Variation Tolerant and Energy-Efficient Charge Domain Compute-in-Memory Array with Binary and Multi-Level Cell Ferroelectric FET

Jiahui Duan[1*], Yixin Xu[2*], Zijian Zhao[1], Anni Lu[3], James Read[3], Mohsen Imani[4], Thomas Kampfe[5], Mike Niemier[1], Xiao Gong[6], Shimeng Yu[3], Vijaykrishnan Narayanan[2], and Kai Ni[1]
[1]University of Notre Dame; [2]Pennsylvania State University; [3]Georgia Institute of Technology; [4]University of California, Irvine; [5]Fraunhofer IPMS; [6]National University of Singapore;
*Equal contribution; (email: jduan3@nd.edu)

***Abstract*— In this work, we present a variation-tolerant and energy-efficient charge-domain Ferroelectric FET (FeFET) based Compute-in-Memory (CiM) array design that is compatible with both binary and multi-level cell memory sensing. We demonstrate that: 1) by exploiting FeFET as a nonvolatile switch, its high ON/OFF ratio in the subthreshold region can suppress the error introduced by the inaccurate ON state conductance, thus realizing robust CiM operations, unlike the current-domain CiM design where the computation results is highly sensitive to the device conductance variation; 2) by leveraging a dense dynamic random access memory (DRAM)-like 1FeFET1C cell structure, the proposed design benefits from the existing high density DRAM establishment while also significantly relaxing the capacitor retention and transistor leakage requirement; 3) the charge-domain CiM supports both binary FeFET with minimum overhead and MLC FeFET with tolerable latency for MLC state sensing, whose efficacy is validated experimentally on both cell-level and array-level; 4) the proposed CiM shows much better device variation resilience than conventional current-domain CiM, and also improves inference accuracy. Macro-level evaluation results demonstrate significantly higher energy efficiency and area efficiency compared to prior CiM works.**

I. Introduction

With the rapid advances in artificial intelligence (AI) models, CiM has attracted attention and been treated as a promising solution for AI applications. In this regard, both binary and MLC non-volatile memory (NVM) based CiM are highly attractive. Existing NVM based CiM methodologies can be roughly classified into two main categories: (i) current-domain CiM (Fig.1(a)); and (ii) charge-domain CiM (Fig.1(b)). More specifically, the current-domain CiM design takes NVM devices (e.g., FeFETs) as conductance and summing up their currents as computed output, thus requiring accurate conductance of FeFETs to distinguish different computation results. It is challenging for binary states and even worse when applying MLC computing (Fig.1(c)). However, FeFETs in the charge-domain CiM act as switches, thus exact ON current doesn't matter as long as capacitors are charged or discharged in time, as shown in Fig.1(d). In this work, we propose a 1FeFET-1C cell to support charge-domain CiM with binary and MLC FeFETs. Such a structure, akin to DRAM cell, can also exploit the decades-long DRAM establishment for high density CiM array (Fig.1(e)). Compared to other charge-domain CiM with different technologies, this 1FeFET1C-based work shows much relaxed requirements on cell capacitors and transistors, lower power consumption, excellent MLC compatibility, and good scalability (Fig.1(f)).

II. 1FeFET1C Cell Integration and Binary CiM Operation

The process integration flow of the 1FeFET1C cell (Fig.2(a)) is shown in Fig.2(b). A $Hf_{0.5}Zr_{0.5}O_2$ (10nm)/Al_2O_3(1nm)/$Hf_{0.5}Zr_{0.5}O_2$(10nm) gate stack is deposited by atomic layer deposition (ALD). Then the via is opened and followed by tungsten (W) sputtering as the bottom electrode of capacitor. Subsequently, a 10nm HfO_2 layer is deposited followed by the top electrode sputtering. The cell top view scanning electron microscopy (SEM) image (Fig.2(c)) shows the FeFET and the capacitor. The gate stack of the FeFET, represented by the cross-sectional transmission electron microscopy (TEM) image (Fig.2(d)) and the atomic composition (Fig.2(e)), clearly show the two layers of $Hf_{0.5}Zr_{0.5}O_2$ separated by the middle Al_2O_3 layer, which is to prevent the ferroelectric film from stabilizing in the monoclinic phase. For the FeFET, a maximum memory window of 2.5V is obtained (Fig.2(f)), which can hold 4 states for MLC FeFET (Fig.2(g)). The capacitance and leakage current of capacitor are shown in Fig.2(h).

Fig.3 shows the basic principles of the proposed binary 1FeFET1C charge-domain CiM operation in the array level. For each cell, one bit of weight is stored in the FeFET as the low-V_{TH} (LVT) state or the high-V_{TH} (HVT) state. In the first cycle, bit lines (BLs) are applied with V_x and input are applied as different word line (WLs, i.e., inputs) voltages to charge each cell capacitor to store the local AND results, i.e., x_iS_{ij}, where S_{ij} is the binary state ('0' or '1') of the FeFET. Then, in the second cycle, a large-enough V_{pass} is applied to WLs, and BLs are floated to enable charge sharing to obtain the final MAC computation results. With the developed cells, such operations are validated experimentally. The charging of the cell capacitor with FeFET in the LVT (Fig.4(a)) and HVT (Fig.4(c)) is studied. As shown in Fig.4(b) and Fig.4(d), successful passing and blocking of the V_{BL} is demonstrated when the FeFET is at the LVT and HVT, respectively. The array operation is also demonstrated. Fig.4(e) shows the V_{BL} transients during the charge sharing step, where the V_{BL} increases linearly with the number of LVT FeFETs (Fig.4(f)), thus validating the proposed operation.

2024 IEEE International Electron Devices Meeting (IEDM) | 979-8-3503-6542-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/IEDM50854.2024.10873305

III. MLC 1FeFET1C Charge-Domain CiM

For conventional MLC NVM-based CiMs, there are two sensing methods: (i) parallel sensing: use a constant and high enough V_G to read different states by current values (Fig.5(a)-(b)); (ii) sequential sensing: use multiple read V_G and distinguish states by the sensed current at each step (Fig.5(c)-(d)). As shown in Fig.5(b) and Fig.5(d), sequential sensing mode shows a large sense margin with tolerable latency as each step can harness the transistor ON/OFF ratio. In addition, it shows a much better tolerance against device variation than the parallel mode, in which the device variation is directly translated into the conductance variation. But for sequential sensing, the memory acts as a switch where the exact conductance value does not significantly impact the operation.

However, the use of the sequential sensing method for conventional current-domain NVM CiM is hindered by the inequality of cell read currents (Fig.5(e)), thus causing inaccurate MLC weight representation. The proposed charge-domain 1FeFET1C-based CiM, applicable with MLC states, can address this challenge. By introducing 1FeFET1C cell structure with MLC FeFETs, the FeFET in each cell acts as a nonvolatile switch to control whether the cell capacitor need to be charged (Fig.6). And the intermediate computation results would be hold in the capacitors, not in the FeFETs. Hence, no precise read currents are needed. In the proposed charge-domain CiM design (Fig.7(a)), weights (2bits) are stored as multi-level states of FeFETs in each 1FeFET1C cell, and the inputs are sent to WLs.

To realize MLC MAC operations, 4 cycles are required. During the first 3 cycles, if the input is bit '0', WLs would be set to V_0 until the 4th cycle, where no charging will happen, irrespective of the FeFET states. If the input is bit '1', different read voltages (V_{read3}/V_{read2}/V_{read1}) are given to WLs sequentially to turn ON/OFF FeFETs. Meanwhile, BLs are asserted to 3 different voltage levels to charge capacitors (Fig.7(b)). In this way, after 3 cycles, cell capacitors would be charged to different voltages which represent the results of dot-product operations between input and MLC weight. In the 4th cycle, the analog summation is performed by floating BLs and then allow charge sharing among all the capacitors to take place such that the stabilized V_{BL} represents the computation results (Fig.7(c)). Compared to current-domain CiM designs, this design is free from static power consumption due to the charge-based computation. Besides, it has better device variation resilience which helps to ensure MAC computation accuracy.

Such an MLC based charge-domain CiM is validated with fabricated 1FeFET1C arrays. Fig.8(a) shows the top view SEM of the array. For demonstration an access transistor is included for charging and sharing processes. By biasing the gate of this access transistor (SL), connection between the BL and floating state of sense node (SN) can be controlled which are necessary for charge and sharing process. The entire operation includes three steps. First, all four FeFETs are initialized to target MLC states. Second, three cycles' charging processes are performed by biasing WL at V_{read3}, V_{read2}, V_{read1}, and BL at V_x/3, 2V_x/3, V_x sequentially so that capacitors in different cells are charged to different voltage based on the MLC state of connected FeFET. Before the charging, all capacitors are discharged to ground. In our measurement, V_x is set at 0.3 V, and V_{read3}, V_{read2}, V_{read1} are chosen 2.1 V, 1.6 V, and 0.9 V, respectively, based on the V_{TH} of FeFET at different MLC states. The Fig.8(c) shows the capacitor voltage transient in a single cell during charging process for different MLC states. The four states can be clearly recognized. And after turning OFF the access transistor, the SN is floated, thus ready for charge sharing of 4 cells enabled by turning on all FeFETs. The computing results can be observed by sensing the SN voltage. Fig.8(d) shows the voltage on the SN, which also shows a good linearity with respect to the theoretical MAC output.

IV. Variation and System Benchmarking

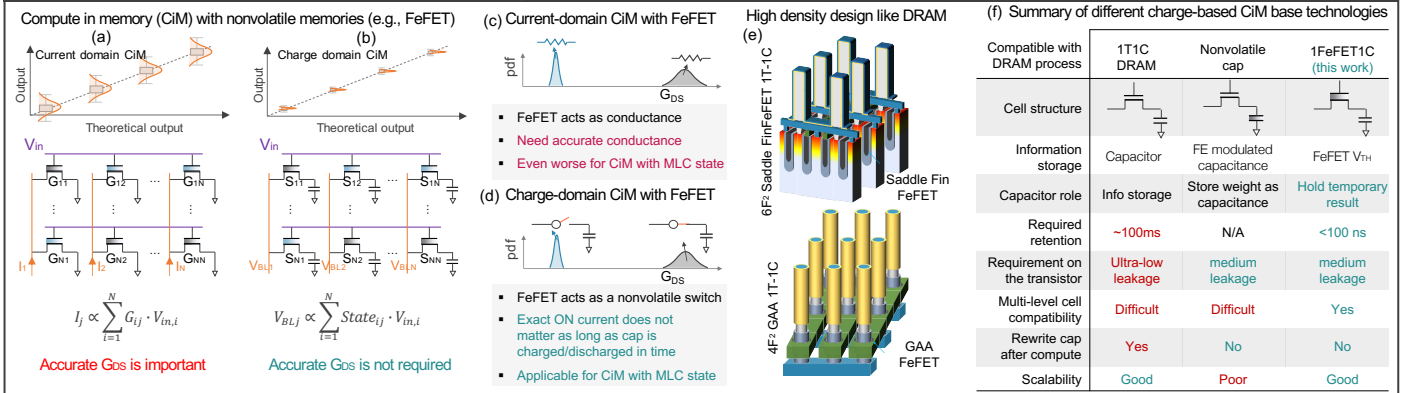
Next SPICE simulations with calibrated FeFET models are conducted to evaluate the potential of scaling to larger scale systems. Fig.9(a) shows that a high degree of linearity of V_{BL} on MAC output is observed with different numbers of cells ranging from 32 to 128 cells in a single column. However, no device variation is considered in this case. To understand the impact of device variation, Monte Carlo simulations with 4σ deviation are conducted. Thanks to its large sensing margin, this design has much better tolerance against V_{TH} variation (Fig.9(b)), as compared with the conventional current domain CiM (Fig.9(c)). As a result, even with a large V_{TH} variation, the array constructed with 1FeFET1C array can successfully maintain a tolerable accuracy loss (Fig.9(d)), compared with current-domain CiM, making it highly promising for emerging NVM technologies. To get a holistic picture of all the device variations, Fig.9(e) studies both the V_{TH} variation and the cell capacitor variation, which shows that the additional capacitor variation does not have a significant effect on the MAC output, again highlighting the robustness of the proposed design. Therefore, the proposed design offers many advantages over the conventional current-domain CiM (Fig.9(f)). Fig.9(g) illustrates the design of a 128x128 1FeFET-1C CiM subarray including peripheral circuitry. The evaluation is conducted with DNN+NeuroSim framework [1] and tested with VGG8 (8-bit input activations and 8-bit weights) on CIFAR10. The evaluation shows this work achieves an energy efficiency of 3200 TOPS/W and an area efficiency of 231.67 TOPS/mm^2 with 1-bit input and 1-bit weight operations. Fig.9(h) shows a comparison with the state-of-the-art charge domain CiMs, showing the excellent performance of our proposed design.

V. Conclusion

In this work, a robust and energy-efficient binary and MLC FeFET-based CiM design is presented by leveraging charge-domain computing. The functionality of binary MAC operations and MLC MAC operations are validated by cell-level and array-level experiments. Besides, the device variation study demonstrates that this design has much better resilience against device variation resilience than conventional current-domain CiM. The macro-level benchmarking also demonstrates that our design shows higher area efficiency and higher energy efficiency over prior CiM works.

Acknowledgement: This work was primarily supported by SUPREME and PRSIM, two of the SRC/DARPA JUMP 2.0 centers, NSF 2344819, 2235366, 2235472, and Singapore MOE Tier 1 A-8001168-00-00 and A-8002027-00-00. Device characterization was supported by U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Energy Frontier Research Centers program under Award Number DESC0021118. **References:** [1] S. Yu et al., IEDM 2019; [2] M. Yu et al., S. VLSI 2022; [3] J. Kulkarni et al., ISSCC 2021; [4] S. Yu et al., JSSC 2022; [5] H. Amrouch et al., Nat Commun 2022

Motivation: Variation-Tolerant Charge Domain Compute-in-Memory with Binary and MLC Ferroelectric FET



(f) Summary of different charge-based CiM base technologies

Compatible with DRAM process	1T1C DRAM	Nonvolatile cap	1FeFET1C (this work)
Cell structure			
Information storage	Capacitor	FE modulated capacitance	FeFET V_{TH}
Capacitor role	Info storage	Store weight as capacitance	Hold temporary result
Required retention	~100ms	N/A	<100 ns
Requirement on the transistor	Ultra-low leakage	medium leakage	medium leakage
Multi-level cell compatibility	Difficult	Difficult	Yes
Rewrite cap after compute	Yes	No	No
Scalability	Good	Poor	Good

Fig.1. Two main categories CiM methodologies exist for nonvolatile memories: (a) current-domain CiM; and (b) charge-domain CiM. (c) Compared to current-domain CiM that requires precise conductance mapping of the weights, (d) proposed charge-domain CiM takes FeFETs as switches, which makes it tolerant against device variation. (e) The proposed 1FeFET1C cell structure can leverage decades of technology know-how of DRAM. (f) The proposed 1FeFET1C structure shows the advantages of MLC compatibility, good scalability and relaxed requirement on the capacitor retention.

1FeFET1C Cell Integration and Characterization

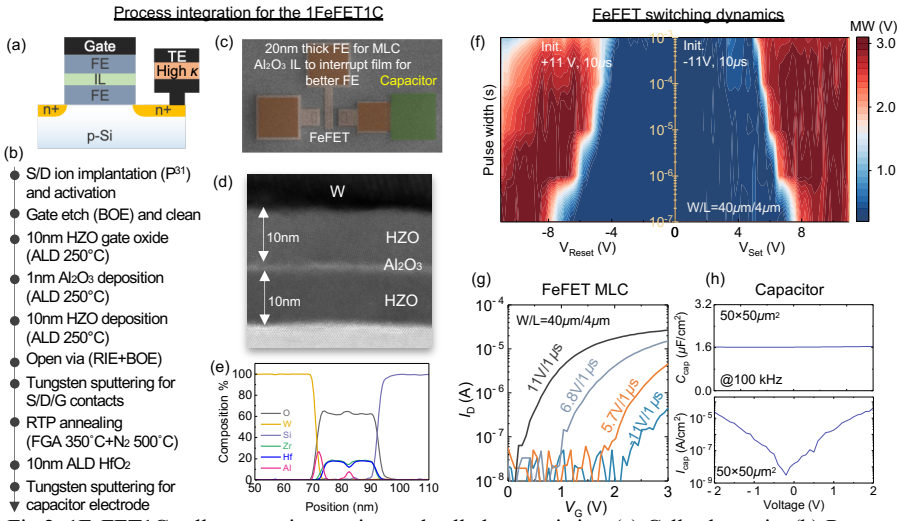


Fig.2. 1FeFET1C cell process integration and cell characteristics. (a) Cell schematic. (b) Process integration flow. (c) Top view SEM. (d) Cross-sectional TEM image and (e) corresponding atomic composition of FeFET gate stack. (f) Switching Dynamics of the FeFET. (g) Measured I_D-V_G curves of four memory states of FeFET. (h) Capacitance and current for the capacitor.

Principles of Binary 1FeFET1C CiM Operation

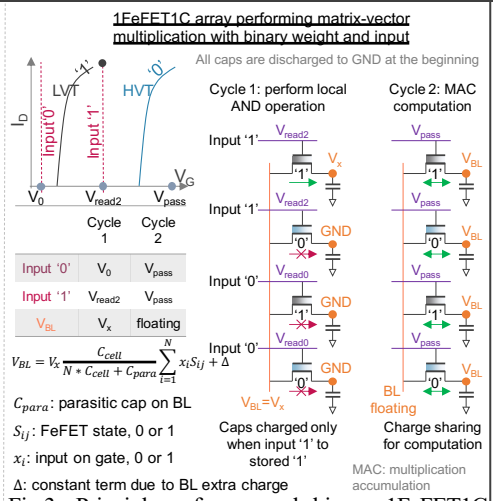


Fig.3. Principles of proposed binary 1FeFET1C CiM. Binary weights are stored in FeFETs. Different voltages are sent to BLs and WLs (inputs) to charge cell capacitors to store $x_i S_{ij}$. Then BL is floated for charge sharing to compute MAC results.

Experimental Validation of Binary 1FeFET1C CiM

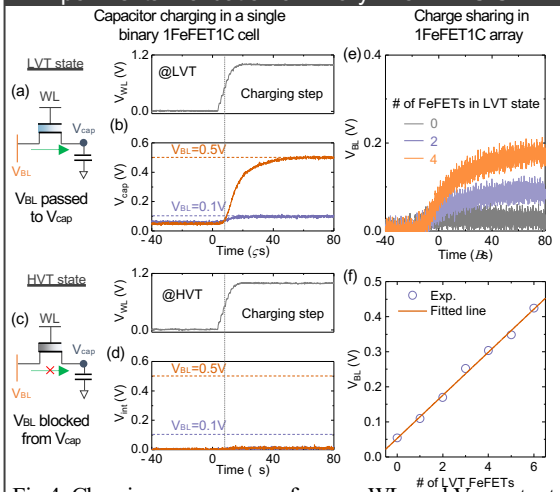


Fig.4. Charging process, waveform on WL, and V_{cap} output in a single binary 1FeFET1C Cell for (a)-(b) LVT state and (c)-(d) HVT state. (e) Output BL voltage after charge sharing in 1FeFET1C array. (f) A linear trend of output BL voltage against number of LVT FeFETs in the array.

Challenges of MLC FeFET for Current-Domain CiM

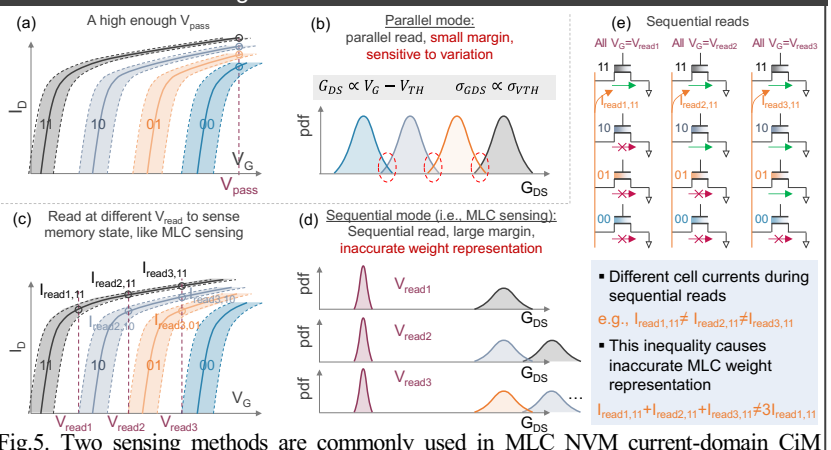


Fig.5. Two sensing methods are commonly used in MLC NVM current-domain CiM designs: (a)-(b) parallel sensing – use a single V_G to read states and distinguish them via current values. This method has a relatively small sense margin and is sensitive to device variation. (c)-(d) sequential sensing – use multiple read voltages and distinguish different states by sensed current of each cycle, which has a large sense margin. (e) However, sequential sensing mode still has unsolved challenges in current-domain CiM -- inaccurate weight representation caused by inequal read cell currents at different read cycles.

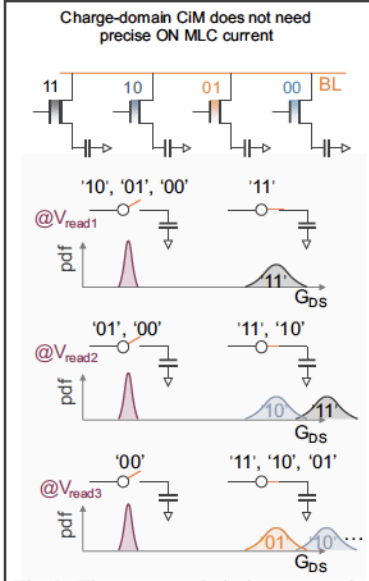
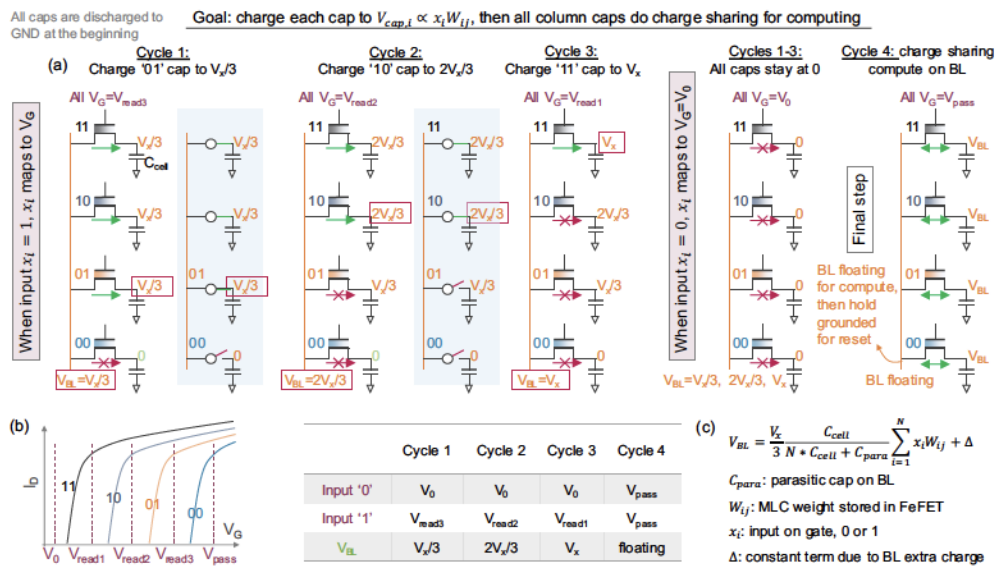


Fig.6. The proposed design can solve the aforementioned challenge by introducing 1FeFET1C cell structure and charge-domain computing since FeFETs act as a switch, not conductance in this design, thus precise read currents are not necessary.



	Cycle 1	Cycle 2	Cycle 3	Cycle 4
Input '0'	V_0	V_0	V_0	V_{pass}
Input '1'	V_{read3}	V_{read2}	V_{read1}	V_{pass}
V_{BL}	$V_x/3$	$2V_x/3$	V_x	floating

Fig.7. (a) In the proposed charge-domain CiM design, 4 cycles are required to realize MLC MAC operations, the first 3 of which are the charging process and the last is the sensing process. During the charging process, if x_i is '1', V_{read3}/V_{read2}/V_{read1} are applied to all WLs sequentially, while WLs are always set to V_0 when x_i is '0'. During the sensing process, a pass voltage (V_{pass}) is applied on all WLs for charge sharing. (b) Different voltages are sent to BLs and WLs (inputs) to charge cell capacitors to different voltages which represent x_iW_{ij} results. (c) Starting from the 4th cycle, BLs are floated and charges in capacitors are shared among all cells so that the final V_{BL} represents the MAC output.

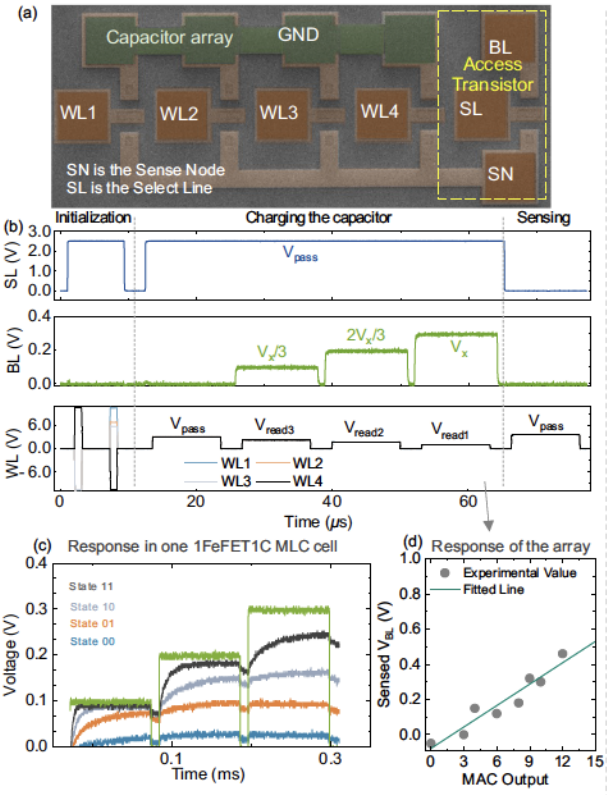
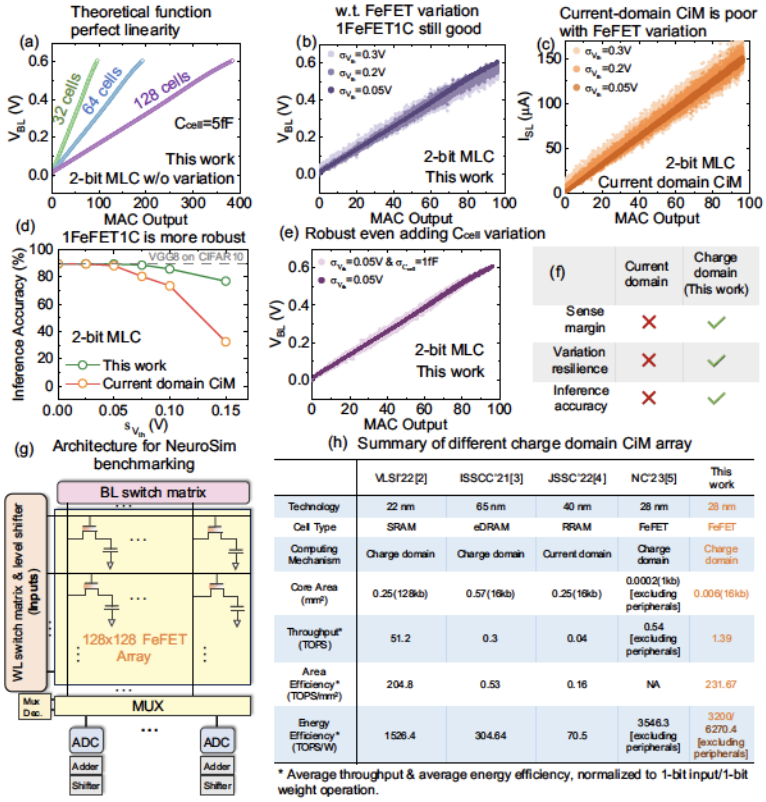


Fig.8. (a) Top view SEM of the developed array. (b) The waveform of charge-domain MLC CiM operation in array, including the initialization, charging, and sensing process. (c) The response of 1FeFET-1C cell during the charging process shows a linear trend. After isolating BL and floating SN by turning off the access transistor, (d) the MAC output can be sensed by the voltage on SN, showing good linearity.



(f)	Current domain	Charge domain (This work)
Sense margin	✗	✓
Variation resilience	✗	✓
Inference accuracy	✗	✓

(h) Summary of different charge domain CiM array

	VLSI'22 [2]	ISSCC'21 [3]	JSSC'22 [4]	NC'23 [5]	This work
Technology	22 nm	65 nm	40 nm	28 nm	28 nm
Cell Type	SRAM	eDRAM	RRAM	FeFET	FeFET
Computing Mechanism	Charge domain	Charge domain	Current domain	Charge domain	Charge domain
Core Area (mm²)	0.25(128kb)	0.57(16kb)	0.25(16kb)	0.0002(1kb) [excluding peripherals]	0.006(16kb)
Throughput* (TOPS)	51.2	0.3	0.04	0.54 [excluding peripherals]	1.39
Area Efficiency* (TOPS/mm²)	204.8	0.53	0.16	NA	231.67
Energy Efficiency* (TOPS/W)	1526.4	304.64	70.5	3546.3 [excluding peripherals]	3200/ 6270.4 [excluding peripherals]

* Average throughput & average energy efficiency, normalized to 1-bit input/1-bit weight operation.

Fig.9. (a) The final V_{BL} of 32/64/128 cells in a single column shows high degree of linearity against the MAC output. (b)-(d) This design has better tolerance against V_{TH} variation on MAC output and inference accuracy than the current-domain CiM. (e) The impact of V_{TH}/ V_{TH}&C_{cell} variation on MAC output of this work is limited. (f) Comparison between charge-domain and current-domain CiM with FeFET. (g) Proposed CiM subarray design. (h) Compared with prior charge-domain CiM designs, this work excels at energy and area efficiency.