Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Eliminating Leakage in Volatile Memory with Anti-Ferroelectric Transistors

**Hongtao Zhong**
Tsinghua University
**Zijie Zheng**
National University of Singapore
**Leming Jiao**
National University of Singapore
**Zuopu Zhou**
National University of Singapore
**Chen Sun**
National University of Singapore
**Xiaoyang Ma**
Princeton University https://orcid.org/0000-0001-6633-9398
**Vijaykrishnan Narayanan**
Pennsylvania State University
**Huazhong Yang**
Tsinghua University
**Kai Ni**
Rochester Institute of Technology https://orcid.org/0000-0002-3628-3431
**Xiao Gong**
National University of Singapore
**Xueqing Li** ( xueqingli@tsinghua.edu.cn )
Tsinghua University

**Article**

**Keywords:**

**Posted Date:** April 17th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2744794/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** There is **NO** Competing Interest.

# Eliminating Leakage in Volatile Memory with Anti-Ferroelectric Transistors

Hongtao Zhong[1*], Zijie Zheng[2*], Leming Jiao[2], Zuopu Zhou[2], Chen Sun[2], Xiaoyang Ma[3], Vijaykrishnan Narayanan[4], Huazhong Yang[1], Kai Ni[5✉], Xiao Gong[2✉] and Xueqing Li[1✉]

[1]Department of Electronic Engineering, Tsinghua University, Beijing, China

[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

[3]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ, USA

[4]Department of Computer Science and Engineering, Pennsylvania State University, PA, USA

[5]Department of Electrical and Microelectronic Engineering, Rochester Institute of Technology, Rochester, NY, USA

[*]These authors contributed equally: Hongtao Zhong, Zijie Zheng;

[✉]To whom correspondence should be addressed; E-mail: kai.ni@rit.edu, elegong@nus.edu.sg, xueqingli@tsinghua.edu.cn

## Abstract

Cache serves as a temporary data memory module in many general-purpose processors and domain-specific accelerators. Its density, power, speed, and reliability play a critical role in enhancing the overall system performance and quality of service. Conventional volatile memories, including static random-access memory (SRAM) and embedded dynamic random-access memory (eDRAM) in the complementary metal-oxide-semiconductor technology, have high performance and good reliability. However, the inherent leakage in both SRAM and eDRAM hinders further improvement towards smaller feature sizes and higher energy efficiency. Although the emerging nonvolatile memories can eliminate the leakage efficiently, the penalties of lower speed and degraded reliability are significant. This article reveals a new opportunity towards leakage-free volatile static memory beyond the known paradigms of existing volatile and nonvolatile memories. By engineering a double-well energy landscape with the assistance of a clamping voltage bias, leakage-free and refresh-free state retention of volatile memory is achieved for the first time. This new memory is highlighted by both the ultra-low leakage of nonvolatile memories and the speed, energy, and reliability advantages of volatile memories. A proof-of-concept memory is demonstrated using in-house anti-ferroelectric field-effect transistors (AFeFETs), delivering an extrapolated endurance of about $10^{12}$ cycles, a retention time of over 10 years, and no subthreshold channel leakage current. Such a new concept of AFeFET-based memory enables an improved balance between density, power, and reliability beyond all existing memory solutions.

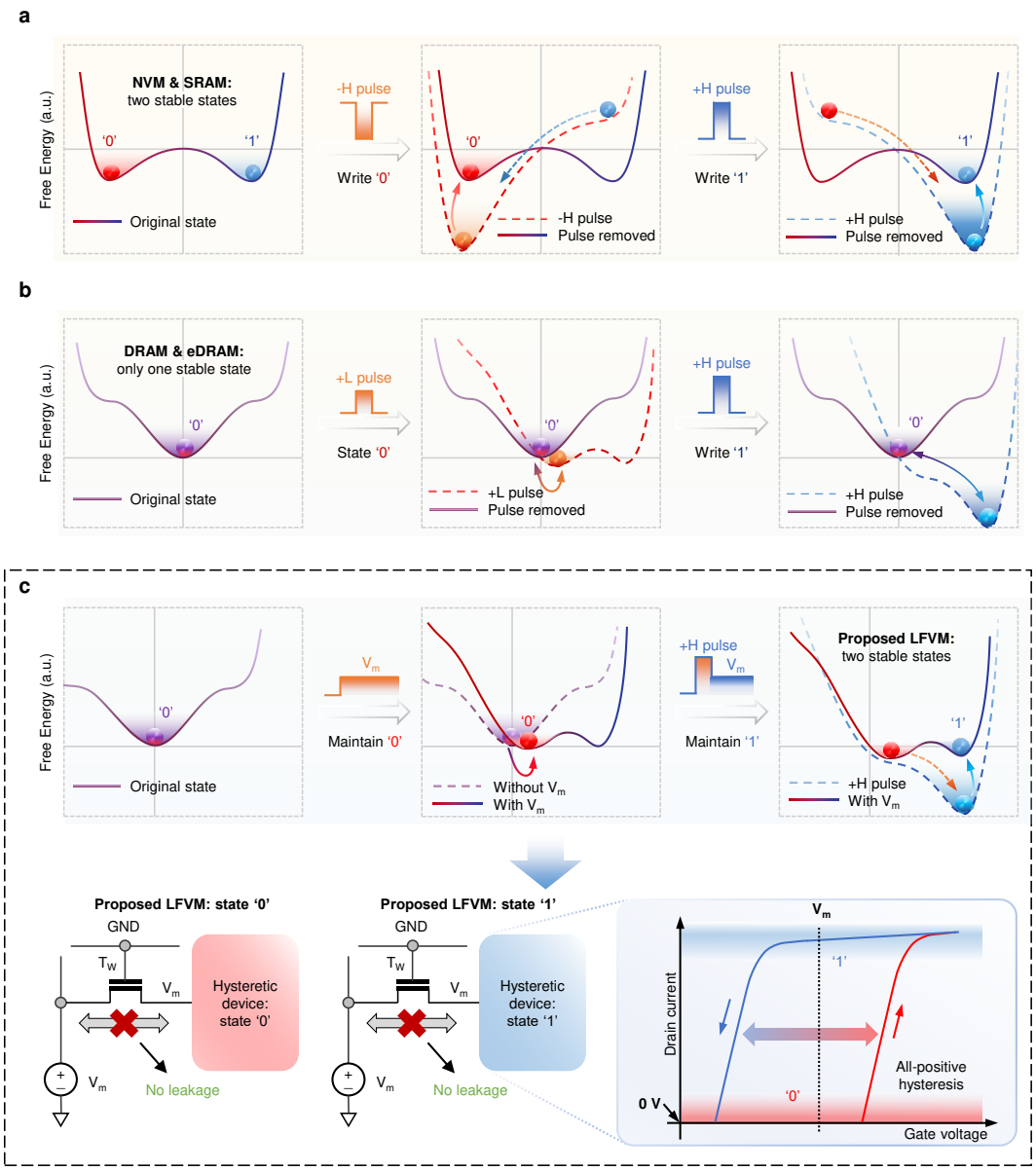



**Fig. 1 | Physical mechanism of different memories**. **a**, State maintaining and switching methods of NVM and SRAM: For NVM, the energy landscape is double-well with two local minimum energy states '0' and '1'. To switch between the states, a high negative or positive pulse is applied. After the pulse is removed, the state will be stabilized to the local minimum energy state '0' or '1'. The energy landscape of SRAM is also double-well but the state switching for SRAM involves VDD and GND only. **b**, State maintaining and switching methods of DRAM and eDRAM: There is only one local minimum energy state '0'. To switch from '0' to '1', a high positive pulse is applied. Returning to '0' involves GND only and does not need a high negative pulse. It is observed that the required voltage range is relatively small. However, the state '1' is unstable and it will return to '0' eventually due to leakage, if the external pulse is removed. Therefore, periodic refresh is adopted in practice to maintain the data. **c**, Proposed LFVM memory based on the hysteretic volatile devices. For a category of volatile devices, with the proposed LFVM operation, both states '1' and '0' can be maintained using the same voltage configuration. Besides, within one such memory cell, it consumes no subthreshold drain-source leakage power, thanks to the same applied voltage at the drain and the source of the turned-off write access transistor. This is achieved by exploiting the volatile hysteresis memory window.

## Introduction

A large, fast, and reliable cache near the processing units boosts the system performance and energy efficiency for data-intensive computing. In the past few decades, the static random-access memory (SRAM) and the embedded dynamic random-access memory (eDRAM) have been the major adopted cache technologies to balance capacity, energy efficiency, reliability, etc. However, as the complementary metal-oxide-semiconductor (CMOS) technology scaling slows down, it has been increasingly challenging to provide larger cache capacity to meet the need of the unprecedented amount of data and computation demands by the emerging applications of Internet of Things[1], computer vision[2], cloud computing[3], etc.

Essentially, both SRAM and eDRAM in CMOS are volatile memories, and the leakage issue is a significant challenge that limits their performance and density scaling for future generations. For example, for SRAM, as the logic transistors approach the physical limits, the increasing subthreshold channel leakage current makes SRAM power-hungry[4-6]. Such an inherent physical mechanism hinders further improvement of the embedded cache capacity and also the performance in many low-power devices such as the Internet of Things and wearable gadgets[7-9]. It is also reported recently that the 3nm-node SRAM has very limited density improvement over the 5nm-node SRAM[10], indicating the challenge of the conventional CMOS process development for SRAM density scaling. For eDRAM, the continual device scaling towards higher density results in a severe challenge in maintaining the refresh rate. Meanwhile, although some emerging nonvolatile memories (NVMs) may potentially deliver higher density without leakage current issue, the inherent penalties of degraded speed and/or reliability are still limiting their application[11]. Given that, developing a new type of energy-efficient, fast, and reliable volatile memory is highly desired and is the focus of this work.

The excitation originates from the re-visiting of the fundamental memory state representation with the underlying energy landscape of a memory device. For both NVM and SRAM, as illustrated in Fig. 1**a**, the energy landscape can be approximated as a double-well potential, where the two local energy minima correspond to the two stable states. Such a landscape is shaped by the powered-on inverter loop in an SRAM at the cost of leakage currents or the inherent energy barrier between states in NVM devices. Initialized at one state, the memory switches to the other state upon a write pulse as the energy landscape tilts and the initial state is no longer stable. After the removal of the

write pulse, the double-well energy landscape is restored and the memory stays at the new state.

Differently, in the case of DRAM, including the embedded DRAM (eDRAM), as illustrated in Fig. 1**b**, the most stable state is when the capacitor is free of charge, which corresponds to the energy landscape with a single energy minimum. When applied with a write pulse to charge up the capacitor, a temporary double-well potential emerges and switching to the other state proceeds. However, with the removal of the external write pulse and the dominance of the leakage current, such a transient energy landscape collapses back into a single energy minimum upon removal of the write pulse. Consequently, refresh operations become necessary to preserve the DRAM memory states.

A fundamental question has thus emerged: is it possible to engineer the volatile memory devices to have a double-well energy landscape, and maintain the memory states without the costs of refresh activities and static holding power? Inspired by this question, this work proposes a new class of leakage-free volatile memory (LFVM), featured with ultra-low-power refresh-free static data retaining capabilities. It exploits the double-well energy landscape with the assistance of a clamping voltage. With the assisting clamping voltage applied to the device gate, the stable double-well energy landscape is maintained. This concept is shown in Fig. 1**c**, in which a write-access transistor ($T_w$) and a hysteretic device are illustrated. The gate of the hysteretic device and the bitline connected to $T_w$ are both biased at $V_m$. Thanks to the stable hysteresis of the hysteretic device, the two memory states can be maintained. One highlight of this new memory concept is that, with the same drain and source voltage in $T_w$, there is no subthreshold leakage current between the drain and the source during the idle state. This feature leads to ultra-high standby energy efficiency. Another highlight is the simplified peripheral control and circuitry. This is because no refresh operation is needed, and the memory-state-independent assisting gate bias can be applied globally to all cells in the array. Lastly, this general memory structure has a much higher density than SRAM with fewer transistors, while also providing the static random access capability.

Given those intriguing highlights, this article further explores the design of candidate hysteretic devices for the proposed LFVM concept. It is exciting to see that both anti-ferroelectric field-effect transistors (AFeFETs) and nanoelectromechanical (NEM) relays fit well with the goal. NEM relays rely on the electrostatically actuated mechanical switching mechanisms[12,13]. The combination of the elastic force, the van der Waals force, and the electrostatic force exhibits the desired hysteresis in Fig. 1**c**. Currently, NEM relays are being actively developed to integrate and scale with the CMOS

technology. In this article, the proof of concept is based on AFeFETs[14,15]. As a derivative of ferroelectric FET (FeFET), AFeFET integrates an anti-ferroelectric (AFE) film into the gate stack of a metal-oxide-semiconductor field effect transistor (MOSFET), which exhibits the hysteresis in the positive gate voltage region, as shown in Fig. 1**c**, with a high ON/OFF ratio, low write power, and compatibility with advanced CMOS (theoretical physical mechanisms are included in the supplementary materials). More importantly, the AFE devices exhibit a significantly improved endurance over their ferroelectric (FE) counterparts, which has been experimentally demonstrated by several recent works[15-17]. Actually, the main challenge for AFeFET in replacing FeFET is the volatility of AFE materials, where the remnant polarization is zero upon the removal of the external gate bias. Although techniques such as DRAM-like periodic refresh or built-in electric field strategy[17,18] can be adopted, the overheads are high. The former approach comes with pricy refresh power consumption and normal access stalls. The latter of engineering built-in bias into the AFE thin film faces constraints by limited tuning range and flexibility[17,18]. This is because the work functions of available CMOS-compatible metals are difficult to cover a wide range, especially considering the Fermi-level pinning. Therefore, the proposed LFVM solves the volatility issue of existing AFeFET memories, and provides a new opportunity to exploit the benefits of high endurance, scalability, and energy efficiency.

Leveraging the appropriate AFE and channel materials, the proposed AFeFET can be integrated within the back-end-of-line (BEOL) thermal budget (≤450 °C), thus allowing multi-tier memory stacking. In this work, we experimentally demonstrate a BEOL-compatible AFeFET using a low thermal budget, wide bandgap, and high electron mobility amorphous-indium-gallium-zinc-oxide (a-IGZO) channel and the BEOL-compatible and scalable $Hf_{1-x}Zr_xO_2$ FE/AFE material system. Our AFeFETs were realized using the metal-ferroelectric-metal-insulator-semiconductor (MFMIS) structure with the flexibility to optimize the gate stack electric field, featuring an extrapolated endurance of about $10^{12}$ cycles, a retention time of more than 10 years with the proposed LFVM operation, a high ON/OFF ratio, and small device-to-device variation. In contrast to the previous impression of impracticality in using AFE devices for memory purposes, we believe that the proposed LFVM opens up a new paradigm for power-efficient and high-performance AFE memory.

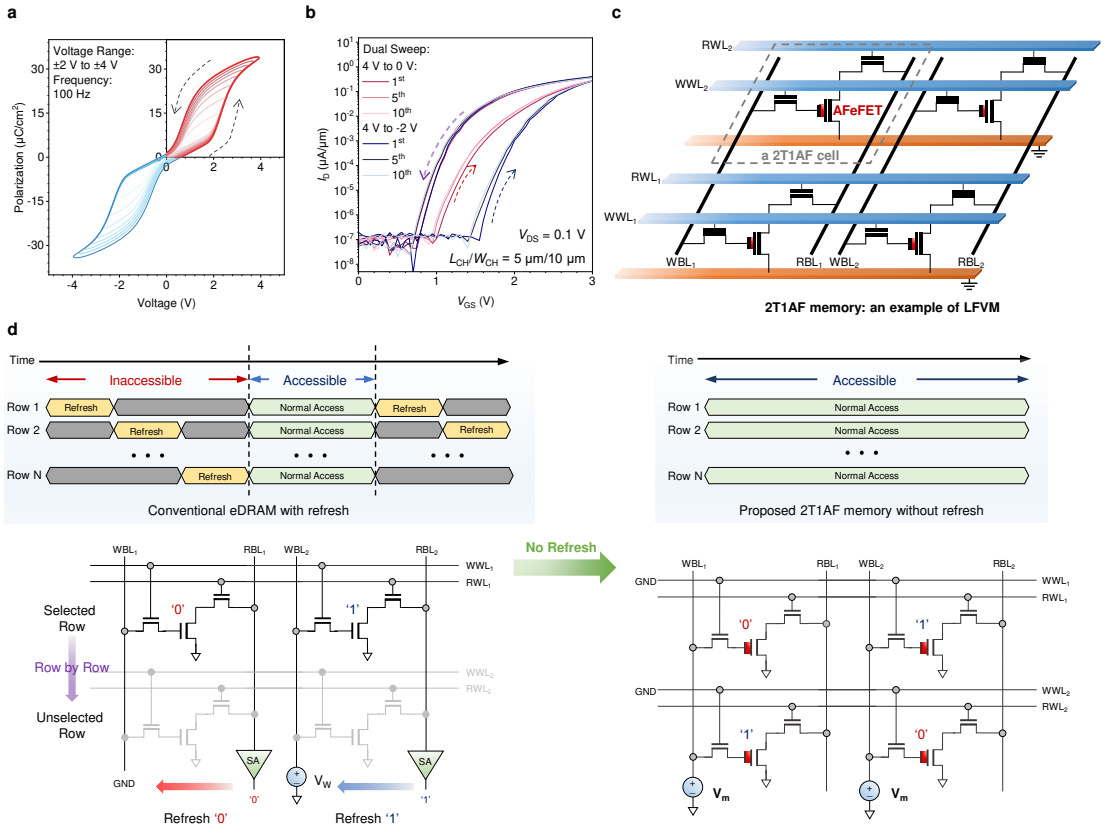



**Fig. 2 | Device electric characteristics and 2T1AF memory. a**, *P*-*V* curves of the $Hf_xZr_{1-x}O_2$-based AFE capacitor. There is a double-hysteresis loop, and the polarization returns to zero at zero voltage, which shows the volatility. **b**, $I_D$-$V_{GS}$ curves measured on the fabricated a-IGZO AFeFETs. The hysteresis loops for the 1st, 5th, and 10th cycle of measurement are shown with repeated and all-positive hysteresis windows, which also implies volatility. **c**, Proposed 2T1AF memory. The write access transistor is controlled by the write word line (WWL), and the input write datum is sent through the write bit line (WBL) to program the AFeFETs. The 2T1AF memory cell evolves from the conventional 3T eDRAM cell with the transistor for data storage replaced by an AFeFET. The 2T1AF memory benefits from non-destructive read and negligible read and write disturbance with the extra access transistors. **d**, Difference in data maintaining method between the 3T eDRAM and the 2T1AF memory. The 3T eDRAM utilizes the conventional refresh operation to maintain data. The refresh operation reads out data and then writes them back row by row. The refresh operation not only consumes high refresh power, but also stalls the normal access requests. The 2T1AF memory utilizes the LFVM operation to maintain data. The states '0' and '1' share the same voltage configuration, so all data stored in the memory array can be maintained simultaneously. The LFVM operation not only consumes ultra-low power but also stalls no access requests. Therefore, the accessible time window for the memory array is extremely extended.

## LFVM Operation with AFeFET in 2T1AF Memory Cell

To reveal how AFeFETs can be exploited to enable LFVM, the AFeFET operation principles are discussed first. It leverages the versatile $Hf_{1-x}Zr_xO_2$ material system, where the parameter $x$ yields mixed FE and AFE characteristics[19]. When $x$ increases from 0.5 to 1, $Hf_{1-x}Zr_xO_2$ with strong ferroelectricity will become more dominated by anti-ferroelectricity with Zr-rich composition. As shown in Fig. 2**a**, the polarization-voltage ($P$-$V$) curves measured on fabricated AFE capacitors show a typical volatile behavior with pinched hysteresis. For Zr-rich $Hf_{1-x}Zr_xO_2$, such characteristic originated from a phase transition induced by the external electric field, which can be observed in both positive and negative branches[20,21]. As introduced earlier, DRAM-like periodic refresh and built-in bias introduction have been reported to address the volatility issue of capacitor-based memory. In addition to the challenges introduced earlier, the effectiveness of those approaches remains unclear for transistor-based memory. By replacing the normal gate dielectric with an AFE thin film, typical AFeFETs can be formed, which also exhibit volatile characteristics. Like the measured drain current to gate voltage ($I_D$-$V_{GS}$) characteristics for a fabricated IGZO AFeFET shown in Fig. 2**b**, an anti-clockwise hysteresis is observed during the 0 V – 4 V – 0 V sweep. When $V_{GS}$ returns to 0 V, the high current state is gone, and the next sweep follows the same path with a more positive threshold voltage ($V_{TH}$), indicating volatility. In practice, the hysteresis window might be limited in the 0 V – 4 V – 0 V sweep, and a small negative voltage, i.e., -2 V in this article, can be applied to enlarge the hysteresis window for a larger initial ON/OFF ratio. The main reason is due to the non-zero voltage drop in the AFE layer with $V_{GS}$ = 0 V, which will be further analyzed in the next section. To address the volatility issue, in this work, we will apply the proposed LFVM operation and demonstrate its intriguing memory operation.

From the array perspective, the LFVM operation in AFeFETs can be understood and evaluated. While denser memory cells are possible, a 2-transistor-1-AFeFET (2T1AF) disturb-free memory cell and corresponding 2T1AF array are taken as an example to illustrate the LFVM operation, as shown in Fig. 2**c**. The 2T1AF cell consists of a write access transistor, a data storage transistor, and a read access transistor. Its structure is similar to that of 3T eDRAM, with the transistor for data storage replaced by an AFeFET. The write and read operations of the 2T1AF memory cell are also similar to those of the 3T eDRAM cell. Detailed voltage configuration is shown in Supplementary Fig. 2 and Supplementary Fig. 3. The use of write access transistor and read access transistor allows

non-destructive read, and prevents write and read disturbance. The peripherals can detect the voltage change on the read bitline (RBL) to sense the stored data.

The main difference between the 3T eDRAM and the 2T1AF memory is the data retention mechanism. The 3T eDRAM utilizes the periodic two-step refresh operation consisting of data read and write-back, as illustrated in Fig. 2**d**. Because the stored data are usually different between rows, the refresh operation is executed row-by-row. Conventional refresh operation not only consumes high refresh power, but also occupies the access time window and degrades the memory performance due to the access stalls. As the memory size increases, the refresh power consumption and performance degradation due to refresh become worse[22]. On the contrary, the 2T1AF memory utilizes the proposed LFVM operation to maintain data. Unlike the refresh operation, the LFVM operation can maintain the cell states '0' and '1' using the same bias voltage configuration. Therefore, the read-and-then-write-back refresh operation is not necessary, and all data stored in the memory array can be maintained. Besides, the LFVM operation itself consumes no subthreshold drain-source leakage and does not stall normal access. This feature saves power and prevents refresh-originated performance degradation, which is highly desired by many low-power applications.

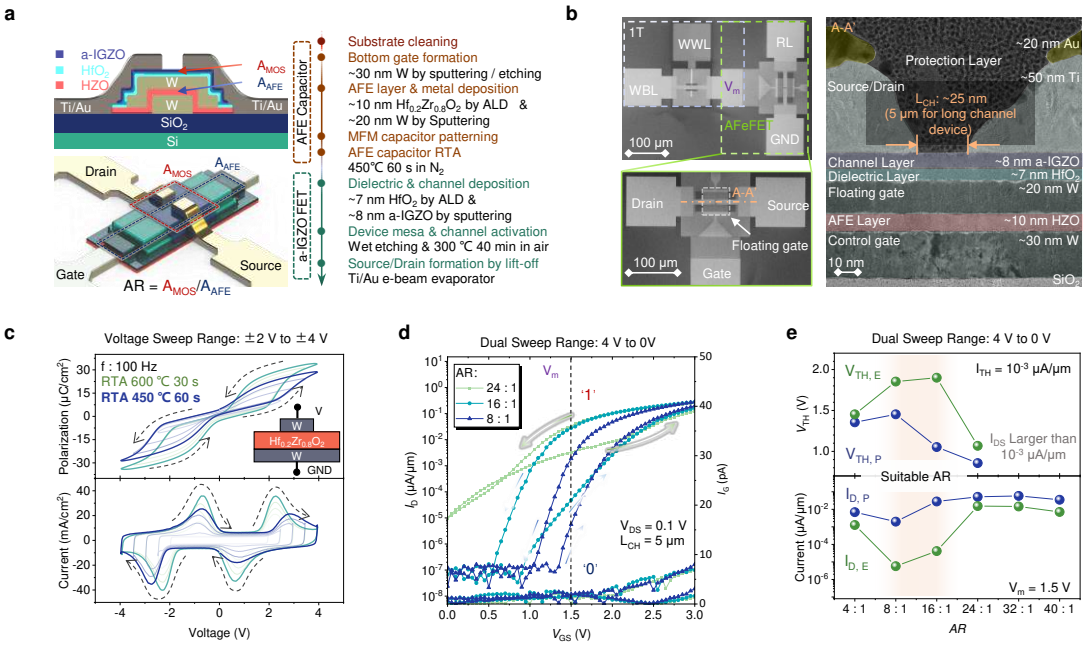



**Fig. 3 | Fabricated device and memory cell with basic characterizations. a**, Device structure and key process steps of the a-IGZO AFeFETs. The cross-sectional view illustrates a gate stack of IGZO/$HfO_2$/W/HZO/W from top to bottom. The schematic illustration utilizes the MFMIS structure. $A_{MOS}$ and $A_{AFE}$ are areas of the AFE layer and the MOSFET, respectively, and the voltage distribution across the gate stack can be tuned by changing the area ratio (AR) of $A_{MOS}$ to $A_{AFE}$. $L_{CH}$ represents the gap between the drain and the source. **b**, The top-view SEM image of the a-IGZO AFeFET and 2T1AF memory, and the TEM image in the channel region with the thickness of each layer. The read access transistor is regarded as a part of the source measure unit (SMU) to detect the AFeFET current. **c**, The *P*-*V* double-hysteresis loops and the *I*-*V* loops of the AFE capacitor with different RTA temperatures. Typical AFE characteristics are validated for both conditions. **d**, $I_D$-$V_{GS}$ curves of the a-IGZO AFeFETs with 5 μm $L_{CH}$ and various ARs. The anti-clockwise hysteresis can be observed with the volatile property induced by the AFE layer within the gate stack. **e,** The $V_{TH}$ with fixed current criteria and $I_D$ at $V_m$ of the a-IGZO AFeFETs with various ARs. The characteristics of the AFeFETs can be tuned by AR, and 8:1 and 16:1 are suitable for the proposed memory cell with a relatively large memory window (MW) and a large ON/OFF ratio.

## AFeFET and 2T1AF Memory Demonstration

Fig. 3**a** summarizes the cross-sectional view, the schematic illustration and the key process steps of the fabricated a-IGZO AFeFET (more details are referred to Supplementary Fig. 4). The AFeFETs use the MFMIS structure with the flexibility to engineer the area ratio (AR) of the MOSFET area ($A_{MOS}$) to the AFE area ($A_{AFE}$) to tune the overall characteristics of the device. Besides, the staggered bottom-gate structure with a top-contacted drain/source is utilized, and $L_{CH}$ is determined by the gap between the drain and the source. The a-IGZO thickness is about 8 nm, and the gate stack comprises 7 nm $HfO_2$ (dielectric layer)/20 nm W (floating gate)/10 nm $Hf_{0.2}Zr_{0.8}O_2$ (AFE layer)/30 nm W (control gate) from top to bottom. Fig. 3**b** shows the scanning electron microscope (SEM) image of the fabricated AFeFET device and the 2T1AF memory cell. The read access transistor is regarded as a part of the source measure unit (SMU) to detect the AFeFET current, so only the write access transistor and the AFeFET device are fabricated to demonstrate the functionality of the LFVM operation. The thicknesses of each layer mentioned above in the channel region (A-A') are confirmed by the high-resolution transmission electron microscopy (HRTEM) image.

Fig. 3**c** illustrates the *P*-*V* and *I*-*V* characteristics of the AFE capacitors with a MIM structure to show the anti-ferroelectricity. The capacitors are fabricated using the same process as the AFeFETs, as shown in Fig. 3**a**. Note that, in the *P*-*V* curves, in contrast to the FE capacitors, there is negligible remnant polarization ($P_r$) at 0 V, implying the volatility of AFE materials. In the *I*-*V* curves, a double current peak in both positive and negative branches corresponding to a pinched hysteresis can be observed. Meanwhile, the rapid thermal annealing (RTA) with both 450 ℃ 60 s and 600 ℃ 30 s can provide strong AFE behavior, indicating the BEOL compatibility of the HZO for AFE applications. With relatively higher annealing temperature, the voltage for AFE to FE phase transition becomes smaller, which might lead to lower operation voltage, as shown in Fig. 3**c**. Fig. 3**d** shows the $I_D$-$V_{GS}$ curves of the a-IGZO AFeFETs ($L_{CH}$ = 5 μm, $V_{DS}$ = 0.1 V) with various ARs of 8:1, 16:1, and 24:1 under dual sweep measurement from 4 V to 0 V. Due to the electrically coupled AFE capacitor on the MOSFET gate, the $I_D$-$V_{GS}$ curves of the AFeFETs exhibit volatile anti-clockwise hysteresis.

Tuning AR is an effective approach to optimize the device characteristics for the proposed memory design, which is equivalent to adjusting the capacitance ratio between the AFE MIM capacitor and the MOS capacitor. Because the two capacitors are in series, the electric field distribution in the AFE thin film can be modulated. As a result, a larger AR results in a larger voltage

drop across the AFE layer, leading to a wider hysteresis window or memory window (MW) in the $I_D$-$V_{GS}$ curves, where MW is defined as the $V_{TH}$ difference between $V_{TH,\,P}$ of the programmed state '1' and $V_{TH,\,E}$ of the erased state '0', at a constant current of $10^{-3}$ μA/μm.

However, although a sufficient AR is necessary for a large MW, the experimental results indicate that the AR of suitable AFeFETs should be selected within a reasonable range for the proposed LFVM. As shown in Fig. 3**d**, the devices with a large AR of 24:1 exhibit a low ON/OFF ratio that is defined as the ratio of $I_D$ at the programmed state to the erased state. This can be understood with consideration of the capacitance matching between the MIM AFE capacitor and the MOS capacitor. The small absolute polarization of AFE capacitors with a small area cannot create enough change in the channel carrier density. As a result, there exists an optimal AR range to maximize both the MW and the ON/OFF ratio simultaneously. More detailed $I_D$-$V_{GS}$ and $C$-$V_{GS}$ characteristics of the fabricated AFeFETs are referred to Supplementary Fig. 5.

Fig. 3**e** summarizes $V_{TH}$ and $I_D$ at $V_m$ of 1.5 V for both states with various ARs. 1.5 V is set for $V_m$ with consideration of the impact of $V_m$ on both the ON/OFF ratio and the leakage power referred to Supplementary Fig. 6. Here, the data are extracted from the abovementioned $I_D$-$V_{GS}$ curves to show the impact of AR on MW and the ON/OFF ratio. It can be observed that as AR increases within a certain range, MW and the ON/OFF ratio improve because of more saturated switching of the AFE layer. The AFeFETs with AR from 8:1 and 16:1 satisfy the proposed memory cell design, which realizes MW and the ON/OFF ratio of ~1 V and ~700 for the fabricated devices, respectively.

Moreover, due to the non-zero flat-band voltage in the gate stack and inevitable charge trapping within the floating gate, it is possible that the back switching from the FE phase to the AFE phase can be displaced from $V_{GS} = 0$ V. In some cases, a negative gate bias may be needed to reverse the FE phase back to the AFE phase. Therefore, as shown in Supplementary Fig. 5**a**, MW without negative biases is relatively narrower in most cases, corresponding to the minor loop operation of the AFeFETs, as illustrated in Supplementary Fig. 7**c**. Hence, a negative write voltage for erase operation will be helpful for the memory operation to enable a large ON/OFF ratio and a large MW.

Moreover, the device variation and scalability are further explored. We found that the fabricated a-IGZO AFeFETs exhibit excellent device-to-device uniformity, and scaled devices maintain proper function with $L_{CH}$ down to ~50 nm with boosted $I_D$ and negatively shifted $V_{TH}$. The experimental results are referred to Supplementary Fig. 8.

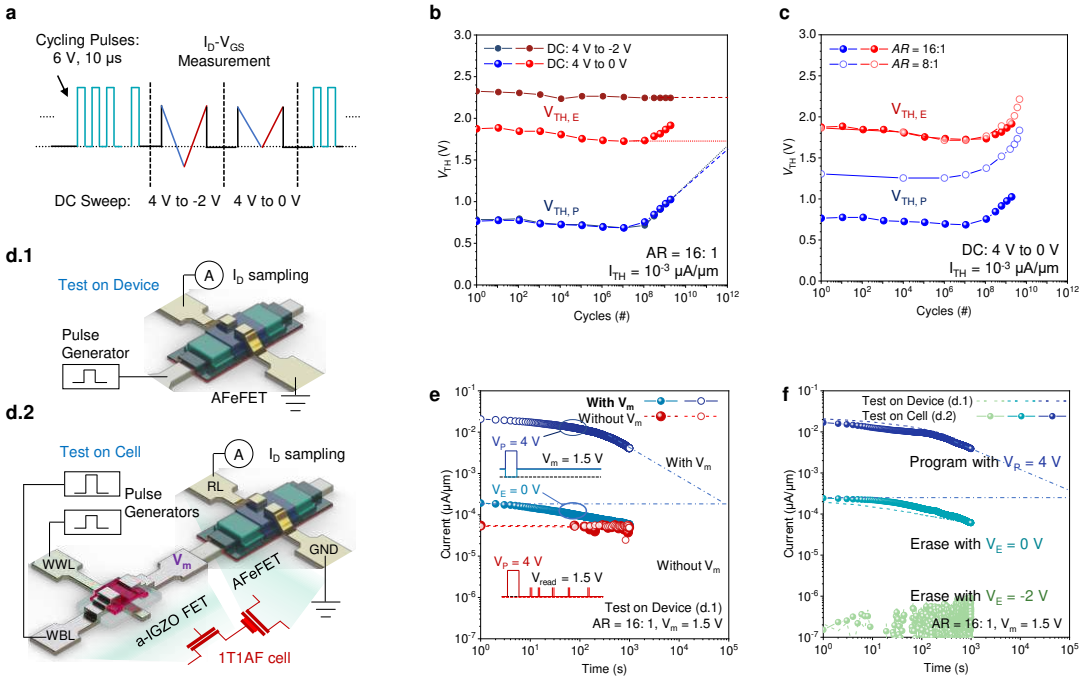



**Fig. 4 | Endurance and retention measurement of the AFeFETs and the memory cell. a**, Schematic of the endurance measurement using unipolar cycling pulses. The DC sweep ranges include both 4 V to -2 V and 4 V to 0 V. **b**, Endurance measurement of the AFeFETs with AR of 16:1, showing high endurance of about $10^{12}$ cycles with the 4 V to -2 V operation and $10^{10}$ cycles with the 4 V to 0 V operation. Here, the endurance limit is defined as the extrapolated MW degradation to half the original. **c**, Endurance measurement of the AFeFETs with ARs of 8:1 and 16:1. The endurance property of the devices with 16:1 AR is better due to the relatively wider MW. **d**, Schematic of the retention measurement for testing on both a single AFeFET device and the LFVM cell. **e**, Retention measurement of the AFeFETs with and without $V_m$ during the standby period. Without $V_m$, the AFeFETs cannot be used as a static memory due to the volatile and dynamic characteristics of the AFE layer, as shown with the two red lines. Here, the read voltage is set to 1.5 V, equal to $V_m$. In contrast, the blue lines are obtained by applying $V_m$ = 1.5 V and sampling $I_D$ after the program and the erase operations. Two distinct states can be observed with the proposed operation scheme. **f**, Retention measurement in the memory cell showing results similar to the device retention and proving a successful operation of the proposed memory cell. With the unipolar operation, the retention time is estimated to be longer than $10^4$ s. By applying a low negative voltage ($V_E$ = -2 V) for erase operation, the extrapolated retention time over 10 years can be achieved.

## Measurement and benchmarking of the 2T1AF Memory

The combination of AFE materials and the proposed LFVM operation yields a memory device with good reliability. Fig. 4**a** illustrates the endurance cycling waveform applied to the gate of the AFeFETs. Unipolar cycling pulses of 6 V and 10 μs are applied to the device gate, followed by the DC sweep ranging from both 4 V to -2 V and from 4 V to 0 V, to obtain the $I_D$-$V_{GS}$ curves. Fig. 4**b** shows the measurement results of the AFeFETs with an AR of 16:1. Our devices maintain proper function after more than $10^9$ cycling without dielectric breakdown or significant MW degradation. This proves that using the AFE material with unipolar stressing for the devices can significantly improve the endurance compared with its FE counterpart[23]. Generally, charge injection into the FE or AFE layer is regarded as one of the primary culprits causing fatigue, and several studies attribute the higher endurance of AFE material to the lower local depolarization field during the switching transient that prevents severe injection effects[24,25].

Besides, as mentioned above, a larger MW can be observed with a DC sweep from 4 V to -2 V because of the more saturated switching of the AFE layer. If we quantify the measured endurance limit when the extrapolated MW is degraded by half, the endurances for the 4 V to -2 V operation and the 4 V to 0 V operation are in the order of $10^{12}$ and $10^{10}$, respectively. Fig. 4**c** shows the endurance measurement of the AFeFETs with AR of both 16:1 and 8:1. With a larger MW, the endurance of the devices with AR of 16:1 is improved. It also shows a parallel $V_{TH}$ shift for both states, suggesting that the absolute $V_{TH}$ shift is the main degradation, rather than the MW shrinking. Since there is no significant imprint of the AFE capacitor as shown in Supplementary Fig. 9, it is more likely caused by the charge injection to the floating electrode or bias temperature instability (BTI) of the IGZO channel, which causes parallel $V_{TH}$ shift for both states. These might be mitigated with more careful device design and optimized device fabrication.

Fig. 4**d** depicts the schematic of the retention measurement setup at both the device level and the memory cell level. To ensure the proper write and read operation for both the device and the memory cell, the voltage pulses utilized here have a fixed pulse width of 10 ms, which is long enough to charge the capacitor and switch the AFE layer. The program pulse amplitude is 4 V, while the erase pulse amplitude varies from 0 V to -2 V. For the proposed design, after each memory write, a hold bias $V_m$ = 1.5 V is applied to hold the memory state. This bias amplitude is selected to avoid significant AFE switching during the hold period so that proper operation with a relatively large

ON/OFF ratio and long retention time can be realized. Fig. 4**e** depicts the retention measurement results with and without $V_m$ during the standby period. It is evident that without $V_m$, the AFeFETs lose the programmed state, and cannot be used as a memory, as suggested by the merged red curves. However, by applying $V_m$, the memory states can be kept in AFeFETs, and two distinct states can be observed with a decent ON/OFF ratio and long retention time.

At last, we demonstrate a successful operation of the memory cell with similar voltage conditions while the write access transistor is turned on during the write operations and turned off during the idle hold time. The transistor gate is set to be floating to eliminate the gate leakage (~10 pA) from the write access transistor during the retention measurement. Fig. 4**f** shows that with the unipolar operation, the retention time is less than $10^5$ s, shorter than the case where $V_E$ = -2 V is applied for memory erase. As mentioned above, it is mainly due to the non-saturated depolarization of the AFE layer at 0 V gate voltage. With the applied negative erase pulse, $I_D$ in the erase state ($I_{D,E}$) decreases and the ON/OFF ratio increases. Therefore, the memory retention improves accordingly with over 10 years of extrapolated retention time. The experimental results have verified the feasibility of the proposed LFVM operation and device design based on AFeFETs. Further scaling and process optimization can be conducted to achieve better performance in terms of speed and energy efficiency.

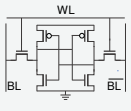
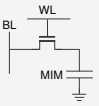
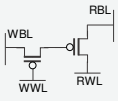
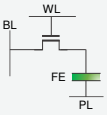
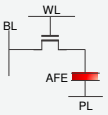
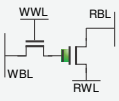
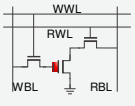
One potential application of the 2T1AF LFVM is the embedded memory integrated with the logic circuits on the same chip for data-intensive edge-computing applications[7-9]. Unlike the stand-alone memories, embedded memories can provide lower access latency but are more limited by the die area. The most common embedded SRAMs (eSRAMs) feature fast speed, mature fabrication, good reliability, high standby leakage power and a large cell footprint of 120-150 $F^2$, where F is the feature size. The eDRAMs, including the capacitor structures (e.g., 1T1C[26] and 2T1C[27]) and transistor-only structures (e.g., 2T[28], 3T[29,30] and 4T[31]), are denser but face the challenges of high power dissipation and performance degradation due to frequent refresh operations.

The embedded NVM (eNVM) can eliminate leakage efficiently. There have been several NVM technologies being produced or developed, including the floating-gate (FG) transistors[32], the spin transfer torque magnetic RAMs (STT-MRAM)[33], the phase change RAM (PCM)[34,35], the resistive RAM (RRAM)[36,37], the ferroelectric capacitors (FeRAM)[38,39], and FeFET[40-42], etc. These NVM cells can store at least two states even when power is off by harnessing different physical mechanisms. However, these devices also face challenges. The FG transistors face the challenges of limited

scalability, high voltages, and low endurance. The resistive NVMs, including STT-MRAM, PCM and RRAM, have high write power and limited write performance because of the current-driven write mechanism. Besides, large variations and a low ON/OFF ratio of the resistive devices make it difficult to distinguish different states, further increasing the error rate[43,44]. FeRAM and FeFET use an electric field and consume no DC power during the write operation. Therefore, the write energy can be significantly reduced. However, FeRAM suffers from the destructive read operation, while FeFET suffers from relatively low endurance[45] and switching stochasticity with device scaling[46].

As summarized in Table 1, the demonstrated 2T1AF memory shows a good balance among various metrics, including density, disturbance, power consumption and endurance. The 2T1AF memory has only 3 transistors without additional capacitors, so the area of the 2T1AF memory is much smaller than the conventional 6T SRAM. With the write and read access transistors, the 2T1AF memory also features non-destructive read operation. Besides, thanks to the DC-free capacitive load, the AFeFET write energy is also ultra-low. With the LFVM operation, no periodic refresh or normal access stall exists in the 2T1AF memory, and the actual standby power approaches zero unless with a non-ideal leaky gate. More detailed comparisons with SRAM and eDRAM are included in Supplementary Fig. 10. Furthermore, the measured endurance of the in-house AFeFET can be more than $10^9$ cycles, while the extrapolated endurance can reach $10^{12}$ cycles, much larger than optimized FeFETs with the same channel structure (~$10^8$ cycles)[23].

It is important to point out that the proposed 2T1AF memory based on AFeFET is just one viable implementation of the LFVM memory. Other designs with different hysteretic devices, e.g., NEM relays[12,13], or different memory circuit topologies, e.g., 2T structure without the read access transistor, can also be explored in the future. Although the fabricated AFeFET device in this article was targeted at demonstrating functionality, rather than optimal performance, recent AFE-related works have shown that the AFE-based devices can embrace high speed, high endurance, low operating voltage, long retention time, etc[15,17]. Besides, the AFE-based memory can also be utilized to achieve multi-bit storage for even higher density[47]. Furthermore, the LFVM memory can also be applied to applications including compute-in-memory (CiM)[48], DNA sequence alignment[49], etc.

| | eSRAM | eDRAM (capacitor)[26,27] | eDRAM (transistor)[28-31] | FeRAM[38,39] | AFeRAM[15-17] | FeFET[40-42] | Proposed 2T1AF memory |
|---|---|---|---|---|---|---|---|
| **Cell structure** |  |  |  |  |  |  |  |
| **Status** | Available [a] | Available | Available | Available | Research [b] | Research | **Research** |
| **Cell size** | 120-150 $F^2$ | 40 $F^2$ | 40-60 $F^2$ | 30-40 $F^2$ | 30-40 $F^2$ | 10-30 $F^2$ | **60 $F^2$** |
| **Non-volatile** | No | No | No | Yes | Yes/No (with/without built-in field) | Yes | **No** |
| **Non-destructive read** | Yes | No | Yes | No | No | Yes | **Yes** |
| **Write voltage** | <1 V | <1 V | <1 V | 1-3 V | 1-3 V | 1.5-4 V | **4 V** |
| **Write energy** | ~1 fJ | ~1 pJ | ~1 fJ | ~100 fJ | ~100 fJ | 1-10 fJ | **1-10 fJ** |
| **Standby power** | High | Medium | Medium | Low | Low | Low | **Low** |
| **Write speed** | <1 ns | >10 ns | 1-5 ns | 1-10 ns | 1-10 ns | 1-10 ns | **1-10 ns** |
| **Read speed** | <1 ns | >10 ns | 1-5 ns | 1-25 ns | 1-25 ns | 1-10 ns | **1-10 ns** |
| **Retention** | Static [c] | Refresh | Refresh | 10 years | 10 years/Refresh (with/without built-in field) | 10 years | **10 years (static with ideal LFVM)** |
| **Endurance** | $>10^{16}$ | $>10^{16}$ | $>10^{16}$ | $>10^{12}$ | $10^9$-$10^{12}$ | $10^5$-$10^{9}$ [d] | **~$10^{12}$** |

[a] Available: The memory has been produced or manufactured.

[b] Research: The memory is still being researched.

[c] Static: The stored data keep maintained with a power supply.

[d] Some FeFETs with special structures (e.g., the back-gated structure[50]) can achieve $>10^{10}$ cycles.

Table 1. Performance comparison between different embedded memories.

## Conclusion

In this article, we propose a novel leakage-free volatile memory (LFVM) based on hysteretic devices with all-positive hysteresis. From the energy landscape perspective, the proposed LFVM memory has a stable double-well energy landscape with a holding voltage bias, which can eliminate the excess subthreshold drain-source leakage current and retain the advantages of the volatile memories. At the cell level, we also propose corresponding LFVM operation that can maintain stored data without the need of refresh operations. We have experimentally demonstrated the LFVM memory in a 2T1AF cell topology based on BEOL-compatible a-IGZO AFeFETs. The proposed 2T1AF memory has much higher endurance than FeFETs by harnessing the inherent AFeFET device capability. Besides, the 2T1AF memory also stands out with ultra-long retention time. Therefore, the LFVM memory opens a new category for a series of volatile memories, and is promising for low-power and endurance-sensitive scenarios.

## Reference


1. Da Xu, L., He, W. & Li, S. Internet of things in industries: A survey. *IEEE Trans. Ind. informatics* **10**, 2233–2243 (2014).
2. Voulodimos, A., Doulamis, N., Doulamis, A. & Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, (2018).
3. Dillon, T., Wu, C. & Chang, E. Cloud computing: issues and challenges. in *2010 24th IEEE international conference on advanced information networking and applications* 27–33 (2010).
4. Roy, K., Mukhopadhyay, S. & Mahmoodi-Meimand, H. Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits. *Proc. IEEE* **91**, 305–327 (2003).
5. Kim, N. S. *et al.* Leakage current: Moore's law meets static power. *Computer.* **36**, 68–75 (2003).
6. Markov, I. L. Limits on fundamental limits to computation. *Nature* **512**, 147–154 (2014).
7. Gollakota, S., Reynolds, M. S., Smith, J. R. & Wetherall, D. J. The emergence of RF-powered computing. *Computer.* **47**, 32–39 (2013).
8. Liu, Y. *et al.* Ambient energy harvesting nonvolatile processors: from circuit to system. in *Proceedings of the 52nd Annual Design Automation Conference* **150** (2015).
9. Ma, K. *et al*. Architecture exploration for ambient energy harvesting nonvolatile processors. in *High Performance Computer Architecture (HPCA), 2015 IEEE 21st International Symposium on* 526–537 (2015).
10. Wu, S.-Y. *et al.* A 3nm CMOS FinFlex™ Platform Technology with Enhanced Power Efficiency and Performance for Mobile SoC and High Performance Computing Applications. in *Proceedings of the 2022 IEEE International Electron Devices Meeting (IEDM)* 27.5.1–27.5.4 (2022).
11. Salahuddin, S., Ni, K. & Datta, S. The era of hyper-scaling in electronics. *Nat. Electron.* **1**, 442–450 (2018).
12. Chong, S. *et al.* Nanoelectromechanical (NEM) relays integrated with CMOS SRAM for improved stability and low leakage. in *Proceedings of the 2009 International Conference on Computer-Aided Design* 478–484 (2009).
13. Chen, C. *et al.* Nano-electro-mechanical relays for FPGA routing: Experimental demonstration and a design technique. in *2012 Design, Automation & Test in Europe Conference & Exhibition*

*(DATE)* 1361–1366 (2012).

14. Jin, C., Jang, K., Saraya, T., Hiramoto, T. & Kobayashi, M. Experimental Study on the Role of Polarization Switching in Subthreshold Characteristics of $HfO_2$-based Ferroelectric and Anti-ferroelectric FET. in *2018 IEEE International Electron Devices Meeting (IEDM)* 31.5.1-31.5.4 (2018).
15. Liang, Z. *et al.* A Novel High-Endurance FeFET Memory Device Based on $ZrO_2$ Anti-Ferroelectric and IGZO Channel. in *2021 IEEE International Electron Devices Meeting (IEDM)* 13–17 (2021).
16. Pešić, M. *et al.* How to make DRAM non-volatile? Anti-ferroelectrics: A new paradigm for universal memories. in *2016 IEEE International Electron Devices Meeting (IEDM)* 11–16 (2016).
17. Chang, S.-C. *et al.* FeRAM using Anti-ferroelectric Capacitors for High-speed and High-density Embedded Memory. in *2021 IEEE International Electron Devices Meeting (IEDM)* 32–33 (2021).
18. Pešić, M. *et al.* Built-in bias generation in anti-ferroelectric stacks: methods and device applications. *IEEE J. Electron Devices Soc.* **6**, 1019–1025 (2018).
19. Muller, J. *et al.* Ferroelectricity in simple binary $ZrO_2$ and $HfO_2$. *Nano Lett.* **12**, 4318–4323 (2012).
20. Reyes-Lillo, S. E., Garrity, K. F. & Rabe, K. M. Antiferroelectricity in thin-film $ZrO_2$ from first principles. *Phys. Rev. B* **90**, 140103 (2014).
21. Min, H. P. & Hwang, C. S. Fluorite-structure antiferroelectrics. *Reports Prog. Phys.* **82**, 124502- (2019).
22. Bhati, I., Chang, M.-T., Chishti, Z., Lu, S.-L. & Jacob, B. DRAM refresh mechanisms, penalties, and trade-offs. *IEEE Trans. Comput.* **65**, 108–121 (2015).
23. Sun, C. *et al.* First Demonstration of BEOL-Compatible Ferroelectric TCAM Featuring a-IGZO Fe-TFTs with Large Memory Window of 2.9 V, Scaled Channel Length of 40 nm, and High Endurance of $10^8$ Cycles. in *2021 Symposium on VLSI Technology* 1–2 (2021).
24. Liu, X. *et al.* Endurance properties of silicon-doped hafnium oxide ferroelectric and antiferroelectric-like thin films: A comparative study and prediction. *Acta Mater.* **154**, 190–198 (2018).

25. Pešić, M., Schroeder, U., Slesazeck, S. & Mikolajick, T. Comparative study of reliability of ferroelectric and anti-ferroelectric memories. *IEEE Trans. Device Mater. Reliab.* **18**, 154–162 (2018).
26. Xie, S. *et al.* 16.2 eDRAM-CIM: compute-in-memory design with reconfigurable embedded-dynamic-memory array realizing adaptive data converters and charge-domain computing. in *2021 IEEE International Solid-State Circuits Conference (ISSCC)* **64**, 248–250 (2021).
27. Chun, K. C., Zhang, W., Jain, P. & Kim, C. H. A 2T1C embedded DRAM macro with no boosted supplies featuring a 7T SRAM based repair and a cell storage monitor. *IEEE J. Solid-State Circuits* **47**, 2517–2526 (2012).
28. Somasekhar, D. *et al.* 2 GHz 2 Mb 2T gain cell memory macro with 128 GBytes/sec bandwidth in a 65 nm logic process technology. *IEEE J. Solid-State Circuits* **44**, 174–185 (2008).
29. Giterman, R., Shalom, A., Burg, A., Fish, A. & Teman, A. A 1-Mbit fully logic-compatible 3T gain-cell embedded DRAM in 16-nm FinFET. *IEEE Solid-State Circuits Lett.* **3**, 110–113 (2020).
30. Chun, K. C., Jain, P., Lee, J. H. & Kim, C. H. A 3T gain cell embedded DRAM utilizing preferential boosting for high density and low power on-die caches. *IEEE J. Solid-State Circuits* **46**, 1495–1505 (2011).
31. Giterman, R. *et al.* An 800-MHz Mixed-$V_T$ 4T IFGC Embedded DRAM in 28-nm CMOS Bulk Process for Approximate Storage Applications. *IEEE J. Solid-State Circuits* **53**, 2136–2148 (2018).
32. Pavan, P., Bez, R., Olivo, P. & Zanoni, E. Flash memory cells-an overview. *Proc. IEEE* **85**, 1248–1271 (1997).
33. Kawahara, T., Ito, K., Takemura, R. & Ohno, H. Spin-transfer torque RAM technology: Review and prospect. *Microelectron. Reliab.* **52**, 613–627 (2012).
34. Wong, H.-S. P. *et al.* Phase change memory. *Proc. IEEE* **98**, 2201–2227 (2010).
35. Mondal, S. Phase change materials for smart textiles--An overview. *Appl. Therm. Eng.* **28**, 1536–1550 (2008).
36. Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. *Nature* **453**, 80–83 (2008).
37. Wong, H.-S. P. *et al.* Metal--oxide RRAM. *Proc. IEEE* **100**, 1951–1970 (2012).

493 38. Scott, J. F. & de Araujo, C. A. Ferroelectric memories. *Science (80-. ).* **246**, 1400–1405 (1989).
494 39. De Araujo, C. *et al.* Fatigue-free ferroelectric capacitors with platinum electrodes. *Nature* **374**,
495 627–629 (1995).
496 40. Miller, S. L. & McWhorter, P. J. Physics of the ferroelectric nonvolatile memory field effect
497 transistor. *J. Appl. Phys.* **72**, 5999–6010 (1992).
498 41. Müller, J. *et al.* Ferroelectricity in $HfO_2$ enables nonvolatile data storage in 28 nm HKMG. in
499 *2012 Symposium on VLSI Technology (VLSIT)* 25–26 (2012).
500 42. Dünkel, S. *et al.* A FeFET based super-low-power ultra-fast embedded NVM technology for
501 22nm FDSOI and beyond. in *2017 IEEE International Electron Devices Meeting (IEDM)* 17–
502 19 (2017).
503 43. Kim, S., Zhou, J. & Lu, W. D. Crossbar RRAM arrays: Selector device requirements during
504 write operation. *IEEE Trans. Electron Devices* **61**, 2820–2826 (2014).
505 44. Grossi, A. *et al.* Fundamental variability limits of filament-based RRAM. in *2016 IEEE*
506 *International Electron Devices Meeting (IEDM)* 4–7 (2016).
507 45. Keshavarzi, A., Ni, K., Van Den Hoek, W., Datta, S. & Raychowdhury, A. Ferroelectronics for
508 edge intelligence. *IEEE Micro* **40**, 33–48 (2020).
509 46. Deng, S. *et al.* A comprehensive model for ferroelectric FET capturing the key behaviors:
510 Scalability, variation, stochasticity, and accumulation. in *2020 IEEE Symposium on VLSI*
511 *Technology* 1–2 (2020).
512 47. Xu, Y. *et al.* Improved Multibit Storage Reliability by Design of Ferroelectric Modulated
513 Antiferroelectric Memory. *IEEE Trans. Electron Devices* (2022).
514 48. Li, S. *et al.* Drisa: A dram-based reconfigurable in-situ accelerator. in *2017 50th Annual*
515 *IEEE/ACM International Symposium on Microarchitecture (MICRO)* 288–301 (2017).
516 49. Angizi, S., Sun, J., Zhang, W. & Fan, D. Aligns: A processing-in-memory accelerator for dna
517 short read alignment leveraging sot-mram. in *2019 56th ACM/IEEE Design Automation*
518 *Conference (DAC)* 1–6 (2019).
519 50. Sharma, A. A. *et al.* High speed memory operation in channel-last, back-gated ferroelectric
520 transistors. in *2020 IEEE International Electron Devices Meeting (IEDM)* 15–18 (2020).
521
522

## Method

### Device fabrication

The process flow for device fabrication is shown in Supplementary Fig. 4. The process began with the deposition of 30 nm W on the $SiO_2$/Si substrate by DC sputtering. Next, a 10 nm $Hf_{0.2}Zr_{0.8}O_2$ (HZO) AFE layer was grown by ALD at 250 ℃ following the patterning and etching process. After depositing another 20 nm W capping layer, the samples were annealed with 450 ℃ for 60 s in an $N_2$ ambient. The AFE capacitors can be fabricated by patterning and etching the top W layer. Furthermore, this layer serves as the floating gate layer for AFeFETs. Then, another 7 nm $HfO_2$ was deposited using ALD at 250 ℃, followed by an 8 nm α-IGZO channel layer deposition using radio frequency (RF) sputtering at room temperature. Device isolation was performed by wet etching of IGZO using the HCl solutions and $HfO_2$ using diluted hydrofluoric acid (HF). The samples were then annealed at 300 ℃ in the air for 30 min before the lift-off process of the Ti/Au S/D electrodes deposited by an e-beam evaporator. The 2T1AF memory cell was fabricated simultaneously using devices with an appropriate AR as the access transistors, while connecting their source terminals with the gate terminals of the AFeFETs.

### Electrical characterization

The electrical properties of the MFM AFE capacitors were measured using an aixACCT TF Analyzer 3000 with the FE module. To obtain the *P*-*V* and *I*-*V* curves, we employed the standard dynamic hysteresis measurement (DHM) with triangular pulses on 100 μm×100 μm square-shaped samples. The endurance of the capacitors was evaluated using the fatigue measurement (FM) with the same tool. For the electrical characterization of AFeFETs, we used an Agilent B1500A semiconductor parameter analyzer. Static characterization was performed using Source/Monitor Units (SMU), while the pulse signals were generated by the Semiconductor Pulse Generator Unit (SPGU) modules. Fig. 4 illustrates the schematic of the measurement setup. For the device endurance and retention measurements, the SPGU module was directly connected to the gate terminal of AFeFETs. For the memory cell demonstration, two SPGU modules were employed. During the write operation, we applied a 5 V bias to the WWL to turn on the access transistor, while the write pulses were applied to the WBL. During the standby period, the WWL was set to floating for mitigating the gate leakage current, while the WBL was maintained at $V_m$. To monitor the state

stored within the memory cell, we kept sampling the source-to-drain current of the AFeFET using the SMU with $V_{DS}$ set to 0.1 V.

## Data availability

The data that support the figures within this paper and other findings of this research are available from the corresponding authors with reasonable requests.

## Acknowledgment

This research is funded in part by the National Natural Science Foundation of China (grant nos. U21B2030, 92264204), in part by Singapore Ministry of Education (Tier 2: MOE2018-T2-2-154), and in part by Industry funding (A-0009450-02-02), and in part by Tsinghua University – Daimler Greater China Ltd. Joint Institute for Sustainable Mobility.

## Author contributions

X.L., X.G. and K.N. proposed and supervised the research. X.L. and H.Z. conceived the LVFM design with corresponding operations. H.Z. conducted the circuit-level simulations. Z.Z. performed the device fabrication and the experimental verification under the advice of X.G., K.N. and X.L. All authors contributed to the discussion and data analysis. H.Z., Z.Z., X.L., X.G. and K.N. wrote the manuscript.

## Competing interests

The authors declare no competing interests.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.pdf