Check for updates

Advanced Science

ADVANCED SCIENCE Open Access

www.advancedscience.com

RESEARCH ARTICLE OPEN ACCESS

A Unified Flash Memory Platform for Mode-Adaptive and Robust AI Computation

Dayeon Yu[1] | Hwiho Hwang[1] | Byeongchan Oh[2] | Junmo Lee[3] | Shimeng Yu[3] | Tae-Hyeon Kim[2] | Hyungjin Kim[1]

[1]Division of Materials Science and Engineering and Department of Semiconductor Engineering, Hanyang University, Seoul, South Korea | [2]Department of Semiconductor Engineering, Seoul National University of Science and Technology, Seoul, South Korea | [3]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

Correspondence: Shimeng Yu (shimeng.yu@ece.gatech.edu) | Tae-Hyeon Kim (teahyeon@seoultech.ac.kr) | Hyungjin Kim (hkim12@hanyang.ac.kr)

Received: 3 February 2026 | **Revised:** 16 June 2026 | **Accepted:** 30 June 2026

Keywords: AND-type array | compute-in-memory (CIM) | dual-mode operation | flash memory | hardware-based neural network

ABSTRACT

Computing-in-memory (CIM) architectures offer a promising route toward energy-efficient artificial intelligence by reducing data-movement overhead. However, most existing CIM hardware operates at a fixed trade-off between accuracy, energy efficiency, and robustness, limiting adaptability to diverse workloads. Here, we present a dual-mode CIM accelerator based on an AND-type charge-trap flash array that enables energy-adaptive operation without device-level structural modification. By integrating transistor-mode current sensing and capacitor-mode charge sensing in the same device structure, the proposed architecture allows flexible switching between high-precision computation and ultra-low-power, noise-resilient operation within a single hardware platform through peripheral switching associated with each sensing mode. Experimental results demonstrate reliable vector–matrix multiplication, hardware neural network inference, and strong tolerance to device and voltage variations. System-level benchmarking further confirms improved energy efficiency and reduced peripheral overhead. This work establishes a practical and scalable CIM platform that dynamically balances performance and robustness, providing a versatile foundation for energy-adaptive artificial intelligence (AI) hardware.

1 | Introduction

The rapid expansion of artificial intelligence (AI) has led to an unprecedented increase in computational demand, exposing the fundamental energy and latency limitations associated with data movement between memory and processing units, commonly referred to as the memory wall [1–9]. As AI models continue to scale in size and complexity, particularly in edge and energy-constrained environments, improving energy efficiency has become as critical as achieving high computational accuracy. Computing-in-memory (CIM) architectures have therefore emerged as a promising paradigm by enabling computation directly within memory arrays, significantly reducing data transfer overhead and energy consumption [10–18]. Despite these advantages, most existing CIM accelerators operate at a fixed design point in the trade-off space between computational accuracy, energy efficiency, and reliability. In practice, however, AI workloads exhibit highly diverse and dynamic requirements. Training, calibration, and accuracy-critical inference demand high precision and throughput, whereas edge inference, always-on sensing, and long-term operation prioritize ultra-low power consumption and robustness against device variability, noise, and aging effects [19–27]. The inability of current CIM hardware to adapt to such heterogeneous workload demands fundamentally limits its energy efficiency and practical applicability at the system level.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Advanced Science* published by Wiley-VCH GmbH

Charge-trap flash (CTF) memory has been widely explored as a CIM platform thanks to its mature fabrication process, excellent retention characteristics, and scalability [28–36]. Conventional NAND flash-based CIM architectures, however, suffer from intrinsic limitations arising from their serial string configuration. As the string length increases, the available current decreases and sensing margins degrade, while access latency and peripheral overhead grow correspondingly [37–46]. These architectural constraints restrict parallelism in vector–matrix multiplication (VMM) and introduce additional energy overhead at both the array and circuit levels, particularly in large-scale implementations. To overcome these limitations, AND-type flash arrays with inherently parallel structures have been proposed [47–62]. In AND-type architectures, memory cells controlled by independent wordlines (WLs) are connected in parallel along shared drain lines, enabling efficient current summation and making them well suited for CIM operations. From an architectural perspective, this parallelism offers clear advantages in both throughput and energy efficiency. Nevertheless, most flash-based CIM implementations, including those based on AND-type arrays, rely exclusively on a single sensing modality: either transistor-mode (current-based) sensing or capacitor-mode (charge-based) sensing.

Transistor-mode sensing supports multilevel weight representation and fast readout, making it suitable for performance-critical workloads. However, it typically incurs higher dynamic energy consumption and increased sensitivity to device-level variations, read noise, and analog-to-digital conversion (ADC) overhead [63]. In contrast, capacitor-mode sensing enables binary operation with near-zero static power consumption and strong tolerance to noise and variability, yet its limited precision and slower sensing dynamics restrict its use in applications requiring high inference accuracy [64–71]. As a result, most current CIM accelerators are confined to a narrow operating regime, forcing designers to choose between energy-efficient but low-precision hardware and high-accuracy but energy-intensive solutions. From an energy-systems perspective, this fixed operating point can be a critical bottleneck. System-level AI implementations would benefit substantially from hardware that can dynamically adapt its operating mode, prioritizing accuracy when power is abundant and switching to robust, ultra-low-power operation under energy-constrained or noisy conditions. However, enabling such adaptability within a single CIM platform remains an open challenge, as most prior approaches require separate device structures, additional peripheral circuits, or fundamentally different memory technologies.

In this work, we address this challenge by presenting a dual-mode CIM accelerator based on an AND-type CTF array that integrates both transistor-mode and capacitor-mode sensing within the same TiN/Al_2O_3/Si_3N_4/SiO_2/poly-Si (TANOS) device stack. Since both sensing schemes are implemented using the same CTF memory cell structure and array stack, the proposed architecture supports flexible selection of the sensing mode without requiring device-level structural modification, while the mode selection is achieved through the corresponding bias and readout conditions. Transistor-mode provides multilevel, high-throughput computation for accuracy-critical tasks, while capacitor-mode offers binary, charge-based operation with near-zero static power consumption and strong robustness against device and voltage variations. By enabling flexible operation across the energy–accuracy–reliability design space, the proposed dual-mode architecture expands the energy-efficiency envelope of flash-based CIM. Through comprehensive evaluation across device, array, and system levels, we demonstrate that the proposed platform can flexibly balance computational accuracy and energy efficiency through workload-dependent sensing-mode selection. This energy-adaptive CIM approach provides a practical route toward workload-adaptive and energy-aware AI hardware.

2 | Device Structure and Operating Principles

The AND-type CTF array with a TANOS stack was fabricated following the process flow illustrated in Figure 1a. A 300 nm buried oxide (BOX) layer was first formed on a p-type (100) silicon wafer by furnace wet oxidation. Subsequently, a 100 nm amorphous silicon layer was deposited on the BOX using low-pressure chemical vapor deposition (LPCVD) at 550°C with a SiH_4 flow rate of 100 sccm, followed by solid-phase crystallization (SPC) at 600°C for 24 h. SPC was employed to obtain a poly-Si channel with uniform electrical characteristics suitable for array-level operation. The active regions and source lines (SLs) and drain lines (DLs) were then defined by photolithography and patterning. A 4.8 nm SiO_2 tunneling oxide was deposited by LPCVD at 800°C using dichlorosilane (DCS, 40 sccm) and N_2O (80 sccm), followed by deposition of a 6.3 nm Si_3N_4 charge-trap layer at 800°C using DCS (40 sccm) and NH_3 (400 sccm). A 9.5 nm Al_2O_3 blocking oxide was subsequently deposited by atomic layer deposition (ALD) at 340°C using trimethylaluminum (TMA) as the precursor. A 50 nm TiN gate electrode was then deposited by metal–organic chemical vapor deposition (MOCVD). Prior to gate patterning, a 15 nm SiO_2 hard mask was deposited by plasma-enhanced chemical vapor deposition (PECVD) at 380°C to improve adhesion between the TiN layer and the photoresist during subsequent etching processes. After photolithography, the SiO_2 hard mask was etched using buffered oxide etchant (BOE, HF:NH_4F = 1:100) at room temperature. TiN gate patterning was performed by inductively coupled plasma (ICP) etching followed by wet etching in diluted hydrogen peroxide (H_2O_2:DI = 1:4) at 67°C. The source and drain regions were formed by As^+ ion implantation at a dose of 2×10^{15} cm^{-2} and an energy of 40 keV, followed by dopant activation via rapid thermal annealing (RTA). Finally, a 300 nm interlayer dielectric (ILD) was deposited by PECVD, and contact-hole patterning and metallization were carried out to complete the back-end-of-line (BEOL) process. All fabrication steps were carried out using CMOS-compatible processes, ensuring scalability and compatibility with conventional flash memory manufacturing.

The array structure, in which devices controlled by independently addressed wordlines (WLs) are connected in parallel along shared source lines (SLs) and drain lines (DLs), is illustrated in Figure 1b. This parallel connectivity enables simultaneous current summation, making the array well suited for efficient VMM operations required for CIM applications. Figure 1c presents a top-view optical microscope image of the fabricated 48 × 24 AND-type flash array, together with a cross-sectional transmission electron microscopy (TEM) image confirming the formation of the TANOS stack. To verify dual-mode operations and eval-

21983844, 2026, 56, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.76619 by Hanyang University Library, Wiley Online Library on [28/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

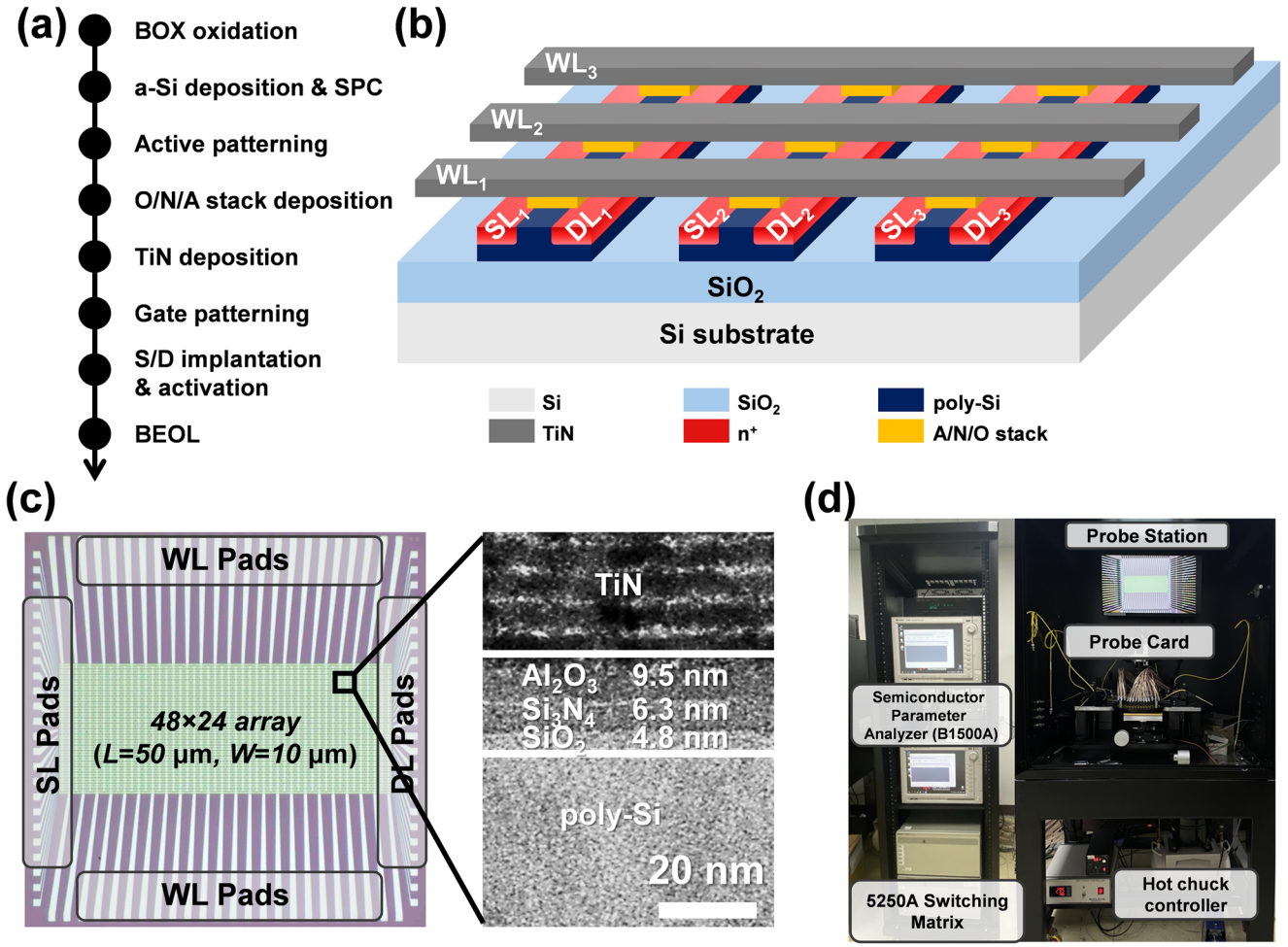

FIGURE 1 | AND-type flash array architecture and experimental setup. (a) Fabrication process flow of the AND flash memory cell with a $TiN/Al_2O_3/Si_3N_4/SiO_2$/poly-Si (TANOS) charge-trap stack. (b) Schematic illustration of the AND-type array architecture, where memory cells controlled by independently addressed WLs are connected in parallel along shared SLs and DLs, enabling current summation for VMM operations. (c) Top-view optical microscope image of the fabricated 48 × 24 AND flash array with a cross-sectional transmission electron microscopy (TEM) image showing the TANOS gate stack. (d) Electrical measurement setup used for transistor- and capacitor-mode characterizations.

uate the electrical characteristics of both transistor-mode and capacitor-mode sensing, the fabricated array was characterized with the measurement system, as shown in Figure 1d. The array was interfaced with a semiconductor parameter analyzer and a switching matrix, and all measurements were controlled using a custom-developed C++ control software. The detailed characterization methods are explained in the Experimental Section.

Although the AND-type flash array is fundamentally based on a transistor structure, it supports both transistor-mode and capacitor-mode sensing through different readout schemes, as illustrated in Figure 2a,b, respectively. In transistor-mode, the SL is grounded, and a read voltage is applied to the DL, allowing the device state to be sensed through the resulting drain current. Multi-level states are realized by modulating channel formation, effectively tuning the device conductance to generate distinct current levels within the on–off window. Owing to its current-based sensing mechanism and fast response, transistor-mode operation is well suited for performance- and accuracy-critical applications. By contrast, in capacitor-mode, the SL is electrically isolated from ground and left floating, and the device state is sensed through the effective capacitance formed between the WL and DL. The SL is intentionally left floating during capacitor-mode operation, enabling mode selection through a simple SL switching scheme without requiring additional bias circuitry. Before channel formation, only the gate-to-drain overlap capacitance (C_{ov}) contributes to the total capacitance, resulting in a low-capacitance state (C_{off}). As the gate voltage increases and channel formation begins, the effective capacitive coupling between the WL and DL expands. Once the gate voltage exceeds the threshold voltage (V_{th}), additional contributions from the source-side C_{ov} and the gate oxide capacitance (C_{ox}) are added, yielding a high-capacitance state (C_{on}) of approximately $2C_{ov} + C_{ox}$. Because capacitor-mode operates with only two well-separated states, it provides strong robustness against device-to-device variation and external noise. Moreover, since the readout relies on charge sensing through capacitive coupling, the static power consumption is negligible. In addition, the simplified binary sensing in capacitor-mode reduces the burden on peripheral readout circuits, contributing to improved system-level energy efficiency. These characteristics make capacitor-mode operation particularly suitable for energy-constrained and long-term stable computing-in-memory applications. Together,

21983844, 2026, 56, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.76619 by Hanyang University Library, Wiley Online Library on [28/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

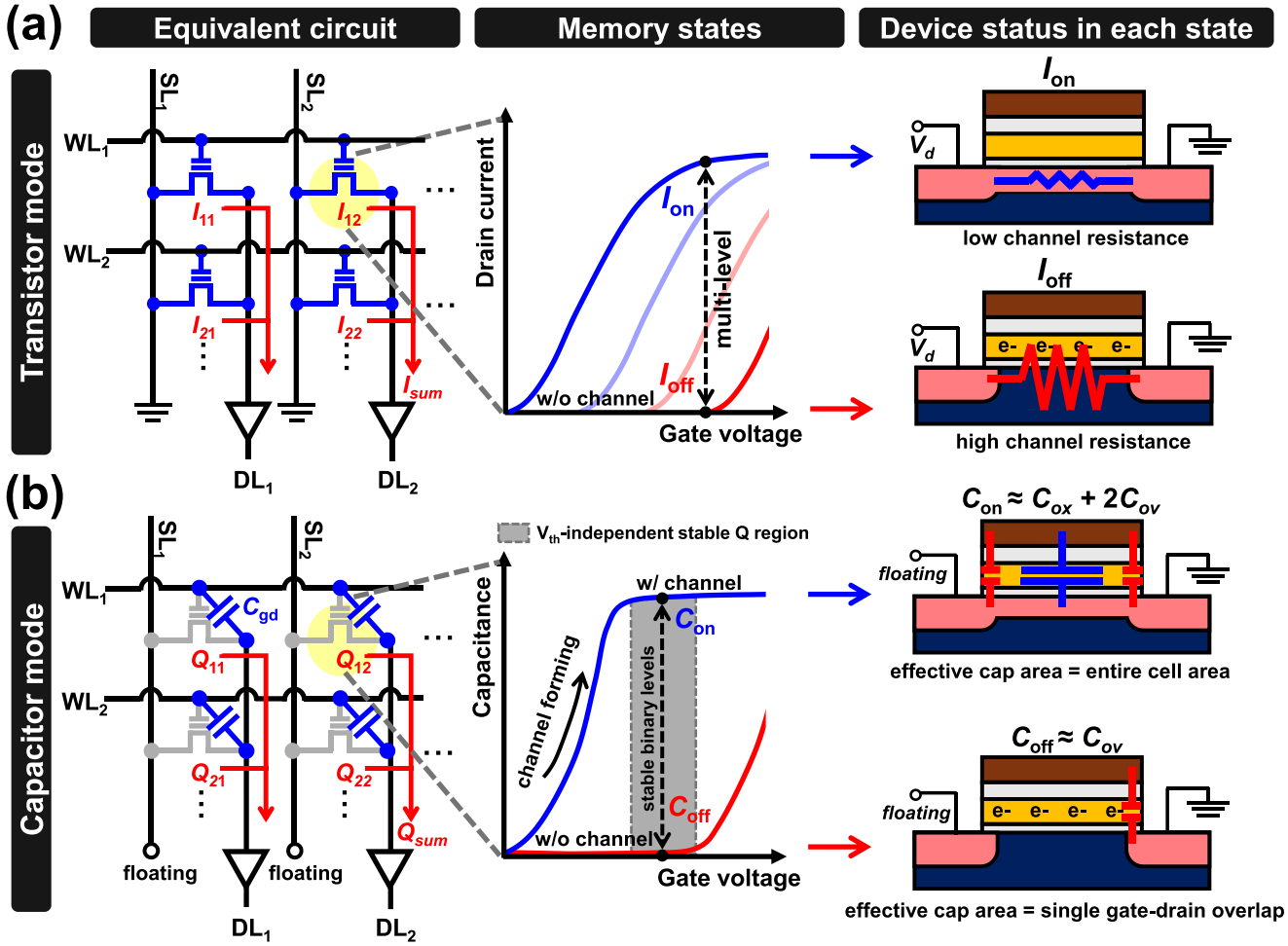

FIGURE 2 | Dual-mode operation mechanisms of AND-type flash memory cell. (a) Transistor-mode sensing scheme, where SL is grounded, and a read voltage is applied to DL to sense the device state through the resulting drain current. Multilevel states are realized by modulating channel formation, enabling current-based weight representation for high-precision computation. (b) Capacitor-mode sensing scheme, where the SL is electrically isolated and left floating, and the device state is sensed through the effective capacitance between WL and DL. Before channel formation, only the gate-to-drain C_{ov} contributes, yielding a low-capacitance state, C_{off}, while channel formation above V_{th} introduces additional capacitive coupling, resulting in a high-capacitance state, C_{on}.

the two sensing modes provide complementary operating characteristics, enabling flexible trade-offs between computational performance and energy efficiency within the same AND-type array platform through peripheral reconfiguration associated with each sensing mode.

3 | Transistor- and Capacitor-Mode Operations

Figure 3a shows the transfer characteristics of the device during programming based on Fowler–Nordheim (FN) tunneling. Incremental step pulse programming (ISPP) [72] was performed by applying 10 μs gate pulses from 9 to 18 V in 0.15 V increments while grounding both the source and drain. As the ISPP voltage increased, V_{th} shifted toward higher values, resulting in a memory window of over 6 V. Conversely, Figure 3b presents the erase characteristics. An initial 18 V, 10 μs program pulse was followed by incremental step pulse erase (ISPE) voltages ranging from −9 to −18 V in −0.15 V increments, with a pulse width of 100 μs applied to the gate. This process induced a leftward shift of V_{th}, eventually restoring the device to a state close to its initial condition. Through these program and erase operations, multiple current levels can be generated at a fixed read voltage, enabling multilevel storage in transistor-mode.

Figure 3c shows the endurance characteristics of the device, demonstrating stable program and erase operation over 10^5 cycles using 15 V, 10 μs program pulses and −21 V, 100 μs erase pulses. Based on this stable write-and-verify scheme, 100 randomly selected devices within a single array were individually tuned to eight distinct current levels ranging from 1 to 8 μA in 1 μA increments, using the bias conditions illustrated in Figure S1a. The resulting current distributions are summarized as histograms in Figure 3d, with the corresponding mean and standard deviation for each target level. All programmed levels exhibit a mean absolute error of approximately 0.1 μA, demonstrating highly precise and reproducible current-level tuning across the array. During the programming or erasing of a selected cell, the program or erase voltage was applied to the corresponding WL, while the surrounding eight neighboring cells were biased using a $V/2$ inhibit scheme. As confirmed in Figure S1b,c, the unselected neighboring cells maintained their original V_{th} without notice-

21983844, 2026, 56, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.76619 by Hanyang University Library, Wiley Online Library on [28/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

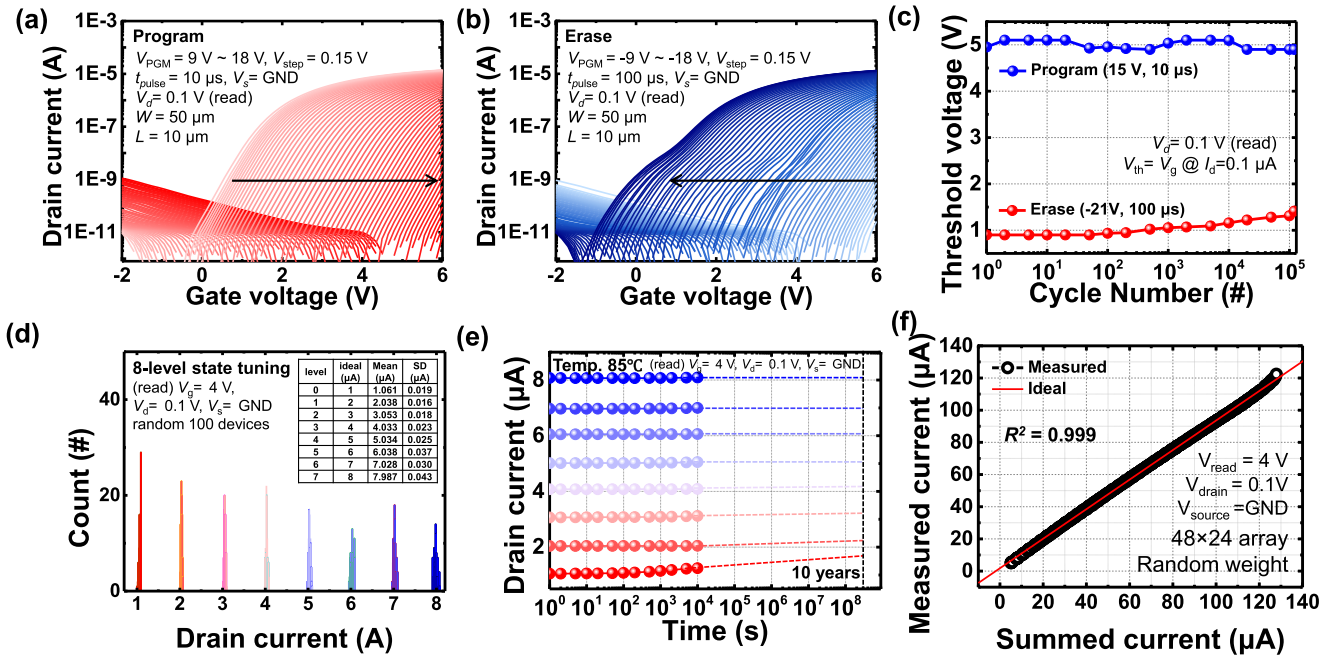

FIGURE 3 | Program/erase characteristics and transistor-mode VMM operation of AND flash array. (a) Transfer characteristics of the AND-type flash memory cell during ISPP, showing a gradual rightward shift of the threshold voltage V_{th} with increasing program pulse amplitude. (b) Transfer characteristics during ISPE, where successive erase pulses induce a leftward shift of V_{th} , returning the device toward its initial state. (c) Endurance characteristics of the device, demonstrating stable program and erase operation over 10^5 cycles. (d) Histogram of eight programmed current levels obtained from 100 randomly selected devices in the array, showing the mean and standard deviation for each target level and confirming precise multi-level current tuning. (e) Retention characteristics of the eight programmed current levels measured at 85°C, indicating stable retention and clear level separation extrapolated to long-term operation. (f) Experimental verification of analog VMM operation in a 48 × 24 array, where the measured DL current exhibits a strong linear correlation with the arithmetic sum of individual device currents.

able disturbance, verifying effective inhibit operation within the AND-type array. With this write-and-verify scheme, a 48 × 24 down-sampled Cameraman image encoded in 3-bit resolution was selected as a target pattern, as shown in Figure S2a. Each device in the 48 × 24 AND-type array was programmed to one of eight discrete current levels corresponding to the target image, and the resulting drain current map was subsequently measured. The resulting map closely reproduces the target image, demonstrating accurate and uniform weight transfer in the AND-type array under current-based transistor-mode operation. This result demonstrates that the AND-type flash array can reliably support multi-level weight programming for large-scale analog CIM applications.

Also, the retention characteristics were evaluated by monitoring these eight programmed levels for 10^4 s at 85°C, as shown in Figure 3e. Extrapolation of the measured data indicates that all eight levels remain clearly distinguishable for up to 10 years (3×10^8 s), confirming excellent charge retention. Such stable endurance and retention characteristics are essential for maintaining weight accuracy and minimizing recalibration overhead in energy-efficient CIM systems. Finally, Figure 3f evaluates the analog VMM capability of the array. After randomly programming device states within a 48 × 24 array, the number of activated WLs was gradually increased, and the measured DL current was compared with the arithmetic sum of individual device currents. The resulting correlation exhibits an R^2 value of 0.999, confirming highly accurate analog current summation suitable for VMM operations.

Figure 4a presents the *C*–*V* characteristics of a device with a channel width of 50 µm and a channel length of 10 µm measured in capacitor-mode, where the SL is left floating and the capacitance between the WL and DL is sensed using a 10 kHz, 200 mV small-signal AC voltage. During the ISPP, 10 µs gate pulses ranging from 9 to 18 V in 0.15 V increments shift V_{th} toward higher values. When the gate voltage is below V_{th}, only the gate-to-drain C_{ov} contributes to the total capacitance, resulting in C_{off} . Once the gate voltage exceeds V_{th} , channel formation enhances the capacitive coupling, and additional contributions from the source-side overlap capacitance C_{ov} and C_{ox} increase the total capacitance to C_{on}, approximately equal to $2C_{ov} + C_{ox}$. Figure 4b shows the erase characteristics obtained by applying an initial 18 V, 10 µs program pulse followed by the ISPE pulses ranging from −9 to −18 V in 0.15 V increments with a pulse width of 100 µs, resulting in a negative shift of V_{th} . Although the measured capacitance varies with both the device state and the applied gate voltage, the device consistently outputs either C_{off} (≈ 0.5 pF) or C_{on} (≈ 2.0 pF) outside the transition region, maintaining clear binary separation.

From the measured *C*–*V* characteristics depending on the device state, V_{th} of each device was extracted by defining V_{th} as the gate voltage at which C_{on} is consistently observed, and the device capacitance value is plotted according to the gate voltage, as shown in Figure 4c. These results confirm that, for a given device, the capacitance remains stable outside the transition region: C_{off} is consistently observed at low gate voltages, while C_{on} is obtained at gate voltages well above V_{th}. This wide separation

21983844, 2026, 56, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.76619 by Hanyang University Library, Wiley Online Library on [28/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

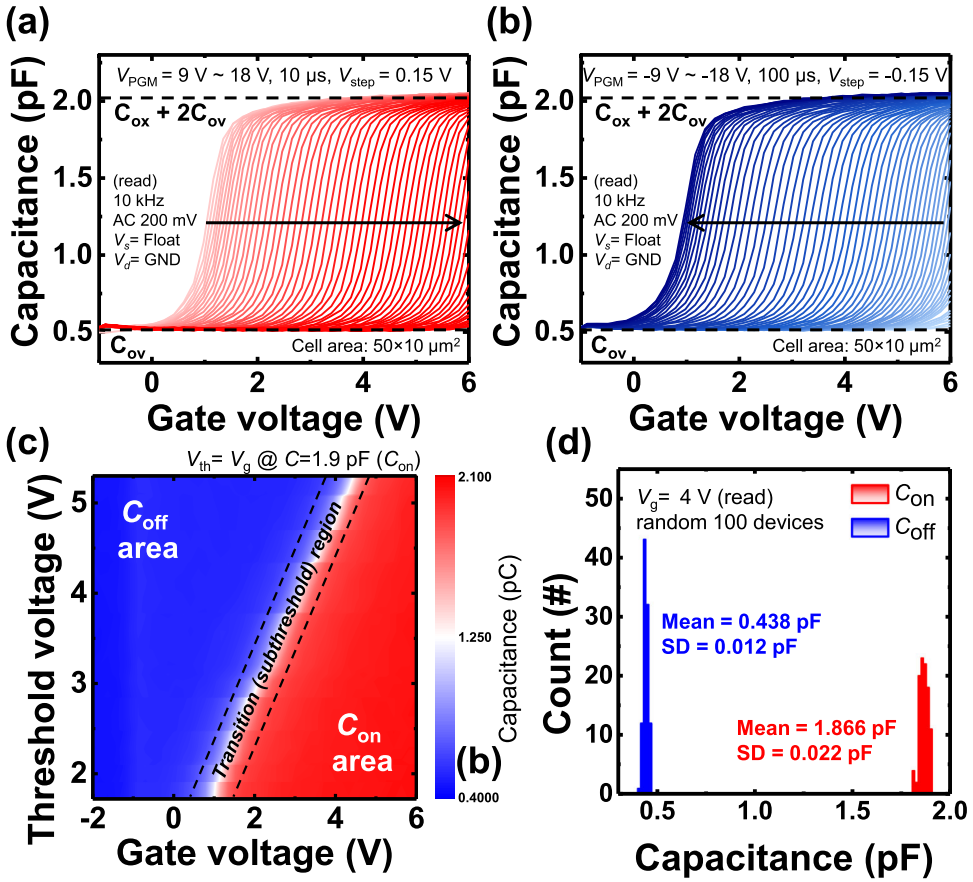

FIGURE 4 | Capacitor-mode sensing characteristics. (a) *C–V* characteristics of the AND-type flash memory cell measured in capacitor-mode under ISPP operation, showing the transition from C_{off} to C_{on} as channel formation occurs above V_{th}. (b) Erase characteristics of the device in capacitor-mode, where ISPE induces a negative shift of V_{th} , modulating the device capacitance state. (c) Extracted device capacitance as a function of gate voltage for different programmed states, illustrating stable capacitance plateaus outside the transition region and enabling robust binary state sensing. (d) Histograms of C_{on} and C_{off} extracted from 100 randomly selected devices in the array, demonstrating tight distributions with small standard deviation and clear binary separation across the array.

enables robust state sensing outside the transition region. Similar behavior can be observed when the gate voltage is fixed. The devices with higher V_{th} reliably maintain the C_{off} state, whereas devices with lower V_{th} consistently exhibit the C_{on} state. In contrast, transistor-mode directly translates threshold voltage and device state variations into output current variation because the drain current is directly affected by V_{th} variation. Conversely, capacitor-mode maintains stable C_{off} and C_{on} values, demonstrating robustness against both gate voltage variation and device-to-device V_{th} or state variation. Consequently, reliable binary sensing can be preserved even when the available memory window is reduced, provided that the memory window remains sufficiently larger than the transition region. Figure 4d further supports the reliability of capacitor-mode sensing by presenting histograms of C_{on} and C_{off} extracted from 100 randomly selected devices. Both capacitance states exhibit a small standard deviation of 0.022 pF, confirming accurate tuning and clear binary separation across the array. To directly demonstrate the array-level implications of this robust binary tuning capability, a 48 × 24 down-sampled Cameraman image encoded in 1-bit resolution was mapped onto a same-sized AND-type array by assigning the binary capacitance states to individual devices. As shown in Figure S2b, the measured capacitance map closely reproduces the target binary image, confirming stable binary state transfer and distinct device-state separation based on capacitance value.

In addition, VMM in capacitor-mode is performed by the discharging current through the DLs, as illustrated in Figure 5a. A falling voltage pulse from the charging voltage (V_c) to the discharging voltage (V_d) is applied to the selected WLs, while the unselected WLs are grounded. The selected cells discharge according to their capacitance values, and the resulting discharging currents are summed along the DLs to realize the VMM operation. When a falling pulse from V_c = 5 V to V_d = 0 V with a fall time of 4 μs is applied, the time-integrated DL current closely matches the charge determined by the effective device capacitance, confirming accurate capacitor-mode VMM operations. Here, the 4 μs fall time was selected to reliably capture the transient discharging current using the waveform generator/fast measurement unit (WGFMU) with a 10 ns current sampling resolution. As such, the selected fall time reflects the measurement condition rather than the intrinsic speed limitation of the proposed VMM scheme. Although capacitor-mode operation requires additional charging and integration processes for charge-domain sensing compared with transistor-mode operation, the associated latency overhead can be mitigated through architecture-level optimizations, including enhanced

21983844, 2026, 56, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.76619 by Hanyang University Library, Wiley Online Library on [28/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

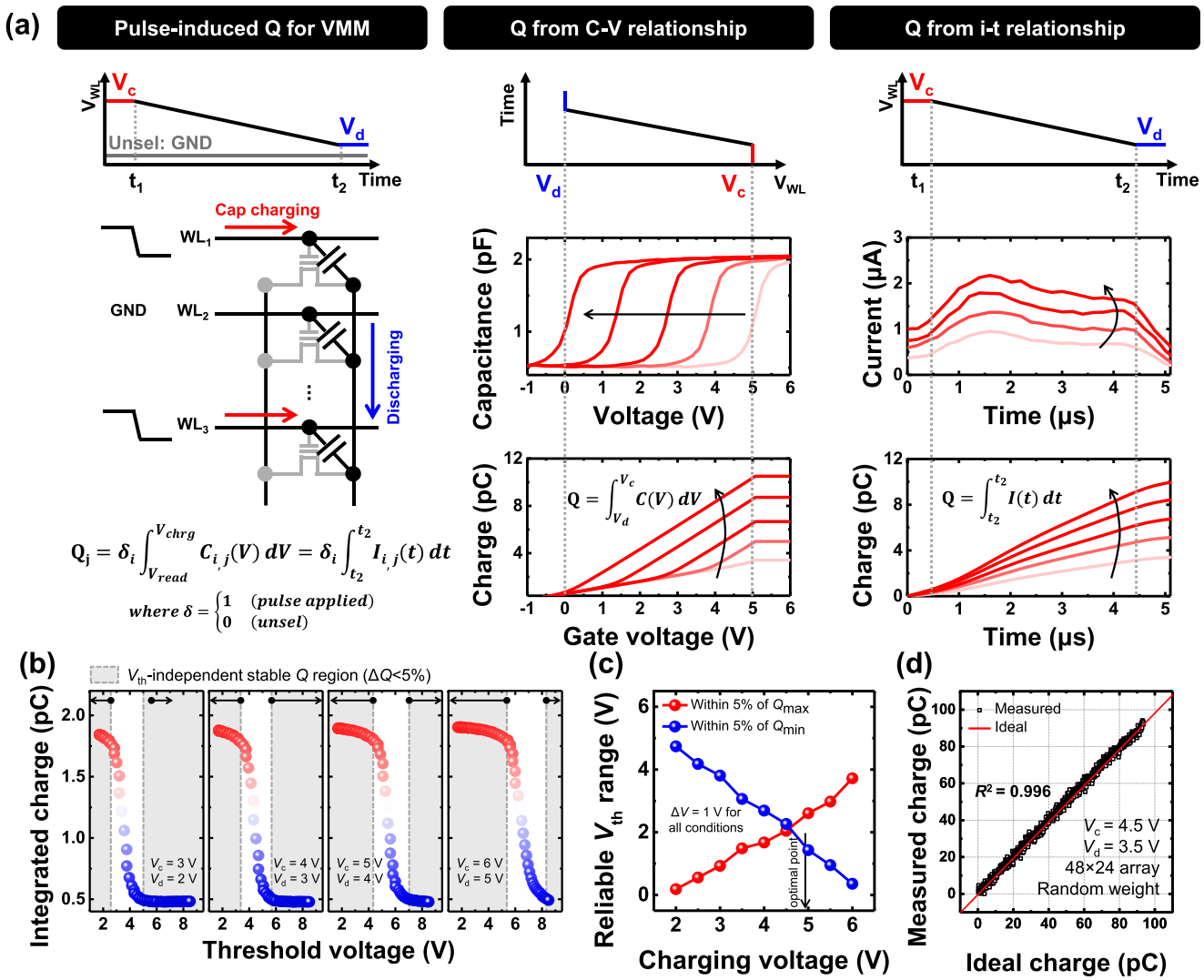

FIGURE 5 | Capacitor-mode VMM operation and read-pulse optimization. (a) Schematic illustration of capacitor-mode VMM operation, where a falling voltage pulse is applied to selected WLs and the resulting discharging currents from individual cells are summed along DLs to perform charge-domain VMM. (b) Integrated output charge as a function of device V_{th} under different charging voltages V_c while maintaining a constant voltage difference $V_c - V_d = 1$ V. The V_{th}-independent stable charge regions ($\Delta Q < 5\%$) for Q_{max} and Q_{min} are highlighted, demonstrating robust binary readout in capacitor-mode. (c) Extracted V_{th} ranges corresponding to stable Q_{max} and Q_{min} (within 5% variation) as a function of V_c, identifying an optimal read-pulse condition that provides balanced stability for both binary states. (d) Experimental verification of capacitor-mode VMM operation in a 48 × 24 array using the optimized read-pulse condition, where the measured output charge exhibits a strong linear correlation with the ideal summed charge.

row and column parallelism [73, 74]. Moreover, because the integration process is performed in parallel across columns, the additional sensing step does not necessarily result in a proportional reduction in throughput. As a result, competitive throughput and latency can be maintained when appropriate architecture-level optimizations and parallel operation strategies are employed.

Also, to verify that the floating-SL configuration does not induce memory-state degradation during repeated read operation, the stability of capacitor-mode read operation under the floating SL condition was evaluated by repeatedly reading two cells sharing the same SL for up to 10^5 cycles. As shown in Figure S3, one cell was tuned to the erase state exhibiting C_{on} by applying a −18 V, 100 μs erase pulse, while the other cell sharing the same SL was tuned to the program state exhibiting C_{off} using a 16 V, 10 μs program pulse. During the repeated read operation, a falling read pulse from 4.5 to 3.5 V with a fall time of 4 μs was applied to the selected WL, while all unselected WLs were grounded. The transient discharging current from each cell was integrated to extract the charge output. No noticeable degradation in the charge output was observed over 10^5 repeated read cycles, indicating stable capacitor-mode read operation under the floating SL condition. This observation further suggests that floating-SL operation does not induce measurable memory-state drift during repeated read operations.

To identify an optimal read pulse that enables stable binary-state output in capacitor-mode, the voltage difference between V_c and V_d was fixed at 1 V, while their absolute voltage levels were

21983844, 2026, 56, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.76619 by Hanyang University Library, Wiley Online Library on [28/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

varied. Under these conditions, the output charge from a single device in the array was measured as a function of device state, as shown in Figure 5b. Specifically, four operating conditions were evaluated by increasing V_c from 3 to 6 V in 1 V increments while maintaining $V_c - V_d = 1$ V. For $V_c = 3$ V and $V_d = 2$ V, devices with lower V_{th} produced a high integrated charge (Q_{max}) of approximately 1.9 pC, remaining within a 5% variation, whereas devices with higher V_{th} consistently yielded a low integrated charge (Q_{min}) of approximately 0.5 pC, also within a 5% variation, once outside the transition region. The V_{th}-independent stable charge regions ($\Delta Q < 5\%$) are highlighted in gray in Figure 5b. Within these regions, the integrated output charge remains nearly constant despite variations in device V_{th}, enabling robust capacitor-mode readout that is inherently tolerant to device-state fluctuations. This behavior confirms the existence of well-defined operating windows in which reliable binary sensing is maintained irrespective of V_{th} variation.

Furthermore, the same analysis was repeated for higher V_c values up to 6 V. As the read pulse voltage levels were shifted upward while maintaining $V_c - V_d = 1$ V, the V_{th} range over which Q_{min} remained stable gradually decreased, whereas the stable V_{th} range for Q_{max} expanded. Figure 5c summarizes these trends by plotting the V_{th} ranges corresponding to stable Q_{min} and Q_{max}, each defined within a 5% variation, as a function of V_c. As a result of these opposing trends, an optimal operating point emerges at an intermediate charging voltage. Specifically, at $V_c = 4.5$ V (with $V_d = 3.5$ V), the stability margins for Q_{min} and Q_{max} become comparable, providing balanced tolerance for both binary states. At this operating point, variations in device state of up to approximately 2 V in V_{th} result in less than 5% variation in both Q_{min} and Q_{max}, ensuring highly stable binary readout. Using this optimal read pulse condition, Figure 5d evaluates capacitor-mode VMM operation in a 48 × 24 array. Devices were randomly programmed to either C_{on} or C_{off} state, and the number of selected WLs gradually increased. The output charge, obtained by integrating the measured DL current, was compared with the calculated summed charge. The measured results closely match the ideal values, yielding an R^2 of 0.996, which confirms highly accurate VMM operation in capacitor-mode. This robust charge-domain operation enables accurate VMM with minimal sensitivity to device variability and analog noise, making capacitor-mode sensing particularly attractive for energy-efficient CIM.

4 | CIM Performance in Dual-Mode Operations

A hardware-based convolutional neural network (CNN) for Fashion-MNIST classification was implemented to compare the performance of transistor-mode and capacitor-mode operation, as shown in Figure 6a. The final fully connected layer with a size of 24 × 10 was mapped onto a 24 × 20 sub-array using a differential-pair scheme. In transistor-mode, 3-bit device states were employed to realize 15-level weight mapping, whereas in capacitor-mode, 1-bit device states were used to implement ternary weight mapping. Accordingly, 15-level quantization-aware training (QAT) was applied in software for transistor-mode operation, while ternary QAT was used for capacitor-mode operation. Figure 6b shows the software-based training results obtained using QAT under different assumed device state levels. Specifically, 15-level QAT was applied for transistor-mode operation, while ternary QAT was used to emulate capacitor-mode operation. For hardware mapping, the inputs to the third fully connected layer were binarized, and network training was performed for 50 epochs. Under these software-level conditions, the 15-level QAT achieved a classification accuracy of 94.53%, whereas the ternary QAT yielded an accuracy of 93.22%. Figure 6c presents the corresponding hardware inference results after transferring the trained weights onto the fabricated array and performing inference using 500 test samples. When transistor-mode was employed for 15-level QAT-based hardware inference, an accuracy of 94.19% was obtained, whereas capacitor-mode hardware inference based on ternary QAT achieved an accuracy of 93.08%. Compared to the software-level results, both modes exhibit a modest accuracy degradation after hardware mapping, which can be attributed to non-idealities such as programming inaccuracies, device-to-device variation, and read noise in the physical array. These results indicate that transistor-mode benefits from higher weight resolution, resulting in higher peak inference accuracy.

To evaluate tolerance to device-state variation, additional simulations were conducted by injecting random noise into the mapped weights. In transistor-mode, weights were represented by the drain current corresponding to each multi-level device state, whereas in capacitor-mode, weights were represented by the capacitance values corresponding to the binary states C_{on} and C_{off}. For each weight $W_{i,j}$, corresponding to the device located at the i-th WL and j-th DL, a random perturbation was sampled from a uniform distribution within the range $[-p \times W_{i,j}, +p \times W_{i,j}]$, where p denotes the normalized noise level. The perturbed weights were then used to evaluate inference accuracy. This procedure, illustrated in the flowchart in Figure 6d, was repeated 30 times for each noise condition using a test set of 1,000 samples, and the results are summarized as box plots. Even when the normalized noise level reached 200%, capacitor-mode inference exhibited a relatively small average accuracy degradation of 19.22% compared to the baseline. In contrast, transistor-mode inference, which relies on higher-resolution weight states and is therefore more sensitive to perturbations, showed a larger average accuracy degradation exceeding 26%. These results demonstrate that transistor-mode enables higher peak accuracy, whereas capacitor-mode provides superior robustness against device-state variation, highlighting the complementary advantages of the proposed dual-mode CIM architecture. This accuracy–robustness trade-off suggests that capacitor-mode operation is particularly advantageous under energy- or noise-constrained conditions, while transistor-mode operation is better suited for accuracy-critical workloads.

Additional system-level simulations were conducted using the CIM benchmarking framework NeuroSim. Specifically, system-level power consumption and chip area were estimated by accounting for the power and area overheads of peripheral circuits and array operations. The proposed AND-flash-based CIM operating in transistor-mode and capacitor-mode was compared with a conventional 1T–1RRAM-based CIM array at the 22 nm technology node under a ResNet-50 workload on the ImageNet dataset [75, 76]. The measured device characteristics of the AND-flash-based CIM in both transistor- and capacitor-modes were appropriately scaled to the 22 nm technology node

21983844, 2026, 56, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.76619 by Hanyang University Library, Wiley Online Library on [28/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

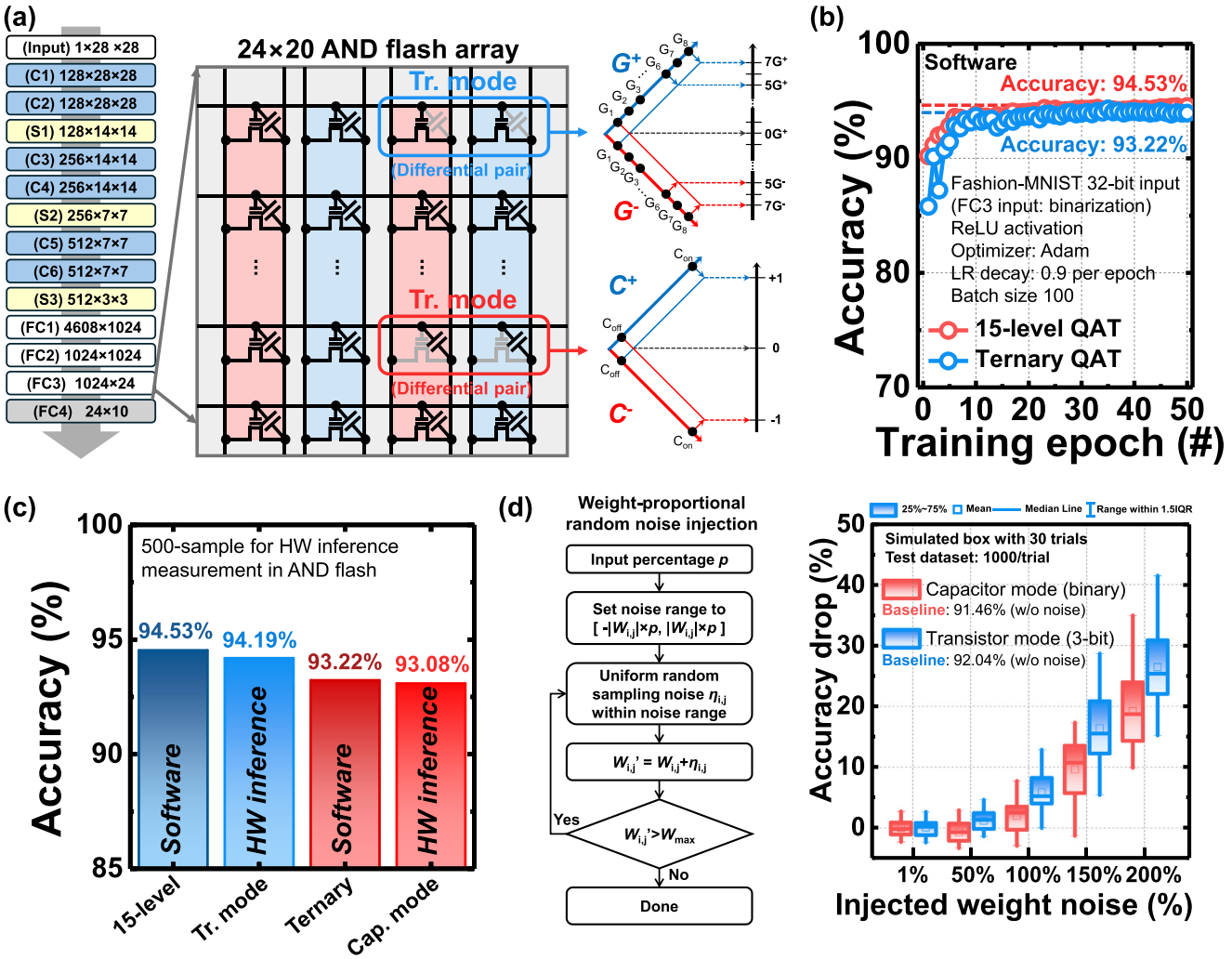

FIGURE 6 | Hardware-based neural network inference. (a) Architecture of the VGG-like CNN used for Fashion-MNIST classification and AND-type array-based hardware implementation. The final fully connected layer is mapped onto the fabricated AND-type flash array. (b) Software-level QAT results assuming different weight resolutions, where 15-level quantization is used for transistor-mode operation and ternary quantization is used for capacitor-mode operation. (c) Hardware inference accuracy obtained after transferring the trained weights onto the fabricated AND-type array and performing inference using 500 test samples, showing modest accuracy degradation compared to the software-level results. (d) Robustness evaluation under device-state variation, where random noise is injected into the mapped weights and the resulting inference accuracy is summarized as box plots. Capacitor-mode inference exhibits reduced sensitivity to weight perturbations compared to transistor-mode inference.

and incorporated into the NeuroSim framework. The device- and system-level parameters used in the simulations are summarized in Figure 7a. To evaluate energy efficiency, the throughput-per-power metric (TOPS/W) was extracted and compared with that of the 1T–1RRAM-based CIM, as shown in Figure 7b. Transistor-mode operation achieves a 2.20× improvement in TOPS/W, while capacitor-mode, benefiting from lower energy consumption per operation, provides a larger improvement of 3.22×. These results indicate that capacitor-mode is particularly well suited for energy-constrained and low-power applications.

Area efficiency was further evaluated using the throughput density metric (TOPS/mm^2), as summarized in Figure 7c. Compared to the 1T–1RRAM baseline, capacitor-mode achieves a 1.33× improvement in TOPS/mm^2, whereas transistor-mode exhibits a higher enhancement of 1.71×. Here, the identical physical array area refers specifically to the memory-cell array region based on the same AND-type flash cell and TANOS stack, rather than the total chip-level area including peripheral circuits. In practical implementations, capacitor-mode sensing may require additional peripheral circuits, such as charge/capacitance readout circuits, reference components, and SL isolation switches for floating-SL operation. Therefore, the reported area efficiency reflects the peripheral configuration assumed in the NeuroSim evaluation, while practical implementations may exhibit different area-efficiency trade-offs depending on the peripheral design. Under the assumed readout configuration, transistor-mode shows higher TOPS/mm^2 than capacitor-mode.

Importantly, as illustrated in Figure 7d, the analog-to-digital converter (ADC) energy consumption is significantly reduced in both modes, with a 6.83× reduction in capacitor-mode and a 2.43× reduction in transistor-mode compared to the 1T–1RRAM baseline. This substantial reduction in ADC energy contributes directly to the improved system-level energy efficiency of the proposed architecture. Taken together, these results demon-

21983844, 2026, 56, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.76619 by Hanyang University Library, Wiley Online Library on [28/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

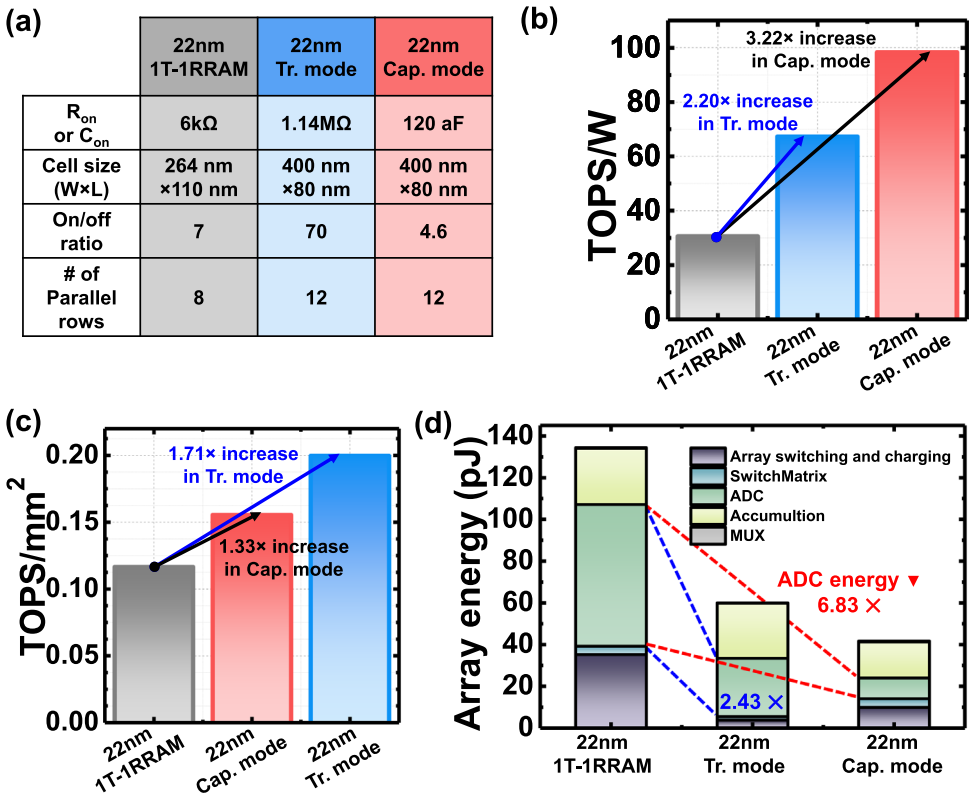

FIGURE 7 | System-level benchmarking of dual-mode AND-flash CIM. (a) Device- and system-level parameters used for array-level simulations in NeuroSim, comparing the proposed AND-flash-based CIM operating in both transistor- and capacitor-modes with a conventional 1T–1RRAM-based CIM at the 22 nm technology node under a ResNet-50 workload on the ImageNet dataset. (b) Energy efficiency comparison in terms of throughput per power (TOPS/W), showing improvements for both operating modes relative to the 1T–1RRAM baseline, with capacitor-mode achieving the highest energy efficiency. (c) Area efficiency comparison using throughput density (TOPS/mm^2), where transistor–mode achieves higher throughput density due to its higher computational speed, while capacitor-mode provides moderate improvement over the baseline. (d) Energy consumption of the proposed dual-mode CIM architecture, showing substantial energy reduction in both operating modes compared to the 1T–1RRAM baseline, with a more significant reduction in capacitor-mode.

strate that capacitor-mode, with its superior energy efficiency and reduced ADC overhead, is well suited for ultra-low-power operation, whereas transistor-mode provides higher throughput and accuracy for performance-critical workloads. These complementary characteristics enable workload-dependent mode selection within the same hardware platform. For example, transistor-mode operation can be employed for accuracy-critical tasks, whereas capacitor-mode operation can be utilized under energy-constrained conditions where robustness and low-power operation are prioritized. Such mode selection can be achieved without modifying the memory-cell structure, requiring only the corresponding biasing and readout configuration. This flexibility highlights the practicality of the proposed dual-mode CIM architecture and enables energy-adaptive operation according to workload requirements. Since both modes employ identical program and erase schemes, the write operation and its associated energy consumption remain unchanged regardless of the selected operating mode. Consequently, the benefits of the proposed dual-mode architecture primarily arise from the read operation and become most pronounced in read-dominant inference workloads employing off-chip training, where the programmed weights are reused over a large number of VMM operations.

5 | Conclusion

In this work, we demonstrated a dual-mode CIM architecture based on an AND-type flash array that supports both transistor-mode and capacitor-mode operation within a unified device structure. By selectively exploiting current-based sensing for high-precision computation and charge-based sensing for ultra-low-power operation, the proposed architecture enables distinct operating characteristics within the same memory-cell and array structure, with the operating mode determined by the corresponding bias and readout conditions. At the device and array levels, transistor-mode provides fine multilevel weight representation and accurate analog VMM, while capacitor-mode offers binary yet highly robust charge-domain operation with near-zero static power consumption. Hardware-based neural network inference confirms that transistor-mode achieves higher peak accuracy owing to its higher weight resolution, whereas capacitor-mode maintains competitive accuracy with significantly improved tolerance to device-state variation and noise. System-level benchmarking using NeuroSim further reveals that capacitor-mode delivers superior energy efficiency and substantially reduced ADC overhead, while transistor-mode achieves higher throughput density. These complementary characteristics

21983844, 2026, 56, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.76619 by Hanyang University Library, Wiley Online Library on [28/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

allow the same hardware platform to operate efficiently across a wide range of workloads, from energy-constrained edge inference to performance-critical computing tasks. Overall, the proposed dual-mode AND-flash CIM architecture provides a practical and scalable approach to workload-adaptive in-memory computing, enabling flexible trade-offs between accuracy, throughput, and energy efficiency. This work highlights how a unified flash memory platform can enable adaptive computing paradigms that dynamically balance accuracy, robustness, and energy efficiency within a single hardware platform.

6 | Experimental Section

6.1 | Electrical Measurements

The fabricated 48 × 24 AND-type CTF array was electrically characterized using a probe station equipped with a 96-channel probe card. The probe card provided access to 48 WLs, 24 SLs, and 24 DLs, which were connected to an E5250A low-leakage switching matrix. The switching matrix was interfaced with a semiconductor parameter analyzer (B1500A) equipped with source measure units (SMUs), a semiconductor pulse generator unit (SPGU), a multi-frequency capacitance measurement unit (MFCMU), and a WGFMU. The capacitance characteristics were measured at 10 kHz using the MFCMU of the B1500 parameter analyzer, where 10 kHz was selected as a standard frequency for *C–V* characterization. The E5250A switching matrix enabled flexible routing among the 96 channels, allowing selective connection of individual WLs, SLs, and DLs to the appropriate measurement units. The entire measurement system was controlled by a custom-developed C++ program, with both the semiconductor parameter analyzer and the switching matrix connected via a general-purpose interface bus (GPIB). Using this setup, the electrical characteristics of the AND-type CTF array were systematically evaluated in both transistor-mode and capacitor-mode.

6.2 | Neural Network Architecture and Hardware Mapping

A VGG-like convolutional neural network was employed for AND-type array-based hardware implementation using the Fashion-MNIST dataset consisting of 1 × 28 × 28 grayscale images. The network is composed of repeated convolutional blocks, each consisting of 3 × 3 convolution layers followed by batch normalization and ReLU activation, interleaved with 2 × 2 max-pooling layers. Specifically, the first convolutional block (C1, C2) processes feature maps with 128 channels at a spatial resolution of 28 × 28, followed by max pooling to down-sample the feature maps to 128 × 14 × 14. The second block (C3, C4) increases the channel dimension to 256 and operates at a 14 × 14 resolution, after which max pooling reduces the feature maps to 256 × 7 × 7. Subsequently, the third block (C5, C6) processes 512-channel feature maps, which are further reduced to a spatial resolution of 3 × 3. After flattening, the network includes four fully connected layers (FC1-FC4) with dimensions of 4608 × 1024, 1024 × 1024, 1024 × 24, and 24 × 10, respectively. Batch normalization and ReLU activation are applied after each fully connected layer except for the final output layer. To account for hardware implementation constraints, deterministic weight quantization was applied to all convolutional and fully connected layers during both training and inference, using 3-bit precision for transistor-mode operation and 1-bit precision for capacitor-mode operation. In addition, the input vector to the final classification layer (FC4) was binarized using a sign-based bipolar representation. The network was trained using the cross-entropy loss function for 100 epochs with the Adam optimizer and an initial learning rate of 0.01. The learning rate decayed by a factor of 0.9 at each epoch. For inference, the final fully connected layer with dimensions of 24 × 10 was physically mapped onto the AND-type CTF array. Weight transfer was performed using drain current levels for 3-bit weights in transistor-mode and device capacitance states for 1-bit weights in capacitor-mode.

Acknowledgements

This work was supported in part by the NRF funded by the Korean government (RS-2023-00260527, 20%, RS-2026-25488900, 20%), in part by the IITP (Institute for Information & Communications Technology Planning & Evaluation) grant funded by the Korean government (RS-2021-II212052, 10%, RS-2024-00506767, 50%), and in part by the Brain Korea 21 Four Program.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. M. Horowitz, "Computing's Energy Problem (and What we can do About it IEEE International Solid-State Circuits Conference (ISSCC) IEEE," 2014), 10–14.

2. J. Park, S. Kim, M. S. Song, et al., "Implementation of Convolutional Neural Networks in Memristor Crossbar Arrays With Binary Activation and Weight Quantization," *ACS Applied Materials & Interfaces* 16, no. 1 (2024): 1054–1065.

3. Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE Journal of Solid-State Circuits* 52, no. 1 (2016): 127–138, https://doi.org/10.1109/JSSC.2016.2616357.

4. V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," *Proceedings of the IEEE* 105, no. 12 (2017): 2295–2329, https://doi.org/10.1109/JPROC.2017.2761740.

5. S. Youn, J. Park, and H. Kim, "Energy-Efficient Ising Solver Implementations in Forming-Free Memristor Crossbar Arrays for Combinatorial Optimization Problems," *Nano Energy* 148 (2025): 111633.

6. A. Gholami, Z. Yao, S. Kim, C. Hooper, M. W. Mahoney, and K. Keutzer, "AI and Memory Wall," *IEEE Micro* 44, no. 3 (2024): 33.

7. P. Mannocci, M. Farronato, N. Lepri, et al., "In-Memory Computing with Emerging Memory Devices: Status and Outlook," *APL Machine Learning* 1 (2023): 1.

8. A. Kumar, M. Krishnaiah, J. Park, et al., "Multibit, Lead-Free Cs_2SnI_6 Resistive Random Access Memory with Self-Compliance for Improved Accuracy in Binary Neural Network Application," *Advanced Functional Materials* 34, no. 16 (2024): 2310780.

21983844, 2026, 56, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.76619 by Hanyang University Library, Wiley Online Library on [28/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

9. Y. S. An, D. Kim, Y. R. Park, et al., "Implementation of Monolithic 3D Integrated TiO Memristor-Based Neural Network for High-Performance in-Memory Computing," *Nano Energy* 139 (2025): 110999, https://doi.org/10.1016/j.nanoen.2025.110999.

10. H. Hwang, S. Youn, and H. Kim, "Recent Advances in Ferroelectric Materials, Devices, and In-Memory Computing Applications," *Nano Convergence* 12 (2025): 55, https://doi.org/10.1186/s40580-025-00520-2.

11. A. K. Rajput, A. K. Tiwari, and M. Pattanaik, "An Energy-Efficient Hybrid SRAM-Based In-Memory Computing Macro for Artificial Intelligence Edge Devices," *Circuits, Systems, and Signal Processing* 42, no. 6 (2023): 3589–3616, https://doi.org/10.1007/s00034-022-02284-0.

12. S. Youn, J. Lee, S. Kim, J. Park, K. Kim, and H. Kim, "Programmable Threshold Logic Implementations in a Memristor Crossbar Array," *Nano Letters* 24, no. 12 (2024): 3581.

13. Z. Sun, S. Kvatinsky, X. Si, A. Mehonic, Y. Cai, and R. Huang, "A Full Spectrum of Computing-in-Memory Technologies," *Nature Electronics* 6, no. 11 (2023): 823.

14. J. Park and H. Kim, "Physical Unclonable Function With 3D Stacked Memristor Crossbar Array Using Self-Differential Pair," *ACS Nano* 19, no. 31 (2025): 28135–28145, https://doi.org/10.1021/acsnano.4c18621.

15. A. Sebastian, M. L. Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory Devices and Applications for In-Memory Computing," *Nature Nanotechnology* 15, no. 7 (2020): 529–544, https://doi.org/10.1038/s41565-020-0655-z.

16. S. Youn, Y. Hwang, T.-H. Kim, et al., "Threshold Learning Algorithm for Memristive Neural Network With Binary Switching Behavior," *Neural Networks* 176 (2024): 106355, https://doi.org/10.1016/j.neunet.2024.106355.

17. K. Roy, I. Chakraborty, M. Ali, A. Ankit, and A. Agrawal, "In-Memory Computing in Emerging Memory Technologies for Machine Learning: An Overview," *ACM/IEEE Design Automation Conference (DAC)* (IEEE, 2020): 1–6.

18. J. Bae, J. Won, and W. Shim, "The Rise of Memtransistors for Neuromorphic Hardware and In-Memory Computing," *Nano Energy* 126 (2024): 109646, https://doi.org/10.1016/j.nanoen.2024.109646.

19. K. Kim, M. S. Song, H. Hwang, S. Hwang, and H. Kim, "A Comprehensive Review of Advanced Trends: From Artificial Synapses to Neuromorphic Systems With Consideration of Non-Ideal Effects," *Frontiers in Neuroscience* 18 (2024): 1279708, https://doi.org/10.3389/fnins.2024.1279708.

20. Y. Wang, X. Yang, Y. Qin, et al., "A 28nm 83.23 TFLOPS/W POSIT-Based Compute-in-Memory Macro for High-Accuracy AI Applications,"*IEEE International Solid-State Circuits Conference (ISSCC) IEEE* (2024): 566–568.

21. G. Kim, S. Youn, J. Park, and H. Kim, "Configurable Kernel Map Implementation in Memristor Crossbar for Convolution Neural Network," *Advanced Intelligent Systems* (2025): 2500032, https://doi.org/10.1002/aisy.202500032.

22. C.-X. Xue, Y.-C. Chiu, T.-W. Liu, et al., "A CMOS-Integrated Compute-in-Memory Macro Based on Resistive Random-Access Memory for AI Edge Devices," *Nature Electronics* 4, no. 1 (2021): 81.

23. T.-H. Kim, S. Kim, J. Park, S. Youn, and H. Kim, "Memristor Crossbar Array with Enhanced Device Yield for In-Memory Vector–Matrix Multiplication," *ACS Applied Electronic Materials* 6, no. 6 (2024): 4099.

24. S. Song, J. Hanamshet, A. Balaji, et al., "Dynamic Reliability Management in Neuromorphic Computing," *ACM Journal on Emerging Technologies in Computing Systems* 17, no. 4 (2021): 1.

25. O. Yousuf, B. D. Hoskins, K. Ramu, et al., "Layer Ensemble Averaging for Fault Tolerance in Memristive Neural Networks," *Nature Communications* 16, no. 1 (2025): 1250.

26. W. Dong, K. Zhou, and Z. Kong, "Efficient Calibration for RRAM-Based In-Memory Computing Using DoRA," (2025), arXiv preprint arXiv:2504.03763.

27. Y. Park, E. Lim, S. Lee, V. Georgiev, S. Kim, and S. Kim, "Ferroelectric Memristor Crossbar Arrays for Highly Integrated Neuromorphic Computing System," *Nano Energy* 141 (2025): 111137, https://doi.org/10.1016/j.nanoen.2025.111137.

28. Y. Xiang, P. Huang, and Z. Zhou, "Analog Deep Neural Network Based on NOR Flash Computing Array for High Speed/Energy Efficiency Computation," *IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2019), 1–4.

29. M. Kim, M. Liu, L. R. Everson, and C. H. Kim, "An Embedded NAND Flash-Based Compute-in-Memory Array Demonstrated in a Standard Logic Process," *IEEE Journal of Solid-State Circuits* 57, no. 2 (2021): 625.

30. D. Yu, S. Ahn, H. Hwang, and H. Kim, "Physical Unclonable Function with Enlarged String Current Variation in NAND Flash Array by WL Selection Scheme," *IEEE Transactions on Electron Devices* 72, no. 10 (2025): 5400–5406.

31. Y. J. Park, H. T. Kwon, B. Kim, et al., "3-D Stacked Synapse Array Based on Charge-Trap Flash Memory for Implementation of Deep Neural Networks," *IEEE Transactions on Electron Devices* 66, no. 1 (2018): 420–427, https://doi.org/10.1109/TED.2018.2881972.

32. D. Kwon, S. Y. Woo, K.-H. Lee, et al., "Reconfigurable Neuromorphic Computing Block Through Integration of Flash Synapse Arrays and Super-Steep Neurons," *Science Advances* 9, no. 29 (2023): adg9123.

33. H. Kim, S. Hwang, J. Park, S. Yun, J.-H. Lee, and B.-G. Park, "Spiking Neural Network Using Synaptic Transistors and Neuron Circuits for Pattern Recognition with Noisy Images," *IEEE Electron Device Letters* 39, no. 4 (2018): 630–633.

34. H.-S. Choi, Y. J. Park, J.-H. Lee, and Y. Kim, "3-D Synapse Array Architecture Based on Charge-Trap Flash Memory for Neuromorphic Application," *Electronics* 9, no. 1 (2019): 57.

35. H. Kim, J. Park, M.-W. Kwon, J.-H. Lee, and B.-G. Park, "Silicon-Based Floating-Body Synaptic Transistor With Frequency-Dependent Short- and Long-Term Memories," *IEEE Electron Device Letters* 37, no. 3 (2016): 249–252, https://doi.org/10.1109/LED.2016.2521863.

36. W. Huang, H. Zhu, J. Li, et al., "A Novel 3D nor Flash with Single-Crystal Silicon Channel: Devices, Integration, and Architecture," *IEEE Electron Device Letters* 43, no. 11 (2022): 1874.

37. H. Hwang, G. Kim, D. Yu, and H. Kim, "Wordline Input Bias Scheme for Neural Network Implementation in 3D-NAND Flash," *Biomimetics* 10, no. 5 (2025): 318.

38. S.-T. Lee, H. Kim, J.-H. Bae, et al., "High-Density and Highly-Reliable Binary Neural Networks Using NAND Flash Memory Cells as Synaptic Devices," *IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2019).

39. H.-T. Lue, P.-K. Hsu, M.-L. Wei, et al., "Optimal Design Methods to Transform 3D NAND Flash into a High-Density, High-Bandwidth and Low-Power Nonvolatile Computing in Memory (nvCIM) Accelerator for Deep-Learning Neural Networks (DNN)," *IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2019).

40. P. Wang, F. Xu, B. Wang, et al., "Three-Dimensional nand Flash for Vector–Matrix Multiplication," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 27, no. 4 (2018): 988–991, https://doi.org/10.1109/TVLSI.2018.2882194.

41. G. H. Lee, M. S. Song, S. Kim, et al., "Ferroelectric Field-Effect Transistors for Binary Neural Network with 3-D NAND Architecture," *IEEE Transactions on Electron Devices* 69 11 (2022): 6438.

42. W. Shim and S. Yu, "Technological Design of 3D NAND-Based Compute-in-Memory Architecture for GB-Scale Deep Neural Network," *IEEE Electron Device Letters* 42, no. 2 (2020): 160–163, https://doi.org/10.1109/LED.2020.3048101.

43. S.-T. Lee and J.-H. Lee, "Review of Neuromorphic Computing Based on NAND Flash Memory," *Nanoscale Horizons* 9 (2024): 1475.

44. M. Oda, K. Sakamawari, S. Seno, et al., "Superior Scalability of Advanced Horizontal Channel Flash for Future Generations of 3D Flash

21983844, 2026, 56, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.76619 by Hanyang University Library, Wiley Online Library on [28/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Memory," *IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2024), 1–4.

45. M. S. Song, H. Hwang, J. Yu, S. Hwang, and H. Kim, "Annealing Effects on Charge Trap Flash with TAHOS Structure," *Journal of Semiconductor Technology and Science* 24, no. 4 (2024): 323.

46. T.-H. E. Yeh, W.-C. Chen, T.-H. B. Hsu, et al., "Z-Interference and Z-Disturbance in Vertical Gate-Type 3-D NAND," *IEEE Transactions on Electron Devices* 63, no. 3 (2016): 1047–1053, https://doi.org/10.1109/TED.2016.2520965.

47. I.-S. Lee, H. Kim, M.-K. Park, et al., "Power and Area-Efficient XNOR-AND Hybrid Binary Neural Networks Using TFT-Type Synaptic Device," *IEEE Electron Device Letters* 44, no. 2 (2022): 325.

48. W.-M. Kang, D. Kwon, S. Y. Woo, et al., "Hardware-Based Spiking Neural Network Using a TFT-Type AND Flash Memory Array Architecture Based on Direct Feedback Alignment," *IEEE Access* 9 (2021): 73121–73132, https://doi.org/10.1109/ACCESS.2021.3080310.

49. R. Han, Y. Xiang, P. Huang, Y. Shan, X. Liu, and J. Kang, "Flash Memory Array for Efficient Implementation of Deep Neural Networks," *Advanced Intelligent Systems* 3, no. 5 (2021): 2000161, https://doi.org/10.1002/aisy.202000161.

50. M. M. Dahan, E. T. Breyer, S. Slesazeck, T. Mikolajick, and S. Kvatinsky, "C-AND: Mixed Writing Scheme for Disturb Reduction in 1T Ferroelectric FET Memory," *IEEE Transactions on Circuits and Systems I: Regular Papers* 69, no. 4 (2022): 1595.

51. H.-S. Choi, H. Kim, J.-H. Lee, B.-G. Park, and Y. Kim, "AND Flash Array Based on Charge Trap Flash for Implementation of Convolutional Neural Networks," *IEEE Electron Device Letters* 41, no. 11 (2020): 1653–1656, https://doi.org/10.1109/LED.2020.3025587.

52. S. Lee, H. Kim, S.-T. Lee, B.-G. Park, and J.-H. Lee, "SiO_2 Fin-Based Flash Synaptic Cells in AND Array Architecture for Binary Neural Networks," *IEEE Electron Device Letters* 43, no. 1 (2021): 142.

53. S. Kim, J. Y. Park, D. K. Lee, et al., "Demonstration of Bias Scheme for Ferroelectric Field-Effect Transistor (FeFET) Based AND Array Operation," *Solid-State Electronics* 216 (2024): 108917, https://doi.org/10.1016/j.sse.2024.108917.

54. K.-H. Lee, D. Kwon, S. Y. Woo, et al., "Highly Linear Analog Spike Processing Block Integrated With an AND-Type Flash Array and CMOS Neuron Circuits," *IEEE Transactions on Electron Devices* 69, no. 11 (2022): 6065.

55. D. Kwon, M.-K. Park, W.-M. Kang, et al., "Hardware-Based Ternary Neural Network using AND-type Poly-Si TFT Array and its Optimization Guideline," *IEEE Transactions on Electron Devices* 70, no. 8 (2023): 4206.

56. W. Shin, J. Im, R. H. Koo, et al., "Self-Curable Synaptic Ferroelectric FET Arrays for Neuromorphic Convolutional Neural Network," *Advanced Science* 10, no. 15 (2023): 2207661.

57. J. Kim, J. Im, S. Oh, et al., "Vertical AND-Type Flash Synaptic Cell Stack for High-Density and Reliable Binary Neural Networks," *IEEE Electron Device Letters* 45, no. 7 (2024): 1369.

58. G. Choe, A. Lu, and S. Yu, "3D AND-Type Ferroelectric Transistors for Compute-in-Memory and the Variability Analysis," *IEEE Electron Device Letters* 43, no. 2 (2021): 304.

59. H.-T. Lue, W. Chen, H.-S. Chang, K.-C. Wang, and C.-Y. Lu, "A novel 3D AND-type NVM Architecture Capable of High-Density, Low-Power In-Memory Sum-of-Product Computation for Artificial Intelligence Application," *IEEE Symposium on VLSI Technology* (IEEE, 2018), 177–178.

60. Y.-T. Seo, D. Kwon, Y. Noh, et al., "3-D AND-Type Flash Memory Architecture with High-κ Gate Dielectric for High-Density Synaptic Devices," *IEEE Transactions on Electron Devices* 68, no. 8 (2021): 3801.

61. H.-T. Lue, G.-R. Lee, T.-H. Yeh, et al., "3D AND: A 3D Stackable Flash Memory Architecture to Realize High-Density and Fast-Read 3D NOR Flash and Storage-Class Memory," *IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2020).

62. J. Im, J. Kim, J. Hwang, et al., "Vertical AND-Type Flash TFT Array Capable of Accurate Vector-Matrix Multiplication Operations for Hardware Neural Networks," *IEEE Electron Device Letters* 45, no. 7 (2024): 1385.

63. W. Shim, H. Jiang, X. Peng, and S. Yu, "Architectural Design of 3D NAND Flash Based Compute-in-Memory for Inference Engine," *International Symposium on Memory Systems (MEMSYS)* (IEEE, 2020), 77–85.

64. J. Yu, H. Hwang, H. Kim, and W. Y. Choi, "Vertically Stackable Memcapacitor Crossbar Array Based on NAND Flash Array Structure," *IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2024), 1–4.

65. K.-U. Demasius, A. Kirschen, and S. Parkin, "Energy-Efficient Memcapacitor Devices for Neuromorphic Computing," *Nature Electronics* 4, no. 10 (2021): 748.

66. S. Hwang, J. Yu, M. S. Song, H. Hwang, and H. Kim, "Memcapacitor Crossbar Array with Charge Trap NAND Flash Structure for Neuromorphic Computing," *Advanced Science* 10, no. 32 (2023): 2303817.

67. J. Sun, H. Zhou, Z. Wang, and Y. Wang, "Memcapacitor-Based Operant Conditioning Neural Network with Deprivation and its Application in Inspection Robots," *IEEE Transactions on Industrial Informatics* 21, no. 10 (2025): 7587–7597.

68. S. Ahn, J. Yu, H. Hwang, et al., "Coupling-Free Readout Scheme for Memcapacitors with NAND Flash Structure," *IEEE Transactions on Electron Devices* 71, no. 8 (2024): 4670.

69. J. Hur, Y.-C. Luo, A. Lu, et al., "Nonvolatile Capacitive Crossbar Array for In-Memory Computing," *Advanced Intelligent Systems* 4, no. 8 (2022): 2100258, https://doi.org/10.1002/aisy.202100258.

70. T.-H. Kim, O. Phadke, Y.-C. Luo, et al., "Tunable Non-Volatile Gate-to-Source/Drain Capacitance of FeFET for Capacitive Synapse," *IEEE Electron Device Letters* 44, no. 10 (2023): 1628.

71. H. Hwang, J. Yu, S. Youn, W. Y. Choi, and H. Kim, "2C-Ternary Content Addressable Memory in Memcapacitor Crossbar Array With NAND Flash Structure," *Small* 21, no. 9 (2025): 2408618, https://doi.org/10.1002/smll.202408618.

72. K.-D. Suh, B.-H. Suh, Y.-H. Lim, et al., "A 3.3 V 32 Mb NAND Flash Memory with Incremental Step Pulse Programming Scheme," *IEEE Journal of Solid-State Circuits* 30, no. 11 (1995): 1149.

73. W. Wan, R. Kubendran, C. Schaefer, et al., "A Compute-in-Memory Chip Based on Resistive Random-Access Memory," *Nature* 608, no. 7923 (2022): 504.

74. Q. Liu, B. Gao, P. Yao, et al., "A Fully Integrated Analog ReRAM Based 78.4 TOPS/W Compute-in-Memory Chip with Fully Parallel MAC Computing," *IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2020), 500–502.

75. W. Li, X. Sun, S. Huang, H. Jiang, and S. Yu, "A 40-nm MLC-RRAM Compute-in-Memory Macro with Sparsity Control, on-chip Write-Verify, and Temperature-Independent ADC References," *IEEE Journal of Solid-State Circuits* 57, no. 9 (2022): 2868.

76. J.-M. Hung, T.-H. Wen, Y.-H. Huang, et al., "8-b Precision 8-Mb ReRAM Compute-in-Memory Macro Using Direct-Current-Free Time-Domain Readout Scheme for AI Edge Devices," *IEEE Journal of Solid-State Circuits* 58, no. 1 (2022): 303.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supporting file: advs76619-sup-0001-SuppMat.docx

21983844, 2026, 56, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.76619 by Hanyang University Library, Wiley Online Library on [28/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License