Nonvolatile Capacitive Crossbar Array for In-Memory Computing

Jae Hur, Yuan-Chun Luo, Anni Lu, Tzu-Han Wang, Shaolan Li, Asif Islam Khan, and Shimeng Yu*

Conventional resistive crossbar array for in-memory computing suffers from high static current/power, serious IR drop, and sneak paths. In contrast, the “capacitive” crossbar array that harnesses transient current and charge transfer is gaining attention as it 1) only consumes dynamic power, 2) has no DC sneak paths and avoids severe IR drop (thus, selector-free), and 3) can be fabricated on top of complementary metal–oxide–semiconductor (CMOS) circuits for 3D-stacking. For the first time, ferroelectric Hf0.5Zr0.5O2 (HZO) capacitive crossbar arrays are experimentally demonstrated. The vector matrix multiplication (VMM) experiments are conducted on the fabricated capacitive crossbar array, showing a linear weighted sum versus numbers of input or on-state weight. The array-level VMM operation could maintain weight pattern reprogramming after 1) thousands of 1 ms/3 V pulses and 2) an extrapolated 10-year retention at 85 °C. Array-level circuit simulation at 22 nm node shows the energy consumption of a capacitive crossbar array is 20–200× lower than the resistive crossbar array counterpart. Moreover, analog-shift-and-add circuits are designed for multibit weight summation, achieving 16.6% less area and 26.9% lower energy consumption than digital-shift-and-add circuits.

1. Introduction

Resistive crossbar array has been intensively studied for in-memory computing,[1] where the parallel VMM shows significantly faster speed and higher energy efficiency when compared

J. Hur, Y.-C. Luo, A. Lu, T.-H. Wang, S. Li, A. I. Khan, S. Yu
School of Electrical and Computer Engineering
Georgia Institute of Technology
791 Atlantic Dr., Atlanta, GA 30332, USA
E-mail: shimeng.yu@ece.gatech.edu
A. I. Khan
School of Materials Science and Engineering
Georgia Institute of Technology
771 Ferst Dr. NW, Atlanta, GA 30332, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202100258.

© 2022 The Authors. Advanced Intelligent Systems published by Wiley-VCH GmbH. 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.

DOI: 10.1002/aisy.202100258
Theoretically, a perfect ferroelectric capacitor should not exhibit capacitance window at DC zero bias (with displacement charges), but only show capacitance window in transient sweep that involves polarization switching. Therefore, interface engineering is required to break the symmetry and open up the nonzero capacitance window at DC zero bias. Prior work on the HfO$_2$-based capacitive synapse\textsuperscript{[7]} needs a DC bias at $\gtrsim 1.5$ V to open the capacitance window, which is not realistic due to the potential read-disturb. In addition, the difference between its high- and low-capacitance state is as low as $\lesssim 1\%$ in the study by Zheng et al.\textsuperscript{[7]} In this work, we demonstrate a HZO capacitive synapse device with $>110\%$ capacitance window, which is free from read-disturb by reading the states at DC zero bias. Here, for the first time, we experimentally integrated the HZO capacitors into a crossbar array structure and set up a measurement system for performing VMM based on the charge transfer mechanism. The experimental results show that the output voltage has high linearity versus both the inputs and weight values. Through circuit simulation, array-level and system-level benchmarking results indicate substantial benefits of capacitive crossbar array over the resistive counterparts at advanced nodes.

Finally, we propose to shift-and-add the weighted sum for multi-bit weight in an analog manner instead of a conventional digital manner to further reduce area and energy consumption.

2. Device Characteristics of Nonvolatile Capacitive Synapse

This section presents the small-signal capacitance–voltage (C–V) characteristics of the single devices from the fabricated HZO crossbar capacitor array. The process flow of the crossbar array can be found in the Experimental Section.

The asymmetric “small-signal” C–V characteristics in Figure 2a show high- and low-capacitance states (HCS/LCS) at DC 0 V. The dielectric constant ($\varepsilon_r$) represents the capacitance states. In our previous work,\textsuperscript{[8]} the small-signal capacitance asymmetry was attributed to excessive oxygen vacancies at the bottom electrode (BE) interface, which induces the domain wall (DW) pinning effect (down-polarized even after positive sweep), resulting in more DWs and thus higher small-signal capacitance. The pinned up-polarized domains do not create DWs as $3$ V pulses tend to flip the domains down-polarized too, resulting in lower capacitance.

Figure 2. a) Asymmetric small-signal C–V shows a memory window at DC 0 V with 10 kHz 100 mV AC small signals applied. b) Physical illustration of the asymmetric C–V at DC 0 V. The positively charged oxygen vacancies (Vo$^{2+}$) at the BE pinned some domains up-polarized. b) While $+3$ V pulses tend to flip the domains down-polarized, more DWs are formed due to the pinned domains, resulting in higher small-signal capacitance. The pinned up-polarized domains do not create DWs as $-3$ V pulses tend to flip the domains up-polarized too, resulting in lower capacitance.
To verify this capacitor as a read-disturb-free and programmable capacitive memory, we further performed small-signal AC measurement at DC = 0 V directly after program and erase with +3 and −3 V write pulses, respectively (Figure 3a), where the cycling endurance shows steady window even after thousands of strong 3 V with 1 ms programming pulses. For inference engine, weight programming is relatively infrequent; thus, 1000 cycles endurance cycles is sufficient for its application. In Figure 3b, a practical read operation applies a 100 mV pulse and integrates the current flowing onto the capacitor. The integrated charges show a distinct margin and an on/off ratio of 113%, similarly as the initial capacitance window in Figure 3a. This is because integrated charge is equal to capacitance × small-signal input voltage. It should be noted that the downward-shifting capacitances over endurance cycles can be a minor issue in an inference-only application. Although the weights need to be occasionally adjusted after deployment, a dummy column of the capacitors or on-chip sensing circuits can be applied to shift the reference voltage in the sense amplifiers correspondingly.

Finally, even though capacitive crossbar array is immune to sneak paths, the half-select write-disturb is still a potential concern. We analyzed the effect of write-disturb under 1/3 write voltage ($V_{\text{w}}$) scheme. The disturbance compared with the total margin is shown in Figure 4a, where the capacitance change (disturbance) is <10% of the total margin with $V_{\text{w}} = 2.5$ V or 3 V.

![Figure 3](image-url)  
**Figure 3.** a) Capacitive memory window measured directly at DC 0 V without sweeping. Distinct memory window exists even after thousands of strong 3 V/1 ms pulses. b) Capacitive memory window measured in displacement charges integrated from current flowing onto the capacitor under small 100 mV input pulse.

![Figure 4](image-url)  
**Figure 4.** a) Small write disturbance compared with the total margin b) under 1/3 $V_{\text{w}}$ scheme and c) programming protocols and the definition of disturbs of LCS and HCS. The measured capacitor size is $200 \times 200 \mu m^2$. 

(c) 2022 The Authors. Advanced Intelligent Systems published by Wiley-VCH GmbH.
after programming stress. The 1/3 $V_w$ scheme is illustrated in Figure 4b and is described as follows. The selected cell experiences the total $V_w$ drop on the capacitor whereas the un-selected cells experience 1/3 $V_w$. The programming protocol for the analysis of the write disturbance in Figure 4a is shown in Figure 4c. We first precycled the device with a hundred of 3 V/1 ms pulses. Then, we applied $V_w$ and measured the small-signal capacitance without disturbance. After that, we applied $V_w$ followed by applying 1/3 $V_w$ with opposite polarity and finally measured the small-signal capacitance with the disturbance. The “disturb” is calculated as the percentage difference between the capacitances with and without disturbance. There were no reset pulses between the measurement of the capacitances with and without disturbance.

3. Array-Level Measurement

We fabricated a small-scale 12 $\times$ 12 HZO-based capacitive crossbar array to demonstrate the in-memory computing functionality as a proof of concept. Figure 5 shows the microscopic image of the fabricated array. The active size of each cross-point cell ranges from 1 $\times$ 1 $\mu$m$^2$ to 100 $\times$ 100 $\mu$m$^2$. The details of the fabrication process are discussed in Experimental Section.

Before moving to the array-level measurement, we need to understand the operating principle of the capacitive array. Figure 6 illustrates the basic concept of the operating principle in two steps. In the first step (Figure 6a), the WL voltages represented by step functions with amplitude $=$ 100 mV as the input vector propagate through WL multiplexer (MUX) and charge the array of ferroelectric capacitance ($C_{FE}$), which is preprogrammed to different capacitances to represent the values in the weight matrix. The product of one input voltage value and one weight capacitance value is encoded as the charges on each $C_{FE}$. In the second step (Figure 6b), the input voltages return to the common-mode voltage ($V_{CM}$) and become the same as the negative input of the operational amplifiers (OPAMPs). At this moment, the voltage drop on each $C_{FE}$ becomes 0 V so the charges are forced to transfer along the corresponding BL onto the reference capacitor ($C_{ref}$). The number of charges on $C_{ref}$ is the weighted sum along the BL and the resulting output voltage ($V_{out}$) determines an entry in the output vector. The ideal analytical equation for weighted sum is as follows.

$$V_{out} = \frac{\sum_{ij} V_{WL_j} \times C_{FE,ij}}{C_{ref}}$$

where $i$ and $j$ represent the number of rows and columns.

In the following experiment, we first demonstrate VMM operations with eight capacitive synapses in one column with all the inputs turned on. The detailed description of the experiment setup is provided in Figure 11. Figure 7a shows the measured weighted sum, $V_{out}$, sensed at the output node of the OPAMP, over 12 measurement trials. The array consists of eight number of 50 $\times$ 50 $\mu$m$^2$ ferroelectric capacitive synaptic cells with all inputs being “1.” The tight distribution in Figure 7a implies low cycle-to-cycle variation. Average $V_{out}$ in Figure 7b shows high linearity versus the number of HCS cells. Arrays with smaller synapses are also demonstrated in Figure 7c,d. Compared with Figure 7b, Figure 7c shows a smaller output swing because the overall charges are reduced with a smaller unit cell area, 25 $\times$ 25 $\mu$m$^2$ with the same $C_{ref} = 22$ pF. To further reduce the unit cell area down to 10 $\times$ 10 $\mu$m$^2$ and maintain output swing $>$ 100 mV, $C_{ref}$ needs to be decreased according to 1 pF, as shown in Figure 7d.

![Figure 5](https://www.advancedsciencenews.com)

**Figure 5.** Microscopic photo of the fabricated capacitive array with unit cell area $=$ 10 $\times$ 10 $\mu$m$^2$ with zoomed-in image of the core crossbar cells. The BE is where the WL input voltages are applied, while the TE serves as columns where output charges flow.

![Figure 6](https://www.advancedsciencenews.com)

**Figure 6.** Illustration of VMM operations with a capacitive crossbar array: a) phase I: charging $C_{FE}$, where the charges on each $C_{FE}$ represent one entry of product in the VMM operation and b) phase II: transferring charges in one BL to the corresponding $C_{ref}$, where the resulting $V_{out}$ represents an entry in the output vector.
A highly linear relationship between \( V_{\text{out}} \) and the number of turned-on WLs is also proven (Figure 8a) with ultratight distribution (Figure 8b). In practice, it is challenging to measure sub-10 \( \mu \text{m} \) capacitor’s response due to parasitics and sensing limit in any off-chip instrument. Therefore, the performance of nanoscale capacitive crossbar arrays with integrated peripheral circuits is projected by simulations in Section 4.

Moreover, reliability tests of the capacitive crossbar array were performed in terms of endurance and retention characteristics at the array level by monitoring \( V_{\text{out}} \). Figure 9a shows the cycling endurance with 3 V/1 ms pulses to reprogram the weight pattern. Even after thousands of such pulses, a sense margin at \( V_{\text{out}} \) still exists. Figure 9b shows the 15-hour retention at 85°C, where a clear \( V_{\text{out}} \) sense margin can be extrapolated to 10 years. Decreasing \( V_{\text{out}} \) in both HCS and LCS over time implies a decreasing capacitance. This might be a result of the imprint effect that is commonly observed in ferroelectrics. As shown in Figure 9c, hours after being programmed to HCS, the small-signal \( C-V \) curve shifts to the negative side, resulting in a lower HCS capacitance. On the other hand, after erased to LCS, the positive-side-shifting \( C-V \) over time (Figure 9d) causes the LCS capacitance to decrease during the retention test.

4. Simulations Toward Large-Scale Systems

To evaluate the latency, energy, and equivalent number of bits (ENOB) under thermal noise of the nonvolatile capacitive array at an advanced technology node, we ran simulation program with integrated circuit emphasis (SPICE) simulation with 22 nm low-power (LP) transistor models by considering wire parasitics. The array schematic and the OPAMP circuits are shown in Figure S1a,b, Supporting Information. The array size and other key parameters in the simulation are listed in Table S1, Supporting Information.

The SPICE simulation results showed an ENOB = 4.3 in Figure S1c, Supporting Information, and latency for the charge transfer is \( \approx 5.1 \text{ ns} \), where latency is defined as the time for \( V_{\text{out}} \) to reach 80% of the steady-state voltage. The ENOB value larger than 4 is achievable, suggesting a 4-bit partial sum quantization, which has been reported to keep reasonable inference accuracy for
Figure 8. Measured output voltage versus number of turned-on inputs with all devices programmed to HCS. a) Scatter plot of 12 measurements for each number of turned-on inputs. b) Average value over 12 consecutive readout.

Figure 9. a) Array-level pulse cycling endurance with all devices programmed to HCS or LCS. 3 V/1 ms pulses are applied before reading the weighted sum, \( V_{\text{out}} \), at end of the column. Even after thousands of such strong write pulses, a distinct memory window still survives, suggesting its feasibility as an inference engine. b) Array-level \( V_{\text{out}} \) retention at 85°C with all devices programmed to HCS or LCS. c) Imprint effect after potentiation pulses leads to \( C-V \) curve shifting left, resulting in HCS decreasing over time. d) Imprint effect after depression pulses leads to \( C-V \) curve shifting right, resulting in LCS decreasing over time.
5. Parallel Processing with Analog-Shift-and-Add

Weights in deep learning algorithms are typically quantized before being mapped to a synaptic array. Conventional hardware implementation of n-bit quantized weight groups n binary synaptic cells together to represent an n-bit weight value. The n cells in one group are in adjacent columns in the same row, where the n outputs of the weighted sums need to be combined later by peripheral circuits based on the bit significance of each cell representation. A common setting of the peripheral circuits is shown in Figure 10a, where MUX selects one bit-line (BL) of the synaptic array among n BLs (assume n-bit weights are mapped to the hardware). Here, n is equal to 4 in this work. The selected BL value is then transformed from an analog representation to a digital representation by the following ADC. The n digital values are processed sequentially and summed based on their bit significance by a digital-shift-and-add circuit.

However, this setting of sequential processing with digital-shift-and-add requires n cycles to combine the weighted sums, which consume n cycles of energy. The sequential processing also requires an array of MUX to select one BL at a time for the input of ADC, which becomes additional area overhead.

To improve the energy and area efficiency, we propose to sum the n BLs with different bit significances in parallel in one cycle using an analog-shift-and-add circuit instead of the digital case. The high-level schematic is shown in Figure 10b, where we can sum up the n BLs in parallel with an analog-shift-and-add circuit and feed the output to an ADC.[13] The parallel processing avoids the need of using MUX and saves the energy and latency by reducing the processing time from n cycles to a single cycle.

The circuits of analog-shift-and-add are shown in Figure 11a. The array of analog voltage buffer serves two purposes. The first purpose is to shield the charges at the outputs of the OPAMPs from leaking, which will lead to significant information loss due to the voltage drop. The second one is to downshift the offset voltage so that the total charges transferred to the next stage can be reduced. Here, the smaller number of charges can result in lower latency. The output voltage of the buffer representing the weighted sum from each BL is connected to an array of capacitors. The voltage representing the least significant bit

| Capacitive [this work] | 1 T-1RRAM[8] | 1 T-1MRAM[11] | 1 T-1FeFET[12] |
|------------------------|--------------|----------------|----------------|
| C_on/R_on [fF Ω⁻¹]    | 120          | 6 k            | 2.5 k          |
| On/off ratio           | 1.125        | 17             | 2.8            |
| Bit/cell               | 1            | 1              | 1              |
| Energy [pJ]            | 0.96         | 88             | 193            |
| Delay [ns]             | 5.1          | <5             | <5             |

Based on the array-level SPICE simulation result, we benchmarked the capacitive crossbar array with the open-source simulator DNN + NeuroSim[10] at 22 and 7 nm to evaluate its system-level performance in Table 2. The simulation setting for NeuroSim is described in Supporting Information. With the capacitive array assumed on top of the peripheral circuits, it is benchmarked with 22 nm 1 T-RRAM, 1 T-1MRAM, 1 T-1FeFET, and 22 nm/7 nm static random-access memory (SRAM). The benchmarking results of the 22 nm capacitive array show a higher energy efficiency and compute efficiency than those of the other resistive counterparts and 22 nm SRAM array. The projection of a 7 nm capacitive array also shows a substantial energy efficiency boost over the 7 nm SRAM.

### Table 1. Subarray-level evaluation with SPICE simulation, compared with those of the representative crossbar arrays obtained using DNN + NeuroSim framework. The energy and latency of the resistive arrays include those from the crossbar structure, while the energy and latency of the capacitive array include those of the crossbar structure and OPAMPs. The results are simulated and averaged assuming all input turned on and weight patterns from a pretrained VGG-8 model. Delay of the capacitive array is defined as the output voltage reaches 80% of the steady-state value. (Subarray energy includes the static and dynamic energy of the crossbar structure. Subarray energy is normalized to 1-bit VMM.)

### Table 2. System-Level benchmarking results show that the capacitive array has the potential of outperforming its resistive counterparts and can be more competitive over 7 nm SRAM. (subarray size = 128 × 128; F = 7 nm or 22 nm for normalizing cell area; thus, it does not indicate the physical feature size. The 7 nm projection of “capacitive” applies 7 nm peripheries while keeping the same cell area as the 22 nm one.)
(LSB) is connected to the capacitor with unit capacitance $C$, while the voltage representing the most significant bit (MSB) is connected to the capacitor with capacitance $2^{n-1} C$. In this way, the MSB voltage can be correctly weighted and results in a larger corresponding number of charges compared to the one with LSB voltage. After the voltages values are successfully weighted based on their bit significance, the charges will be transferred to the summed-up capacitance ($C_{\text{sum}}$) shunting of an OPAMP with its output connected to an ADC. The value of $C_{\text{sum}}$ should be designed based on the value of $n$ (n-bit weight) and the desired output voltage range. It should be pointed out that the capacitive array for the analog-shift-and-add requires voltage as input, which is compatible with the output of the capacitive crossbar array. In contrast, a resistive array which outputs current requires additional circuits to transform the current to voltage with additional area and energy overhead.

To evaluate the advantages based on the analog-shift-and-add at the system level, we extracted the energy and latency from SPICE simulation and estimated the area of the analog-shift-and-add circuit, from the voltage buffers to the output of the OPAMP shunting $C_{\text{sum}}$. In the system-level evaluation, we compare the performance between designs based on digital- and analog-shift-and-add. Both systems apply a 4-BL-sharing-1-ADC setting and an array size of $128 \times 128$. The unit capacitance for the analog-shift-and-add is 100 aF. For the ADC precision in the context of digital-shift-and-add, although the number of rows connecting to one column in the synaptic array is equivalent to 7 bits, 3-bit loss at the ADC side can be tolerated without severely degrading the inference accuracy according to analysis in the study by Jiang et al.\cite{13}. Hence, a 4-bit flash ADC ($7 - 3 = 4$) is applied for the digital case. For the analog case, the number of rows is also of 7 bits. However, 4 columns are merged before connecting to the ADC, so the total number of bits is 11. According to the study by Jiang et al.,\cite{13} because the analog way avoids the additional information loss during the process of digital quantization, 7-bit reduction (more than the 3-bit reduction in the digital case) at the ADC side is tolerable without severely degrading the inference accuracy. Therefore, the ADC precision is also set to 4 bits ($11 - 7 = 4$). The delay of the analog-shift-and-add takes less than 2 ns, from receiving the output voltages of the capacitive array to delivering the final output.

Figure 11b shows the performance improvement achieved by the analog-shift-and-add over the digital-shift-and-add. Because of the parallel configuration, the read dynamic energy is 26.9% lower. The overall area is 16.6% lower because BL MUX is no longer needed and the area of the analog-shift-and-add circuit is relatively compact.

6. Conclusion

In this work, the nonvolatile capacitive crossbar array was experimentally demonstrated for in-memory computing for the first time. Essentially, the capacitor crossbar structure is based on
the ferroelectric HZO, which is sandwiched between top and bottom TiN electrodes. Due to the intrinsic asymmetric interfaces of the plasma-enhanced atomic layer deposition (PEALD)-grown HZO capacitor, it was found that nonidentical capacitance values at DC 0 V can be obtained after program and erase pulses, which could be utilized as two distinct memory states. Depending on the capacitance values, corresponding charges could be stored in the ferroelectric capacitor, which will later be transferred to the reference capacitor shunting the OPAMP and generate the respective output voltage. Based on the charge transfer mechanism, our experimental results of the crossbar showed high linearity with respect to both the inputs and weight values to perform the VMM. Moreover, the reliability test of the capacitive crossbar array showed promising features for the inference engine in terms of cycling endurance of >1000 cycles and extrapolated retention of >10 years at 85 °C.

To evaluate the key parameters, that is, latency, energy, ENOB, etc., the subarray-level SPICE simulation was conducted while considering wire parasitics. Finally, based on the SPICE simulation results, we benchmarked the capacitive crossbar array using DNN + NeuroSim at 22 and 7 nm to evaluate its system-level performance. As capacitors only consume dynamic power but with negligible static power, ≈2× energy efficiency boost could be achieved in the 22 nm capacitive array when compared with the SRAM at 7 nm. Furthermore, the capacitive crossbar array can become selector free due to its capacitor nature without DC sneak paths and also take advantage of the BEOL process, thus achieving high-area efficiency compared with other resistive counterparts and conventional SRAM array. Finally, peripheral circuits with analog-shift-and-add, compatible with the capacitive crossbar array, are proposed. Compared with the design of digital-shift-and-add, the analog way can achieve 16.6% lower area and 26.9% lower energy consumption.

7. Experimental Section

Fabrication Method: Figure 12a shows the key fabrication process of the HZO crossbar capacitor structure. The capacitive crossbar array was fabricated on 100 nm of thermal oxide on top of the p-type (100) silicon wafer. For the capacitor stacks, we used the Fiji G2 ALD system from Veeco. First, 25 nm of PEALD TiN at 250 °C was deposited to form the BE. Then, 21 nm of PEALD Al2O3 was deposited afterward to use it as a hard mask. The photoresist (S1805) was spin coated and baked on the experiment sample to define the BE structure. The Heidelberg MLA150 was used for photolithography and patterning the sample. The Al etchant was used to pattern Al2O3 under the photoresist. Finally, the bottom TiN was patterned using diluted H2O2 at ≈50 °C. The photoresist and Al2O3 were removed.

Figure 12. a) Fabrication process flow of a TiN/HZO(10 nm)/TiN capacitive memory as the capacitive synapse device. b) Illustration of top and lateral views of the HZO crossbar capacitors.

Figure 13. a) The weighted-sum operation requires two steps: 1) charging each entry of CFE and 2) transferring all charges in a column to Cref. Illustration of experimental setup for b) reading weighted sum and c) program/erase (switches around OPAMPs are not shown.) d) Photo of the entire measurement setup.
afterward with the aforementioned chemicals. Then, 10 nm of the ferroelectric layer, PEALD HZO, and 25 nm of the top electrode (TE) PEALD TiN, were deposited. Again, 21 nm of PEALD Al₂O₃ was deposited to form the hard mask for TE patterning. After patterning the top electrode with the identical procedure for BE, the pad area was opened using diluted HF. Finally, the sample was rapidly annealed at 450 °C for 30 s to crystalize the ferroelectric orthorhombic phase of the HZO layer. Figure 12b shows the schematics of a single crossbar capacitor in the fabricated array from lateral and top viewpoints.

**Array-Level Measurement Setup:** Figure 13a-b shows the schematic of our measurement setup for “reading weighted sum,” where the rows received input voltage in parallel through Keithley 707B semiconductor switch matrix. The columns were externally connected to OPAMPS on a printed circuit board. VMM was performed in two phases: 1) charging the array and 2) charge transfer to the output. In our setup, input pulses with 100 mV and 0 V represented binary 1 and 0, respectively, in the input vector using Keysight 81150 A pulse function arbitrary generator. The input vector was multiplied with a column of capacitive weight, resulting in product of charges on each capacitor. After the input was returned to ground, there was no voltage drop across each capacitor; thus, every entry of the product (charges) was then transferred to the reference capacitor (Cref) shunting the OPAMP. The transferred charges resulted in the output voltage of the OPAMP (Vout) as the weighted sum, which was read from the oscilloscope. With more devices in HCS or more input activated, there were more charges and thus a higher Vout was obtained. In Figure 13c, the “program/erase” sequence of the capacitive crossbar array for write was processed with Keithley 4200-SCS parameter analyzer and Keithley 707B semiconductor switch matrix. The write pulse was applied from the 4200-SCS parameter analyzer and multiple cells were selected using the switch matrix. The Cascade-12 K semiauto probe station was used while probing the sample. The photo image of the entire setup for the array-level measurement is shown in Figure 13d.

**Supporting Information**

Supporting Information is available from the Wiley Online Library or from the author.

**Acknowledgements**

J.H. and Y.-C. L. contributed equally to this work. This work was supported in part by ASCENT, one of the six SRC/DARPA JUMP centers. The device fabrication was performed in part at the Georgia Tech Institute for Electronics and Nanotechnology, a member of the National Nanotechnology Coordinated Infrastructure (NNCI), which was supported by the NSF-ECCS-2025462.

**Conflict of Interest**

The authors declare no conflict of interest.

**Data Availability Statement**

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

**Keywords**

compute-in-memory, device reliability, ferroelectric field-effect transistors, ferroelectric random access memory, Hf₀.₅Zr₀.₅O₂ (HZO)

Received: December 11, 2021
Revised: January 16, 2022
Published online: February 23, 2022

[1] S. Yu, Proc. IEEE 2018, 106, 260.
[2] S. Yu, P. Chen, IEEE Solid-State Circuit. Mag. 2016, 8, 43.
[3] P. Jain, U. Arslan, M. Sekhar, B. C. Lin, L. Wei, T. Sahu, J. Alzate-Vinasco, A. Vangapaty, M. Meterelliyo, N. Strutt, A. B. Chen, P. Hentges, P. A. Quintero, C. Connor, O. Golonzka, K. Fischer, F. Hamzaoglu, in IEEE Int. Solid- State Circuits Conf. - (ISSCC), IEEE, Piscataway, NJ 2019, pp. 212–214.
[4] Y.-C. Luo, A. Lu, J. Hur, L. Shaolan, S. Yu, in IEEE 11th Int. Memory Workshop (IMW), IEEE, Piscataway, NJ 2021, pp. 1–4.
[5] Z. Wang, M. Rao, J.-W. Han, J. Zhang, P. Lin, Y. Li, C. Li, W. Song, S. Asapu, R. Midya, Y. Zhuo, H. Jiang, J. H. Yoon, N. K. Upadhyay, S. Joshi, M. Hu, J. P. Strachan, M. Barnell, Q. Wu, H. Wu, G. Qiu, R. S. Williams, Q. Xia, J. J. Yang. Nat. Commun. 2018, 9, 3208.
[6] J. Hur, Y.-C. Luo, N. Tasneem, A. I. Khan, S. Yu, IEEE Trans. on Electron Dev. 2021, 68, 3176.
[7] Q. Zheng, Z. Wang, N. Gong, Z. Yu, C. Chen, Y. Cai, Q. Huang, H. Jiang, Q. Xia, R. Huang, IEEE Electron Dev. Lett. 2019, 40, 1309.
[8] Y.-C. Luo, J. Hur, P. Wang, A. I. Khan, S. Yu, Appl. Phys. Lett. 2020, 117, 073501.
[9] Y.-C. Luo, A. Lu, J. Hur, S. Li, S. Yu, IEEE Trans. Circuits Syst. II: Express Briefs 2021, https://doi.org/10.1109/TCSII.2021.3108148.
[10] X. Peng, S. Huang, Y. Luo, X. Sun, S. Yu, in IEEE Int. Electron Devices Meeting (IEDM), IEEE, Piscataway, NJ 2019, p. 32.5.1–32.5.4.
[11] J. G. Alzate, U. Arslan, P. Bai, J. Brockman, Y. J. Chen, N. Das, K. Fischer, T. Ghani, P. Heil, P. Hentges, R. Jahan, A. Littlejohn, M. Mainuddin, D. Ouellette, J. Pellegrin, T. Pramanik, C. Puls, P. Quintero, T. Rahman, M. Sekhar, B. Sell, M. Seth, A. J. Smith, A. K. Smith, L. Wei, C. Wiegand, O. Golonzka, F. Hamzaoglu, in IEEE Int. Electron Devices Meeting (IEDM), IEEE, Piscataway, NJ 2019, p. 2.4.1–2.4.4.
[12] K. Ni, B. Grisaf, W. Chakraborty, A. K. Saha, S. Dutta, M. Jerry, J. A. Smith, S. Gupta, S. Datta, in IEEE Int. Electron Devices Meeting (IEDM), IEEE, Piscataway, NJ 2018, p. 16.1.1–16.1.4.
[13] H. Jiang, W. Li, S. Huang, S. Cosemans, F. Catthoor, S. Yu, IEEE Design Test 2021, 1.