of hardware spiking neural network based on synaptic transistors and I&F neuron circuits. IEEE Electron. Device Lett. **39**(9), 1441–1444 (2018). https://doi.org/10.1109/LED.2018.2853635
52. B. Rueckauer, I.-A. Lungu, Y. Hu, M. Pfeiffer, S.-C. Liu, Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. Front. Neurosci. 11–2017 (2017). https://doi.org/10.3389/fnins.2017.00682
53. T. Bu, W. Fang, J. Ding, P. Dai, Z. Yu, T. Huang, Optimal ANN-SNN conversion for high-accuracy and ultra-low-latency spiking neural networks. arXiv preprint arXiv:230304347. 2023
54. E.O. Neftci, H. Mostafa, F. Zenke, Surrogate gradient learning in spiking neural networks. IEEE Signal. Process. Mag. **36**(6), 51–63 (2019)

Publisher's note

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