Experimental Demonstration of Non-volatile Capacitive Crossbar Array for In-memory Computing

Yuan-Chun Luo*, Jae Hur*, Tzu-Han Wang, Anni Lu, Shaolan Li, Asif Islam Khan, Shimeng Yu
*equally contributed; Georgia Institute of Technology, Atlanta, GA 30332, USA
Email: shimeng.yu@ece.gatech.edu

2021 IEEE International Electron Devices Meeting (IEDM) | 978-1-6654-2572-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEDM19574.2021.9720508

***Abstract*—**Resistive crossbar arrays for in-memory computing suffer from high static current/power, serious IR drop, and sneak paths. To overcome these challenges, “capacitive” crossbar array that relies on transient current and charge transfer is becoming attractive since it (1) consumes only dynamic power, (2) has no DC sneak paths and avoids severe IR drop along wires, thus is selector-free, (3) can be fabricated on top of the CMOS circuits for potential 3D stacking. In this work, for the first time, we experimentally demonstrated a ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ (HZO) capacitive crossbar array. The asymmetry of the HZO electrode interfaces leads to the small-signal capacitance on/off ratio >110% that could be read at 0V DC, and a read-disturb-free operation is achieved. The vector-matrix multiplication (VMM) experiments are performed on a fabricated small-scale capacitive crossbar array, showing a linear weighted sum versus either numbers of input or on-state weight. The array-level VMM operation could sustain weight pattern reprogramming after (a) thousands of strong 1ms/3V pulses and (b) an extrapolated 10-year retention at 85 ºC. Array-level SPICE simulation at 22nm node shows that the energy consumption of a capacitive crossbar array is 20~200× lower than that of a resistive crossbar array counterpart.

I. 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 than the traditional von Neumann architecture does. The representative resistive synapses include two-terminal phase change memory (PCM), resistive memory (RRAM), magnetic memory (MRAM), and three-terminal flash transistor or ferroelectric transistor (FeFET) (w/ channel conductance encoding the weight). However, with multiple rows turned on simultaneously and the resistance of synapses being usually of only a few kΩ, the arrays typically conduct high static current and consume large static power. When the array size increases, the wire resistance (R_{wire}) becomes more comparable to the low cell resistance (R_{cell}), resulting in a serious IR drop along wires. Furthermore, the low R_{cell} also induces sneak-path current, causing further degradation in compute accuracy. To counter the sneak paths, an extra access transistor is used for 1T1R cell, but the width of the access transistor is usually sized up to deliver the high write current of the resistive synapse [2], leading to a large cell area (>60F^2 [3]) and a lower area efficiency.

Inspired by the charge transfer principle, capacitive crossbar array has been proposed to overcome the disadvantages of the resistive crossbar array (Fig. 1). Capacitive array utilizes programmable small-signal capacitance states (at DC zero bias) as weight. Since capacitors consume only dynamic power, a high static power is eliminated. Moreover, the open-circuit nature of a capacitor can effectively block the sneak-path current and prevents IR drop along wires. Without the need of an access transistor, the cell size of a capacitive synapse can ideally be kept at the minimum 4F^2. Since HZO capacitors have been proven both CMOS and BEOL compatible [4], the crossbar structure can be potentially fabricated on top of the peripheral circuits, leading to a high area efficiency.

Theoretically, a perfect ferroelectric capacitor should not exhibit capacitance window at zero DC bias (w/ non-switching charges), but only show capacitance window in transient sweep (involving polarization switching). Therefore, interface engineering is required to break the symmetry and to open up the non-zero capacitance window at DC zero bias. Prior work on HfO_2-based capacitive synapse [5] needs a DC bias at ~1.5V, which is not realistic due to the severe read disturb. In Ref. [5], the difference between its high and low capacitance state is also low ~1%. In this work, we will demonstrate a HZO capacitive synapse with on/off ratio >10× larger than that in Ref. [5]. For the first time, we will integrate the HZO capacitors into crossbar array, and set up a measurement system for performing VMM using the charge transfer mechanism. The experimental results show that the output voltage has high linearity versus both the inputs and weight values. Finally, array-level simulation and system-level benchmarking results indicate substantial benefits over the resistive counterparts at advanced node.

II. Device Characteristics of Non-volatile Capacitive Synapse

Fig. 2 shows the fabrication process of the HZO crossbar capacitor structure. The TiN/HZO/TiN capacitor stacks were deposited using plasma-enhanced atomic layer deposition (PEALD). Fig. 3 shows the schematics of a single crossbar capacitor in our fabricated array from lateral and top viewpoints. The asymmetric “small-signal” C-V in Fig. 4 shows higher and lower capacitance states (HCS/LCS) at DC 0V. The asymmetry is attributed to excessive oxygen vacancies at the bottom electrode interface, which induces the domain wall pinning effect (down-polarized even after positive sweep) [6], resulting in more domain walls thus more charges in HCS (Fig. 5). To verify this capacitor as a read-disturb-free and programmable capacitive memory, we further performed small-signal AC measurement at DC=0V directly after program/erase with +3/-3V write pulses (Fig. 6(a)), where the cycling endurance shows steady window even after thousands of strong 3V/1ms programming pulses. For inference engine, weight programming is infrequent thus 1000 cycles satisfy the needs. In Fig. 6(b), a practical read operation applies a 100mV 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 Fig. 6(a). 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 V_w write scheme in Fig. 7(a). The half-selected cells show <10% error in capacitance (Fig. 7(b)). The programming protocols and the definition of error are illustrated in Fig. 7(c).

III. Array-level Measurement Setup

We fabricated a 12×12 capacitive crossbar array (Fig. 8). Fig. 9(a) shows the schematic of our measurement setup, where the rows receive input voltage in parallel through a switch matrix. The columns are externally connected to operational amplifiers (OPAMPs) on a printed circuit board. VMM is performed in two phases: 1) Charging the array; 2) Charge transfer to the output. In our setup, input pulses with 100mV/ 0V represent binary 1/0 in the input vector. The input vector is multiplied with a column of capacitive weight, resulting in product of charges on each capacitor. After the input is returned to ground, there is no voltage across each capacitor so every entry of the product (charges) will then be transferred to the reference capacitor (C_{ref}) shunting the OPAMP. The transferred charges result in the output voltage of the OPAMP (V_{out}) as the weighted sum. With more devices in HCS or more input activated, there will be more charges and thus a higher V_{out}. Measurement setup for "reading weighted sum" and "program/erase" is shown in Fig. 9(b-c). The actual setup for the array-level measurement is shown in Fig. 10.

IV. Array-level Measurement Results

Fig. 11(a) shows the weighted sum, V_{out}, over 12 measurement trials. The array consists of eight 50×50 μm^2 synaptic cells with all input being "1". The tight distribution in Fig. 11(a) implies low cycle-to-cycle variation. Averaged V_{out} in Fig. 11(b) shows a high linearity versus the number of HCS cells. Arrays with smaller synapses are also demonstrated in Fig. 11(c-d). Compared to Fig. 11(b), Fig. 11(c) shows a smaller output swing because the overall charges are reduced with a smaller unit cell area, 25×25 μm^2 with the same C_{ref}. To further reduce the unit cell area down to 10×10 μm^2 and remain an output swing > 100mV, C_{ref} needs to be decreased accordingly, as shown in Fig. 11(d). Highly linear relationship between V_{out} and the number of turned-on WLs is also proven (Fig. 12(a)) with ultra-tight distribution (Fig. 12(b)). Practically, it is challenging to measure sub-10 μm capacitor's response due to parasitics and sensing limit in any off-chip instrument. Therefore, the performance of nanoscale capacitive crossbar arrays is projected by simulations in Section V.

Reliability tests of the array are performed. Fig. 13 shows the cycling endurance with 3V/1ms pulses to reprogram the weight pattern. Even after thousands of such pulses, a sense margin at V_{out} still exists. Fig. 14(a) shows the 15-hour retention at 85 °C, where a clear V_{out} sense margin can be extrapolated to 10 years. Decreasing V_{out} in both HCS and LCS over time implies a decreasing capacitance. This might be a result of the imprint effect. As shown in Fig. 14(b), hours after programmed to HCS, the small-signal C-V curve shifts to the left, resulting in lower HCS capacitance. On the other hand, after erased to LCS, the right-shifting C-V over time (Fig. 14(c)) causes the LCS capacitance to decrease during retention test.

V. Simulations towards large-scale system

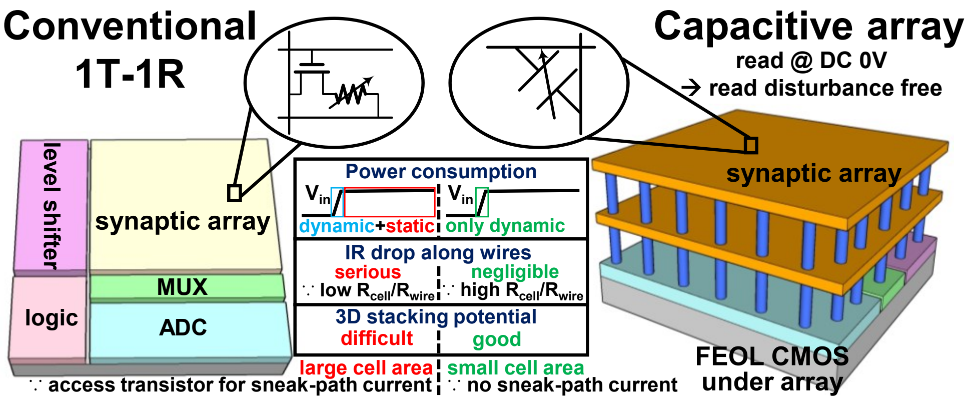
To evaluate the latency, energy, and equivalent number of bits (ENOB, under thermal noise) of the non-volatile capacitive array at an advanced technology node, we run SPICE simulation with 22nm LP transistor models and wire parasitics considered. The array schematic and the OPAMP circuits are shown in Fig. 15(a-b). The array size and other important parameters in the simulation are listed in Fig. 15(c). Supply voltage of the OPAMP is boosted to 1.5V to increase the voltage swing, speed, and ENOB. This setting will not cause severe stress on the transistors in the OPAMP because the cascode structure limits the V_{gs} and V_{ds} below 1V. The SPICE simulation results show an ENOB=4.3 (Fig. 15(d)) and latency for the charge transfer ~5.1ns, where latency is defined as the time for V_{out} to reach 80% of the steady-state voltage. ENOB>4 is achievable, suggesting a 4-bit partial sum quantization, which has been reported to keep a reasonable CIFAR-10 inference accuracy [8]. Subsequently, we compare the capacitive subarray results with those of the resistive subarrays in terms of energy and latency. Due to the suppression of static energy during steady-state readout, the capacitive subarray consumes a much lower total energy, 20~200× lower energy compared to 1-bit/cell 1T-RRAM [3], 1T-1MRAM [9], and 1T-1FeFET [10] (Fig. 16(a)). The subarray energy is normalized to 1-bit multiply-accumulate (MAC).

Based on the array-level SPICE simulation result, we benchmark the capacitive crossbar array with DNN+NeuroSim at 22nm and 7nm to evaluate its system-level performance (Fig. 16(b)). With the capacitive array assumed on top of the peripheral circuits, it is benchmarked with 22nm 1T-RRAM, 1T-1MRAM, 1T-1FeFET, and 22nm/7nm SRAM. The benchmarking results of the 22nm capacitive array show a higher energy efficiency and compute efficiency than those of the other resistive counterparts and 22nm SRAM array. The projection of a 7nm capacitive array also shows a substantial ~2× energy efficiency boost over the 7nm SRAM.

Acknowledgement: This work is supported in part by ASCENT, one of the six SRC/DARPA JUMP centers. The device fabrication is done at IEN cleanroom at Georgia Tech that is supported by the NSF-NNCI program.

References

[1] S. Yu *et al.*, *Proc. IEEE*, vol. 106, no. 2, pp. 260-285, 2018.
[2] S. Yu, P.-Y. Chen, *IEEE Solid State Circuits Magazine*, vol. 8, no. 2, pp. 43-56, 2016.
[3] P. Jain *et al.*, *IEEE ISSCC,* 2019.
[4] J. Hur, *et al.*, *IEEE Trans. Electron Devices*, vol. 68, no. 7, pp. 3176-3180, 2021.
[5] Q. Zheng *et al.*, *IEEE Electron Device Lett.*, vol. 40, no.8, pp. 1309-131, 2019.
[6] Y.-C. Luo *et al.*, *Appl. Phys. Lett.*, 117, 073501, 2020
[7] Y.-C. Luo *et al.*, *IEEE IMW,* 2021.
[8] X. Peng *et al.*, *IEEE IEDM,* 2019.
[9] J. G. Alzate *et al., IEEE IEDM,* 2019.
[10] K Ni, *et al., IEEE IEDM,* 2018.



Power consumption	
V_{in} dynamic+static	V_{in} only dynamic
IR drop along wires	
serious ∵ low R_{cell}/R_{wire}	negligible ∵ high R_{cell}/R_{wire}
3D stacking potential	
difficult	good
large cell area ∵ access transistor for sneak-path current	small cell area ∵ no sneak-path current

Fig. 1. Non-volatile capacitive crossbar array with HZO capacitor for in-memory computing that (1) only consumes dynamic energy due to the capacitive nature, (2) shows low IR drop due to minimum sneak-path current and high ratio between R_{HZO} and R_{wire}, (3) can be fabricated with high density on top of the peripheral circuits, (4) can be designed without access transistors, and (5) allows zero read disturb due to its read operation at DC 0V.

Fabrication process flow for the capacitive synapse

- RCA cleaning of p^+ wafer
- PEALD TiN (12 nm) for bottom electrode (BE)
- BE patterning w/ diluted H_2O_2
- PEALD HZO (10 nm) for ferroelectric layer and PEALD TiN (12 nm) for top electrode (TE)
- TE Patterning w/ diluted H_2O_2
- Patterning of the HZO for pad opening with diluted HF
- RTA at 450 °C for 30 s

Fig. 2. Fabrication process flow for a TiN / HZO(10nm) / TiN capacitive memory as the capacitive synapse.

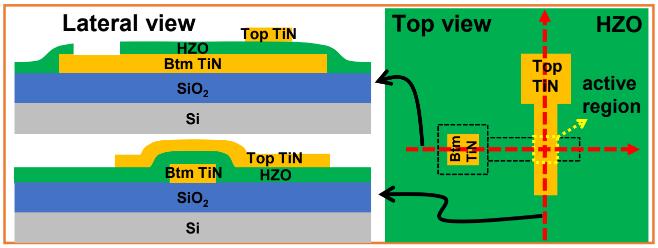


Fig. 3. Illustration of top and lateral views of the HZO crossbar capacitors.

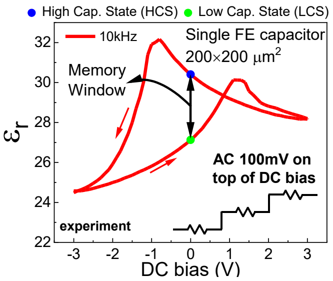


Fig. 4. Asymmetric small-signal C-V shows a memory window at DC 0V with 10kHz 100mV AC small signals applied.

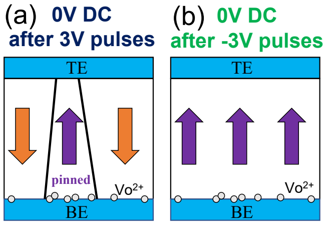


Fig. 5. Physical illustration of the asymmetric C-V at DC 0V. The positively charged oxygen vacancies (Vo^{2+}) at the bottom electrode (BE) pinned some domains up-polarized. (a) While +3V pulses tend to flip the domains down-polarized, more domain walls (DWs) are formed due to the pinned domains, resulting in higher small-signal capacitance. (b) The pinned up-polarized domains do not create DWs as -3V pulses tend to flip the domains up-polarized too, resulting in a lower capacitance.

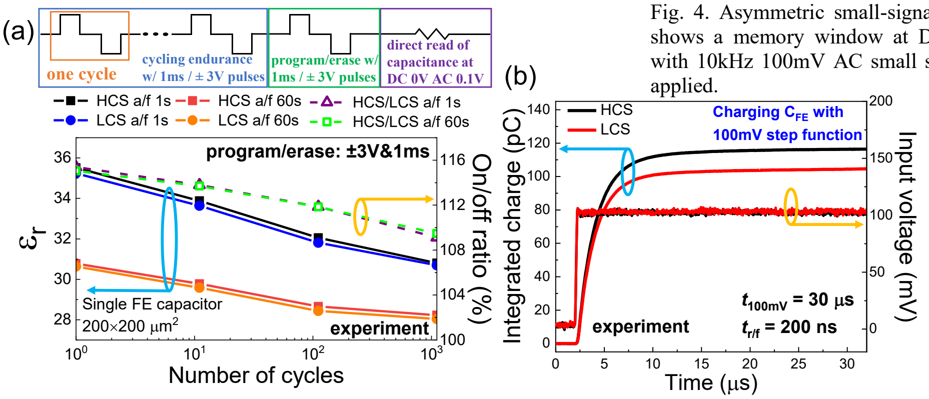


Fig. 6. (a) Capacitive memory window measured directly at DC 0V without sweeping. Distinct memory window exists even after thousands of strong 3V/1ms pulses. (b) Capacitive memory window measured in displacement charges integrated from current flowing onto the capacitor under small 100mV input pulse.

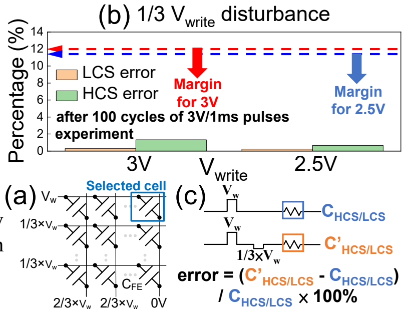


Fig. 7. (a) 1/3 V_w scheme (b) Small write disturbance compared to the total margin under 1/3 V_w scheme and (c) programming protocols and the definition of errors of LCS and HCS in (b).

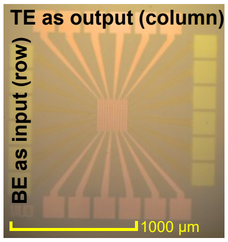
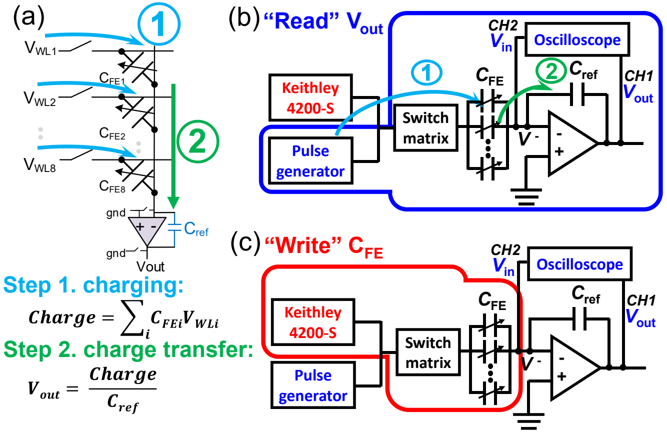


Fig. 8. Microscopic photo of the fabricated capacitive array with unit cell area = 10×10 μm². The bottom electrode (BE) is where the WL input voltages are applied while the top electrode (TE) serves as columns where output charges flow.



Step 1. charging:

$$Charge = \sum_i C_{FEi} V_{WLi}$$

Step 2. charge transfer:

$$V_{out} = \frac{Charge}{C_{ref}}$$

Fig. 9. (a) The weighted-sum operation requires two steps: (1) charging each entry of C_{FE} and (2) transferring all charges in a column to C_{ref}. (b-c) Illustration of experimental setup for (b) reading weighted sum and (c) program/erase (switches around OPAMPs are not shown.)

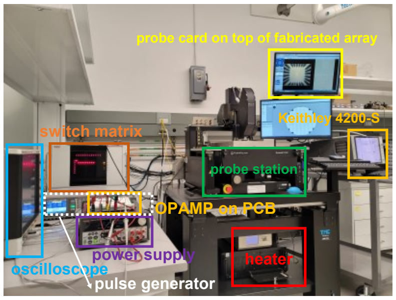


Fig. 10. Photo of the entire measurement setup.

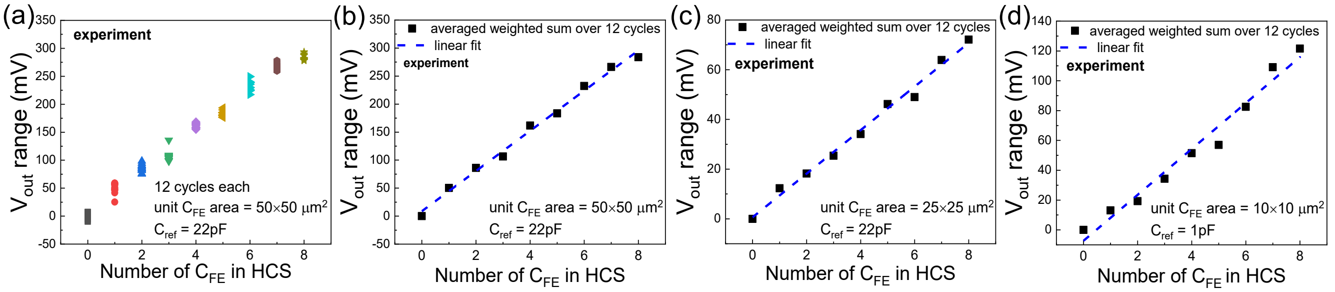


Fig. 11. Array-level VMM experiments: measured output voltage range represents weighted sum versus the number of high capacitance states (HCS) among a total of eight capacitive synapses in a column of an array with different unit cell sizes (a-b) 50×50 μm² (c) 25×25 μm² (d) 10×10 μm², with all the input turned on. (a) scatter plot of 12 measurements for each number of C_{FE} in HCS. (b-d) Each point is an averaged value over 12 consecutive readout.

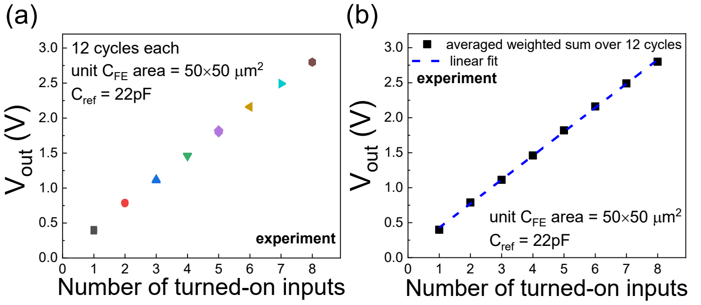


Fig. 12. Measured output voltage versus number of turned-on inputs with all devices programmed to HCS. (a) scatter plot of 12 measurements for each number of turned-on inputs. (b) averaged value over 12 consecutive readout.

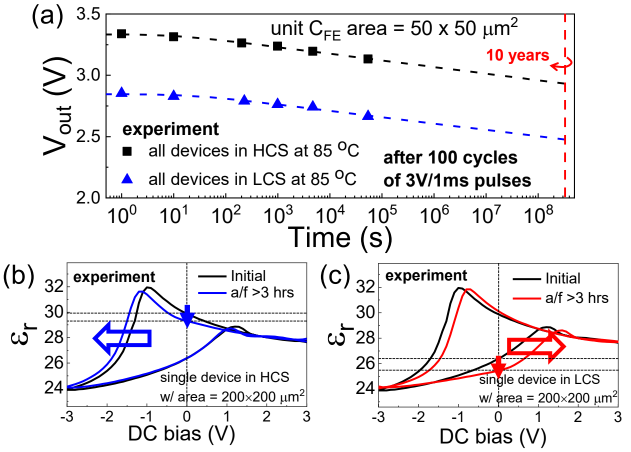


Fig. 14. (a) Array-level V_{out} retention at 85 ºC with all devices programmed to HCS or LCS. (b) Imprint effect after potentiation pulses leads to C-V curve shifting left, resulting in HCS decreasing over time. (c) Imprint effect after depression pulses leads to C-V curve shifting right, resulting in LCS decreasing over time.

(a) Subarray-level Benchmark

	Capacitive (this work)	1T-1RRAM [3]	1T-1MRAM [9]	1T-1FeFET [10]
C_{on}/R_{on} (aF/Ω)	120	6k	2.5k	67k
On/off ratio	1.125	17	2.8	100
Bit/cell	1	1	1	1
Energy (pJ)	0.96	88	193	21
Delay (ns)	5.1	<5	<5	<5
BEOL stacking	Yes	No	No	No

Subarray energy includes the static & dynamic energy of the crossbar structure.
Subarray energy is normalized to 1-bit vector-matrix multiplication.

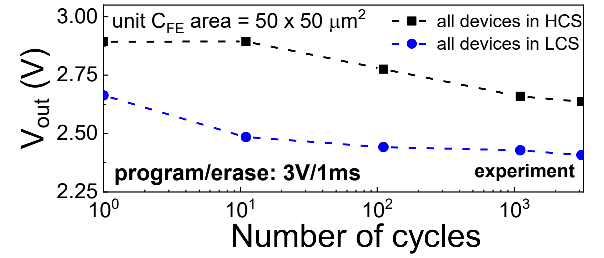
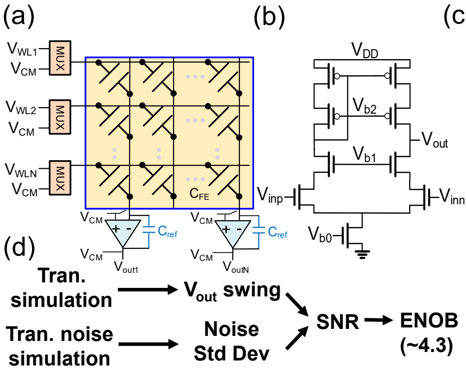


Fig. 13. Array-level pulse cycling endurance with all devices programmed to HCS or LCS. 3V/1ms pulses are applied before reading the weighted sum, V_{out}, at end of the column. Even after thousands of such strong write pulses, a distinct memory window still survives, suggesting its feasibility as an inference engine.



(c)

parameter	value
Technology node	22nm
C_{FE} in HCS	120aF
On/off ratio	1.125
Array size	128x128
C_{ref}	3pF
Boosted suppy voltage of OPAMP	1.5V
V_{WL} amplitude	60mV

Fig. 15. (a) Schematic of the capacitive crossbar array and its peripheral circuits (b) 9T OPAMP is used for small area and low power consumption as well as reasonable latency. (c) Key parameters for the SPICE simulation. (d) ENOB extraction from SPICE simulation.

(b) System-level Benchmark

VGG-8 (8-bit activation; 8-bit weight) on CIFAR10, with Novel Weight Mapping and Dataflow							
Technology node (LSTP)	7nm		22nm				
Device	SRAM	capacitive (this work)	SRAM	capacitive (this work)	1T-1RRAM [3]	1T-1MRAM [9]	1T-1FeFET [10]
Cell area (F^2)	600	92 (BEOL)	200	9 (BEOL)	60	100	26
ADC	4b-ML-VSA						
Chip area (mm^2)	8.4	4.1	55.2	25.6	60.0	88.6	38.0
Latency (ms)	0.95	0.87	1.04	0.98	1.32	1.88	0.90
Throughput (TOPS)	1.30	1.41	1.19	1.25	0.94	0.66	1.37
Energy efficiency (TOPS/W)	59	105	24	44	27	20	38
Compute efficiency ($GOPS/mm^2$)	154	341	22	41	16	7	36
Power (mW)	21	13	47	28	34	33	36

Subarray size = 128*128;
F = 7nm or 22nm for normalizing cell area, doesn't indicate physical feature size.
The 7nm projection of "capacitive" applies 7nm peripheries while keeping the same cell area as the 22nm one.

Fig. 16. (a) Subarray-level evaluation with SPICE simulation, compared to those of the representative crossbar arrays obtained using DNN+NeuroSim framework. The energy and latency of the resistive arrays includes 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 pre-trained VGG-8 model. Delay of the capacitive array is defined as the output voltage reaches 80% of the steady-state value. (b) System-level benchmarking results show that the capacitive array has the potential of outperforming its resistive counterparts and can be more competitive over 7nm SRAM.