SCIENCE CHINA
Information Sciences

• LETTER •

October 2025, Vol. 68, Iss. 10, 209405:1–209405:2
https://doi.org/10.1007/s11432-025-4454-2

An energy-efficient FeFET-based computing-in-memory macro using BEOL-integrated HZO ferroelectric capacitors

Weizeng LI[1,2], Zhidao ZHOU[1,2], Linfang WANG[1,2], Junyu ZHU[1,2], Junzhe SHEN[1,2], Hongyang HU[1], Baihan WANG[1,2], Zhi LI[1,2], Wang YE[1,2], Zhongze HAN[1,2], Hanghang GAO[1,2] & Chunmeng DOU[1*]

[1]*Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China*
[2]*University of Chinese Academy of Sciences, Beijing 100049, China*

Received 6 January 2025/Revised 23 April 2025/Accepted 19 May 2025/Published online 17 September 2025

Citation Li W Z, Zhou Z D, Wang L F, et al. An energy-efficient FeFET-based computing-in-memory macro using BEOL-integrated HZO ferroelectric capacitors. Sci China Inf Sci, 2025, 68(10): 209405, https://doi.org/10.1007/s11432-025-4454-2

Traditional von Neumann artificial intelligence (AI) edge processors face challenges in energy efficiency and latency, primarily due to frequent data transfers between computation and memory units. Non-volatile computing-in-memory (nvCIM) addresses these challenges by performing computations inside the memory array. However, in order to boost its energy efficiency and throughput, there are several key requirements: (1) large memory window to ensure sufficient signal margin [1]; (2) high endurance to enable frequent weight updating [2]; (3) low power readout scheme to minimize the energy- and area-costs of analog-to-digital converters (ADC) [2, 3].

In recent years, HZO-based ($Hf_{0.5}Zr_{0.5}O_2$) FeFETs have been advancing rapidly, offering key advantages such as large memory window, high endurance, and low programming power [4–6]. Depending on the integration scheme of the ferroelectric layers, FeFETs can be classified into front-end-of-line (FEOL) and back-end-of-line (BEOL) types. Compared to the FEOL FeFETs, BEOL FeFETs feature a good process compatibility with the standard CMOS process. Besides, by eliminating the silicon dioxide interfacial layer in the FEOL ones, BEOL FeFETs exhibit improved endurance [5]. Therefore, BEOL FeFETs can potentially offer a competent platform to implement nvCIM, which remains largely unexplored.

This study presents an 8 kb HZO-based BEOL FeFET nvCIM macro for classification using low power- and area-cost ReLU-winner-take-all (WTA) circuits for non-linear processing and analog readout. A test-chip is also demonstrated in silicon using the standard CMOS process.

Proposed FeFET CIM macro. Figure 1(a) illustrates the overall structure of the HZO-based BEOL FeFET nvCIM macro. The macro consists of a 256×32 FeFET array (FA), selectors and drivers for the word line (WL), bit line (BL), and source line (SL), a WTA with the ReLU activation function (ReLU-WTA), as well as the timing and mode control unit (Ctrl). Ternary weight data are encoded using a pair of HZO BEOL FeFET cells. For a weight of 0, both cells in the even positive and odd negative BLs are set to high threshold voltage (HVT). For a weight of +1, the cell in the even-numbered positive BLs is set to low threshold voltage (LVT), while the cell in the odd-numbered negative BLs is set to HVT. Conversely, for a weight of −1, the cell in the even-numbered positive BLs is set to HVT, while the cell in the odd-numbered negative BLs is set to LVT. Clamping transistors are used to stabilize the BLs to the clamping voltage (V_{CLP}), suppressing voltage fluctuations that could affect computing accuracy. The input voltages ($V_{\mathrm{IN}}[0:255]$) are applied to the FA, and the output currents ($I_{\mathrm{OUT}}[0:31]$) from the even-numbered and odd-numbered BLs are routed to the ReLU-WTA. After generating the activation currents based on the ReLU function, the ReLU-WTA compares the currents and recognizes the largest one. The output corresponding to the largest current is pulled up, while the outputs for all other currents are pulled down.

Figure 1(b) illustrates the schematic of the ReLU-WTA circuit, which comprises 16 branches. Each branch includes a current mirror (CM[15 : 0]) to generate ReLU activation currents ($I_{\mathrm{ACT}}[15:0]$) in the analog domain and a WTA cell (WTA[15 : 0]) to compare these currents, select the largest one, and output the classification result ($\mathrm{WTA_{OUT}}[15:0]$). The current mirror subtracts the current on the odd-numbered negative BL ($I_{\mathrm{OUT}}[2N+1]$) from the even-numbered positive BL ($I_{\mathrm{OUT}}[2N]$). For the activation current generated in a given branch ($I_{\mathrm{ACT}}[N]$), it equals the following equation:

$$I_{\mathrm{ACT}}[N] = \max(0, I_{\mathrm{OUT}}[2N] - I_{\mathrm{OUT}}[2N+1]). \quad (1)$$

* Corresponding author (email: douchunmeng@ime.ac.cn)

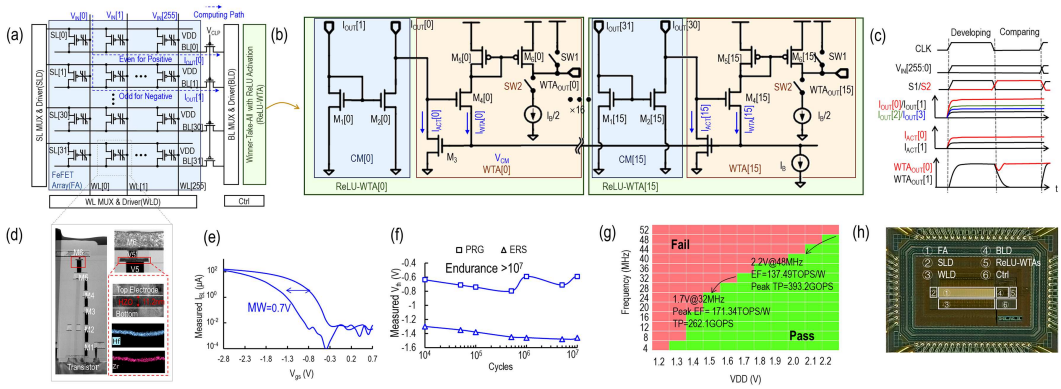

Figure 1 (Color online) (a) The overall structure of the HZO-based BEOL FeFET nvCIM macro; (b) the schematic of the ReLU-WTA; (c) the waveform of the macro; (d) the fabrication process of the FA; (e) the I-V characteristics of the HZO BEOL FeFET; (f) the endurance evaluation of the HZO BEOL FeFET; (g) shmoo plots of the chip; (h) die photo of the chip.

Figure 1(c) illustrates the waveform of the macro having two phases which are the developing phase and the comparing phase.

During the developing phase, the input data are encoded into $V_{\rm in}[255:0]$ and fed into the array, generating the output current ($I_{\rm OUT}[0:31]$). Subsequently, $I_{\rm ACT}[15:0]$ are produced during the current development process. The switch SW1 is activated, initializing the ReLU-WTA by charging its output nodes ($\rm WTA_{OUT}[15:0]$) to the supply voltage ($V_{\rm DD}$).

Once the developing phase is complete, the comparing phase is triggered by activating switch SW2, enabling the WTA (WTA[15 : 0]) to compare the input currents ($I_{\rm ACT}[15:0]$). During this phase, the common wire voltage ($V_{\rm CM}$) is determined by the WTA cell with the largest input current (the winning cell). The bias current (I_B) flows through transistors M4 and M5 of the winning cell, maintaining its $\rm WTA_{OUT}$ charged to $V_{\rm DD}$. In contrast, the $\rm WTA_{OUT}$ of the losing cells is discharged to the ground (GND) by a current source of $I_B/2$.

As a result, only the $\rm WTA_{OUT}$ of the winning cell outputs $V_{\rm DD}$, directly representing the classification result.

Test-chip and measurements. Figure 1(d) describes the fabrication process of the HZO BEOL FeFET, which employs the standard 180 nm CMOS process with one polysilicon layer and six metal layers (1P6M) in a commercial foundry and an in-house developed HZO ferroelectric capacitor technology in a pilot process line. The HZO ferroelectric capacitors are positioned between the fifth (M5) and sixth (M6) metal layers. Each cell consists of a titanium nitride (TiN) bottom electrode, an HZO ferroelectric layer, and a TiN top electrode. Figure 1(d) also includes an elemental analysis map, illustrating the composition of the HZO ferroelectric layer. The read and write operation tables of the proposed FA array are provided in Appendixes A and B.

Figure 1(e) shows the current-voltage (I-V) characteristics of the HZO BEOL FeFET. It shows a large threshold-voltage ($V_{\rm TH}$)-shift, or memory window (MW), between the HVT and LVT cell, which is approximately 0.7 V. This large $V_{\rm TH}$-shift can ensure a sufficient signal ratio for highly parallel analog computing. Figure 1(f) indicates that the MW remains stable over 10^7 programming and erasing cycles without considerable degradation. This result confirms the good endurance characteristics of the HZO BEOL FeFET. It can potentially enable nvCIM to support frequent weight updating. Figure 1(g) shows the shmoo plot of the test chip. It operates at frequencies ranging from 4 to 48 MHz by adjusting the supply voltage ($V_{\rm DD}$) between 1.3 and 2.2 V. A peak energy efficiency (EF) of 171.34 TOPS/W can be achieved at 32 MHz with a $V_{\rm DD}$ of 1.7 V. Additionally, a peak throughput of 393.2 GOPS can be achieved at 48 MHz with a $V_{\rm DD}$ of 2.2 V. Figure 1(h) shows the die photo of the chip, fabricated using a standard 180 nm CMOS process in a commercial foundry and an in-house developed HZO ferroelectric capacitor technology in a pilot process line. The macro occupies an area of 0.74 $\rm mm^2$.

Further chip summary and comparison table are provided in Appendixes C and D. It achieves a cell endurance exceeding 10^7 cycles, a peak EF of 171.34 TOPS/W, a peak area efficiency of 531.35 $\rm GOPS/mm^2$, and a peak throughput of 393.2 GOPS.

Conclusion. This study presents an 8 kb HZO-based BEOL FeFET nvCIM macro, which supports neural networks with 256 inputs and 16 outputs. It employs a low-power and area-efficient ReLU-WTA circuit to perform nonlinear activation and analog classification, achieving competent cell endurance, energy- and area-efficiency reported among BEOL FeFET implementations. This work shows the HZO BEOL FeFET as a competent technology platform for implementing nvCIM macro for AI edge devices.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 92364202, U2441247, 62488101).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Chen W H, Dou C, Li K X, et al. CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors. Nat Electron, 2019, 2: 420–428
2 Wang L F, Li W Z, Zhou Z D, et al. A flash-SRAM-ADC-fused plastic computing-in-memory macro for learning in neural networks in a standard 14 nm FinFET process. In: Proceedings of IEEE International Solid-State Circuits Conference, San Francisco, 2024. 582–584
3 An J J, Wang L F, Ye W, et al. Design memristor-based computing-in-memory for AI accelerators considering the interplay between devices, circuits, and system. Sci China Inf Sci, 2023, 66: 182404
4 Wang Y, Tao L, Guzman R, et al. A stable rhombohedral phase in ferroelectric $\rm Hf(Zr)_{1+x}O_2$ capacitor with ultralow coercive field. Science, 2023, 381: 558–563
5 Salahuddin S, Ni K, Datta S. The era of hyper-scaling in electronics. Nat Electron, 2018, 1: 442–450
6 Khan A I, Keshavarzi A, Datta S. The future of ferroelectric field-effect transistor technology. Nat Electron, 2020, 3: 588–597