Adaptive Circuit Approaches to Low-Power Multi-Level/Cell FeFET Memory

Juejian Wu, Yixin Xu, Bowen Xue, Yu Wang, Yongpan Liu, Huazhong Yang, and Xueqing Li

The Department of Electronic Engineering
Tsinghua University, Beijing 100084, China
E-mail: xueqingli@tsinghua.edu.cn

Abstract – Ferroelectric FETs (FeFETs) have emerged as a promising multi-level/cell (MLC) nonvolatile memory (NVM) candidate for low-power applications. This originates from the advantages of both efficient memory access and intrinsic device-level in-memory computing flexibilities. However, there still exist challenges for FeFET MLC NVM: (i) high power consumption in read operations due to high-gain requirement for sense amplifiers during sensing, and (ii) high latency and energy consumption in write operations with conventional recursive program-and-verify. Targeting at lower power, less latency, and higher density, this work investigates and optimizes the read and write approaches to MLC FeFET NVM design: (i) Adaptive FeFET memory State Mapping (ASM) between the FeFET drain-source current and the digital states to increase the sensing margin; (ii) Adaptive FeFET Gate Biasing (AGB) read methods that adopt the optimized FeFET gate voltage to boost the sensible dynamic range and to store more levels of states per cell; (iii) Adaptive Prediction-based Direct (APD) write methods that minimize the program-and-verify activities. Evaluations show significant latency and energy improvement. Furthermore, the number of sensible levels of states per cell is also increased with an enhanced dynamic sensing range and an enhanced sensing margin.

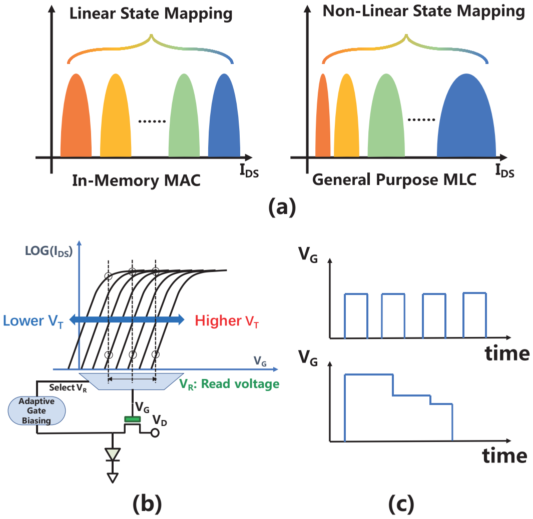

Fig. 1. Adaptive circuit approaches to FeFET MLC NVM: an overview.

I. Introduction

In modern data-intensive computing systems, memory has become a critical component that affects the overall system performance, cost, and the power consumption [1]. This is primarily because the memory and computing units are separated in the conventional von Neumann architectures and that the memory access and even data storage can be time-consuming and/or power-hungry [2]. Such a phenomenon has become the so-called bottleneck of "the memory wall" in many applications [1][3]. To mitigate this memory bottleneck, various efforts in the embedded memory design have been made, including (i) the adoption of nonvolatile memory (NVM) to eliminate the standby leakage power [4][5][6], (ii) multi-level/cell (MLC) storage to increase the data density for lower area cost [7][8], (iii) memory-centric computing solutions to reduce the memory traffic, in particular, the in-memory computing (CiM) techniques [9][10][11].

While these efforts have been tactically helpful and led to MLC NVM and "computable" NVM [9][10][11], there still exists large gaps of power and performance between the available memory solutions and the desired system metrics. These gaps are caused by both device-mechanism limitations and peripherals-related overheads. Meanwhile, a few new trends have made it urgent and promising to further explore new design and optimization space. One big trend is the emerging of new NVM devices, in particular, the ferroelectric field-effect-transistor (FeFET) [12]-[15]. FeFETs are essentially unique in bearing both a computing transistor switch and a multi-level NVM cell in one nanoscale device. Related research is still in its infancy in collectively exploiting the FeFET characteristics, including the NVM-switch integration, the read-write isolation, the DC-power-free write access, and the high on/off state ratio. Another trend from the application perspective, is that the memory design goals include both general-purpose storage and the computation capability with application-dependent accuracy requirement. Designs with such flexibility can support more scenarios and reduce the system implementation cost.

With these trends in mind, this work investigates the device-circuit co-design and proposes low-power approaches to MLC memory design using the emerging FeFETs. More specifically, as illustrated in Fig. 1, the insights and contributions include:

- Adaptive FeFET memory State Mapping (ASM) between the FeFET drain-source current state and the digital values to increase the sensing margin for applications of both general-purpose data storage and CiM. This originates from the insights of how the different FeFET NVM states interact with each other to achieve the minimum sensing overheads;
- Adaptive FeFET Gate Biasing (AGB) read, which chooses the optimum FeFET gate biasing in a read to enable more levels of states. This originates from the insights of boosting the dynamic range considering the FeFET read sensitivity under different gate biasing;

978-1-7281-4123-7/20/$31.00 © 2020 IEEE

Table I. Comparisons between FeFET NVM and Other Typical Memory Technologies [16]-[24][29][30]

Memory Technology	Cell Structure	Non-volatility	Read Time	Cell Write Time	Cell Write Energy	Endurance
SRAM	6T/8T/10T	NO	~1 ns	~1 ns	~10^{-16} J/bit	10^{16}
RRAM	1T-1R	YES	~10 ns	~10 ns	~10^{-13} J/bit	10^{8} - 10^{12}
STT-MRAM	1T-1MTJ	YES	~10 ns	~10 ns	~10^{-13} J/bit	>10^{15}
PCM	1T-1PCM	YES	~10 ns	~10 ns	~10^{-11} J/bit	10^{8} - 10^{15}
FeRAM	1T-1C	YES	~10 ns	~10 ns	~10^{-14} J/bit	10^{10} - 10^{15}
FeFET	1T/2T	YES	~10 ns	~10 ns	~10^{-14} J/bit	10^{4} - 10^{12}

- Adaptive Prediction-based Direct (APD) write, which prevents reset and minimizes the program-and-verify operations to reduce the power consumption and latency.

In the rest of this paper, section II reviews MLC NVM technologies and discusses the FeFET MLC NVM design exploration space. Section III presents the FeFET MLC NVM read techniques, including ASM and AGB. Section IV introduces the proposed APD write operations. Section V concludes this paper._

II. FeFET MLC NVM: Background and Design Space

This section reviews NVM and FeFET device basics, investigates the FeFET-based MLC NVM design space, and discusses the design challenges and opportunities illustrated in Fig. 1. The target applications include the general-purpose nonvolatile data storage and the emerging domain-specific CiM accelerators.

A. Why FeFET: NVM and FeFET Basics

Fig. 2 illustrates the FeFET concept of device structure and general I_{DS}-V_G curves. FeFETs are essentially a MOSFET with an embedded ferroelectric layer at the gate stack [15][18],[25]-[27]. The polarization state of the ferroelectric domains in the integrated ferroelectric layer could be modulated conveniently by the external gate voltage. By doing so, an n-type FeFET may exhibit a small and even negative V_T with positive polarization, and a large V_T with negative polarization. Accordingly, an FeFET may achieve a tunable and stable threshold voltage V_T, which could be used as a nonvolatile method to store multi-level data.

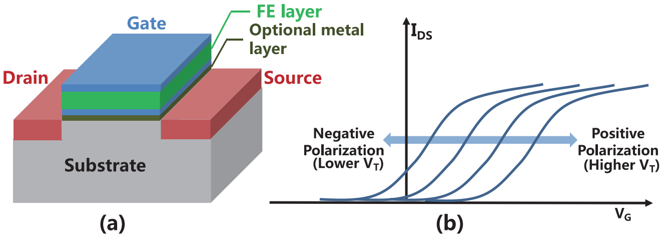
Gate
FE layer
Optional metal layer
Drain
Source
Substrate
(a)
I_{DS}
Negative Polarization (Lower V_T)
Positive Polarization (Higher V_T)
V_G
(b)

Fig. 2. FeFET basics. (a) Device structure; (b) General I_{DS}-V_G characteristics.

The FeFET device concept is not new and dates back to decades ago. Yet it becomes fascinating when researchers found out that the doped hafnium dioxide (HfO_2), a mature and widely used material in the CMOS process, could exhibit the desired ferroelectricity [28]. This indicates that FeFETs could be fabricated in a way that is CMOS-compatible, scalable and low-cost, as demonstrated with recent FeFET experimental reports with ~10nm FinFET CMOS processes [13][14].

Table I summarizes the comparison between FeFET and other memory technologies, including CMOS SRAM, resistive random-access memory (RRAM), magnetic random-access memory (MRAM), phase-change memory (PCM), and the ferroelectric (capacitor) random-access memory (FeRAM) [29][30]. In addition to the excellence of process compatibility and scaling capability, FeFETs are particularly intriguing in a few aspects among these memory technologies:

- High circuit distinguishability of different states with a typical on/off ratio above 10^5. This is intrinsically enabled by the high transconductance gain between the drain-source current I_{DS} and the internal gate voltage. This feature does not exist in other NVM technologies and makes FeFET highly promising with MLC applications, as to be further explored in this work.
- Energy-efficient DC-power-free write operations. This is enabled by the fact that the load of the FeFET NVM write operation is capacitive and does not consume DC currents. In comparison, the resistive load in other resistive NVM technologies consumes DC power and deteriorates the energy efficiency with the presence of device write-speed variations.

In addition to the abovementioned outstanding features, FeFET also exhibit moderate endurance, write speed, and operation voltage adaptability. All these features put together, FeFET has become a promising candidate for both NVM storage and computing applications [31]-[35].

The motivation of adopting MLC NVM instead of single-level/cell (SLC) NVM is straightforward: higher density for general-purpose storage and higher CiM computing accuracy. Nevertheless, generally, MLC designs still face a big challenge of device variations and yield. This necessitates the device-circuit co-efforts to ensure the required accuracy, as to be further investigated subsequently.

B. FeFET MLC NVM and CiM Circuits Architectures

MLC techniques increase the data density to meet the high demand of storage and computing in memories. MLC has already been used for NAND, RRAM, etc. [8][36]-[39]. It has also been used with FeFET for neuromorphic applications [40]-[43].

Fig. 3 shows the circuit architecture for general MLC NVM storage and CiM acceleration applications. In addition to the main memory cell array, it includes the read and write peripheral interface, in which the critical task is to optimize the balance of density, power, speed, etc. For faster sensing, sense amplifiers (SA) are usually needed to read out different states. A smaller safe margin between memory states requires a higher SA gain. Practically, an MLC read could use one-step parallel sensing with multi-reference amplifiers or multi-step sequential sensing with fewer-reference amplifiers. Therefore, there is a tradeoff between the read latency and the peripheral circuit overheads [36]. To optimize the MLC read and write, it is essential that both the device and peripheral circuits support the required accuracy with balanced energy consumption, latency overheads, and yield. As to be further revealed, FeFET-based MLC NVM design exhibits both challenges and opportunities.

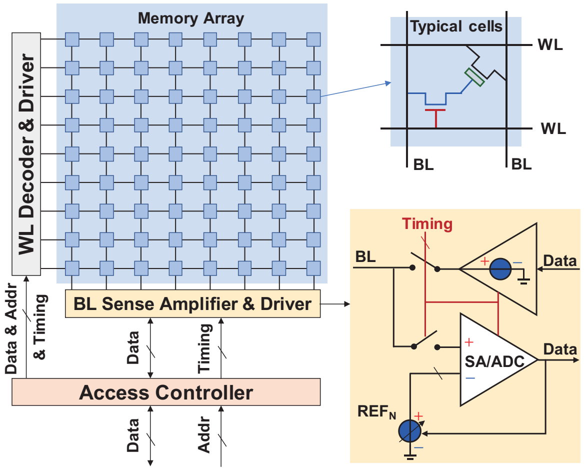

Fig. 3. MLC circuit architectures for NVM storage and CiM acceleration.

C. Opportunities and Challenges

Before enjoying the benefits of higher memory storage density through MLC implementations, there are two major challenges to overcome. For each challenge, we also show opportunities with the proposed FeFET circuit approaches.

The first challenge is the high power consumption due to the need of using high-gain SAs to distinguish MLC states in the read operations. This is simply because of smaller safe margins between adjacent states in MLC implementations compared with those in SLC. In this work, we reveal out two opportunities. The first one is optimizing the state mapping between the device physical storage parameter or the device access behavior (e.g. the polarization state or I_{DS}) and the digital values of states (e.g. '01'/'10'). We propose to use Adaptive FeFET memory State Mapping (ASM) to maximize the safe margin between different states. The adaptivity here is incorporated in the custom mapping for general-purpose NVM storage and CiM applications. The second opportunity that we find out in this work, is the exploit of extra device tuning knobs that do not exist in other NVM devices, in particular, the FeFET gate voltage biasing during read, to enhance the sensible dynamic range. We propose Adaptive FeFET Gate Biasing (AGB) for this purpose, originating from the insights that the FeFET I_{DS} is could be modulated with different gate biasing.

The second challenge lies in lowing the energy and time consumption for MLC NVM write operations. As the program-and-verify iterations usually are needed to achieve the final accuracy, it is helpful to reduce the number of iterations for faster settling down and lower energy consumption. In this regard, we propose Adaptive Prediction-based Direct (APD) write techniques to use prior knowledge of the FeFET program behavior to set the write pulse height and duration to achieve inter-data direct programming. This prevents reset operations and minimizes the program-and-verify activities for lower power consumption and latency.

III. FeFET MLC NVM: Sensing Approaches

This section presents approaches of adaptive state mapping (ASM) and adaptive gate biasing (AGB) to sensing the FeFET MLC NVM, so as to improve the dynamic sensing range and the sensing margin between memory states. The theory, implementation and quantitative evaluation of the approaches based on circuit simulations are included.

A. Adaptive State Mapping

As discussed above, the safe margin between the MLC memory states is much smaller than that of the SLC memory, making it more difficult to read out the memory data accurately. Practically, it is observed that the smallest sensing margin determines the sensing costs. Therefore, it is straightforward to optimize the locality of the memory states within the entire dynamic sensing range, such that the sensing safe margin between all adjacent memory states is the same. This strategy ensures that the minimum sensing safe margin is the highest, i.e. producing the widest memory window and leading to the minimum required SA gain and power.

However, applications my exert specific memory data locality. For example, in multiply-and-accumulation (MAC) operations, it is naturally to assign the memory state to be a linear function of the sensed output current (or voltage), which ensures the linear accumulation functionality. Therefore, we propose to use an adaptive mapping strategy between the stored bits and the sensed output, as illustrated in Fig. 4.

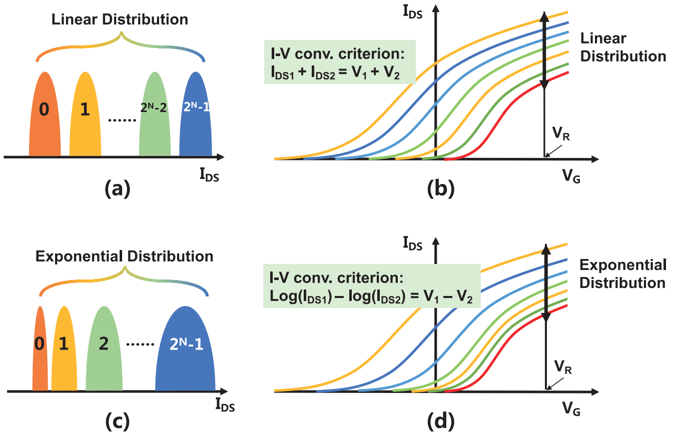

Fig. 4. Adaptive State Mapping for NVM storage and CiM acceleration. (a-b) Linear state mapping for CiM applications. (c-d) Exponential state mapping for general-purpose MLC NVM storage applications.

The "adaptivity" refers to the automatic state mapping for different applications. For computing-in-memory (CiM) applications, e.g. MAC, we use the drain source current I_{DS} to linearly represent the memory state, so that the summation or subtraction of currents can be linearly mapped to the summation or subtraction of the memory data. For general purpose memory storage, because the sensing interface is usually built with a diode-based logarithm current-voltage (I-V) converter, as illustrated in Fig. 4, we map the memory states in an exponential way such that the sensed converted voltages (the logarithm of I_{DS}) of adjacent memory states have a constant absolute difference.

B. Adaptive FeFET Gate Biasing (AGB) Read Methods

As discussed above, there is a strong motivation to achieve higher-density NVM through increasing the number of levels per cell with affordable overheads, if possible. However, in existing FeFET-based MLC NVM designs, it could be costly for SA to distinguish too many states with a subtle difference of sensed currents when reading an MLC NVM. Therefore, it will be of great significance to explore a new dimension of read methods to increase the sensible number of states without putting too many overheads to SA.

When revisiting the prior methods of reading an FeFET MLC NVM cell, it is found out that they use a fixed FeFET gate voltage biasing for reading all memory states. While conventional two-terminal NVM devices do not have opportunity to modulate a third terminal, FeFET, on the contrary, is unique in having the extra gate terminal and making the read operation more elegant. As a matter of fact, tuning the FeFET gate voltage is a convenient approach to increasing the I_{DS} dynamic range and increasing the number of sensible states. This is the finding that leads to the proposed adaptive gate biasing (AGB) for read operations.

The theory behind ABG could be illustrated in Fig. 5. If only one fixed gate biasing voltage is used, either V1 or V2, the sensible dynamic range is limited. For example, when the biasing voltage is V1, the sensible dynamic range is determined by α, which is the ratio of I_{DS} @ V_{DS} = V1 to the minimum sensible current I_0, e.g. 1μA (I_0 is limited by the SA gain). In other words, the current states below I_0 could not be sensed. Similarly, if only V2 is adopted as the gate biasing voltage, the originally insensible states with V1 as the gate bias now becomes sensible, but the states originally within the sensible dynamic range of α with V1 as the gate bias now shrinks to α', a much smaller value than α.

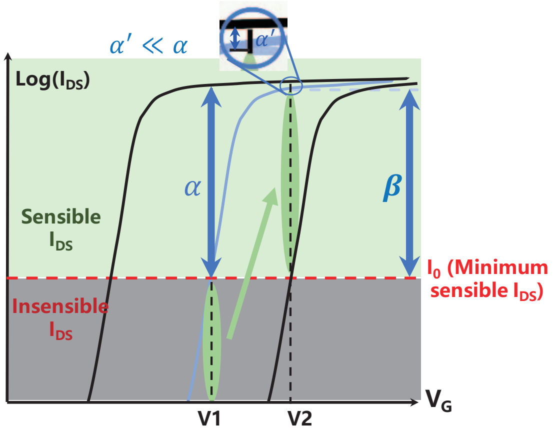

Fig. 5. The scheme of proposed multi-gate-voltage read method.

The proposed ABG can combine the use of both V1 and V2: for the states between I_0 and α*I_0 at V1, V1 is adopted as the gate bias; for the states below I_0 at V1, V2 is adopted as the gate bias. Therefore, the overall dynamic sensing range is α*β, which could be much large than the dynamic sensing range of α at the fixed V1 gate bias and the dynamic sensing range of α'*β at the fixed V2 gate bias.

There are a few restrictions for AGB when dynamically tuning the gate bias for read. First, the gate bias should be sufficiently low to prevent disturbing the stored polarization states. Second, the number of tunable gate bias should not be too many to prevent high complexity.

Fig. 6 illustrates the practical read procedure with ABG, using gate bias voltages V1 and V2 (V1<V2) as an example. Both V1 and V2 can be used first as the read voltage. For example, if V1 is adopted first and the sensing current is above the minimum sensing current I_0, the state can be read out directly; otherwise, increase the read voltage to V2 to read out the state. If V2 is adopted first and the sensing current is below β*I_0, the state can be read out directly; otherwise, lower the read voltage to V1 to read out the state.

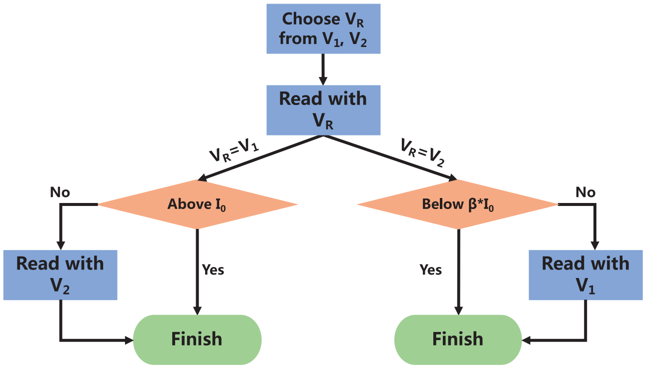

Fig. 6. Proposed AGB read procedure (two read bias voltage example).

TABLE II. FeFET Model Parameters

Parameter	Nominal Value	Description
Ps	2.5e-5 C/cm^2	Saturation polarization
Pr	2.0e-5 C/cm^2	Remnant polarization
Ec	9.0e5 V/cm	Coercive field
AR	1	Area ratio: A_{FE}/A_{MOS}
Epison_FE	20	FE linear dielectric constant
Tauo	5.07e-9 s	Switching time constant
Vo	5.86 V	Switching time voltage acceleration constant
m	1.34	Switching time exponent

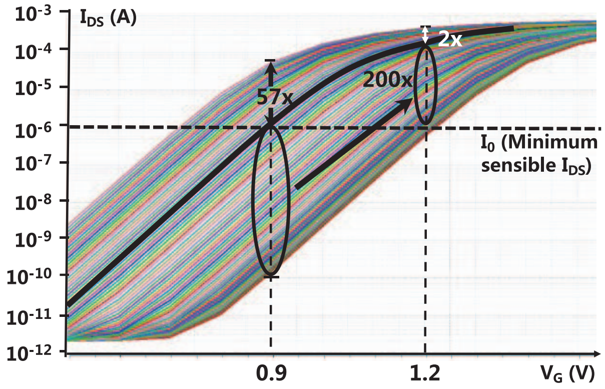

Fig. 7. I_{DS} vs V_G for different memory states in one memory cell (read mode).

In order to quantitatively benchmark the effectiveness of the proposed ABG method, the experimentally calibrated FeFET device SPICE model for simulation from [45] is adopted for circuit simulations. The parameter settings are listed in Table II. In the simulations, the write time is limited within the range of 1ns to 3.9μs. We choose 0.9V and 1.2V as two example read voltages, i.e. V1 = 0.9V and V2 = 1.2V. The minimum sensing current of sense amplifiers, I_0, is set to 1.0µA.

Fig. 7 shows the simulation results. In the case of V_G = 0.9V, I_{DS} ranges from a very small value (<nA) to an end of ~57µA, showing a sensible dynamic range of 57 and the number of sensible states is 42 assuming 10% adjacent sensed current difference. In the case of V_G = 1.2V, I_{DS} ranges from less than 1.0µA to an end of 400µA, showing a sensible dynamic range of 400 and the number of sensible states is 62. With ABG, the total sensible dynamic range increases to α * β = 57 * 200 = 11,400, and the total number of sensible states increases to 98 (compared with the baseline of 62).

IV. FeFET MLC NVM: Write Approaches

In this section, write operation optimizations for FeFET-based MLC will be introduced. The Adaptive Prediction-based Direct (APD) write method is proposed to reduce the energy consumption and operation latency of MLC write. The improvement is also quantitatively evaluated in this section.

A. Adaptive Prediction-Based Direct (APD) Write Method

The method of program-and-verify (P&V) has already been applied to MLC memories to ensure that the cell has been written to the exact range of state. The write operation ends when the distance between the current memory state and the target state is close enough. Most of the existing P&V methods have adopted fixed pulse amplitude and duration [43]. The proposed APD method not only verifies the latest memory state, but also calculate the distance to the desired state and shapes the subsequent write pulse, if needed, in both amplitude and duration, to achieve the minimized number of P&V activities.

For FeFET-based NVM, the write energy is ultra-low due to the zero DC current within the array. Charging the parasitic capacitance along the interconnections consumes most energy during a write operation [44]. The write operation of an MLC memory may consist of a series of write pulse. Meanwhile, the capacitor needs be charged and discharged for several times. Therefore, reducing the pulse number can significantly lower the energy consumption of the write operation.

The proposed improves the flexibility of the write pulse to reduce the pulse number. Compared to the existing methods, this process-aware write supports adaptive pulse amplitude and duration, depending on the latest verifying result. For instance, the pulse amplitude and duration can be stored in a look-up table (LUT) based on the prior knowledge of the device characteristics. If the distance to the target state is large, a write pulse with larger amplitude and duration will be applied. Otherwise, the amplitude and duration of the pulse should be small to fine tune the memory state, as shown in Fig. 8. Therefore, the number of pulses could be significantly reduced.

The procedure of the proposed APD write method is shown in Fig. 9. Each cycle starts with a verify operation to sense the current memory state. Then the control module judges whether the distance to the target is close enough. If yes, the write operation ends. Otherwise, the control module will decide the amplitude and duration of the next pulse based on the distance between the current state and the target state. Then the pulse generator module will output the pulse to implement the write operation, which is the last step of the loop.

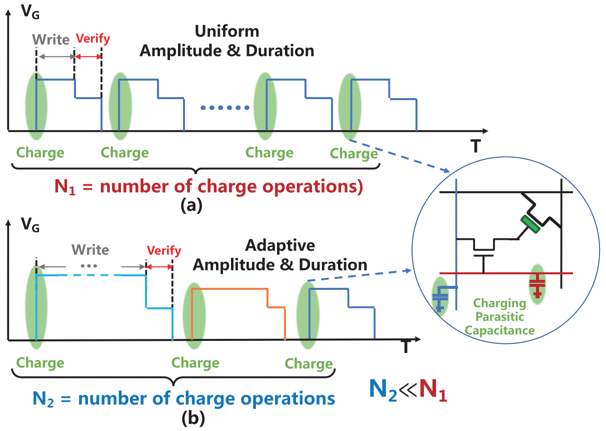

Fig. 8. Write MLC FeFET NVM. (a) Traditional approach; (b) Proposed APD approach with adaptive amplitude and duration.

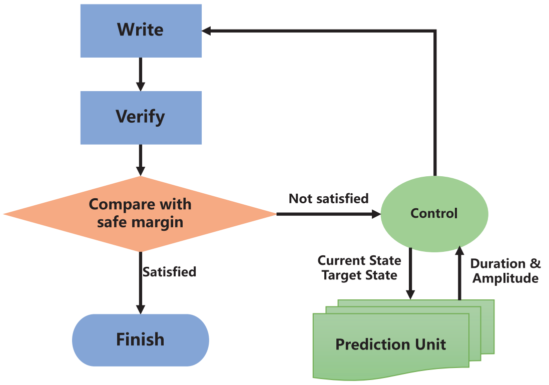

Fig. 9. The procedure of the APD program-and-verify.

B. Benchmark

The energy and latency of the proposed write strategy have been simulated based on the FeFET SPICE model from [45]. The parameters are listed in Table II. Both the proposed APD write approach and the baseline approach using a fixed write pulse amplitude and duration are investigated.

As mentioned in section III, the memory states are defined at V_G = 0.9V and V_G = 1.2V as an example. In the simulation, four typical states are selected from the 98 states of 6.6-bit/cell design for a close look. These four states are shown in Table III. For the baseline simulation, the write duration is set to 3ns to ensure that one state can switch to the closest state through only one write pulse. The verify duration of both baseline and proposed methods is set to 1ns.

Table III. FeFET State Definition

State	Definition
A	I_{DS}=32 µ A@V_G=1.2V
B	I_{DS}=38 µ A@V_G=1.2V
C	I_{DS}=165 µ A@V_G=1.2V
D	I_{DS}=15 µ A@V_G=0.9V

In transient simulations, the cell is initialized to one of the four states. The capacitor of each word line or bit line in a large array is set to 50fF. During a write, both latency and energy are monitored. The energy consumption has considered write, verify, and the parasitic capacitor charging. As the energy of each write pulse depends on the pulse amplitude and the load capacitor, the charging energy can be derived by the charging amplitude and iterations.

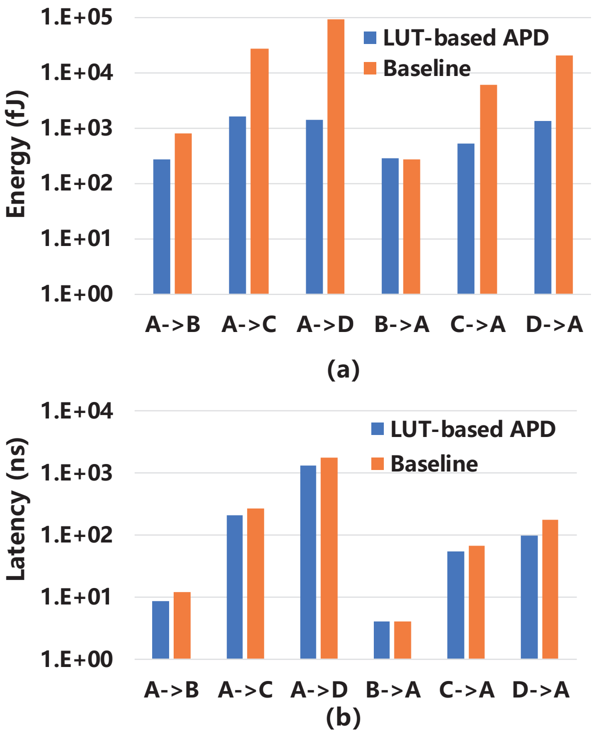

Fig. 10. Proposed APD write approach evaluation: (a) Energy; (b) Latency.

The energy and latency of the two methods during six write procedures between the four states are shown in Fig. 10. Benefiting from the adaptive pulse amplitude and duration, the number of pulses is significantly reduced. The average energy saving is 91%. Overwriting the state A with the state B, the energy consumption of the proposed method is 270fJ, while the baseline method consumes 820fJ. The difference becomes even larger when the baseline write operation requires more iterations. For instance, in a write from the state A to the state D, the baseline write energy reaches 91pJ due to 435 program-and-verify iterations. The proposed method consumes only 1.42pJ because only 4 iterations are needed.

Moreover, the latency of the proposed design also shows improvement because of fewer write pulses and verify operations. The average latency reduction is 25%. In the scenarios of a small distance between the current memory state to the target state, the proposed method and the baseline method may have similar or the same latency, as the APD write approach may adopt the same pulse as the baseline. An example is the B->A write operation. In the scenarios of a large distance transition, the improvement could be significant, for example, 45% latency reduction for the D->A write.

V. Conclusions

This paper has investigated the multi-level/cell NVM design space using FeFETs. The design challenges and opportunities have been discussed in detail. In order to achieve higher density, lower the energy consumption and access latency, this paper has proposed a few effective approaches, including the adaptive state mapping (ASM) and adaptive gate biasing (AGB) read, and the adaptive prediction-based direct (APD) write. The theory, implementation and evaluation of these approaches have been presented in detail, showing promising memory and computing paradigms enabled by the FeFET-based MLC technologies.

Acknowledgments

This work is supported in part by NSFC under grants 61874066 and 61720106013, and in part by BNRist and The Beijing Information Center for Future Chips. Juejian Wu and Yixin Xu contributed equally to this work. The authors would like to thank Prof. Suman Datta and Prof. Kai Ni for the device model support.

References

[1] W. A. Wulf and S. A. McKee, “Hitting the Memory Wall: Implications of the obvious,” *ACM SIGARCH computer architecture news*, vol. 23, no. 1, pp. 20–24, 1995.

[2] N. R. Mahapatra and B. Venkatrao, “The Processor-Memory Bottleneck: Problems and solutions,” *XRDS: Crossroads, The ACM Magazine for Students*, vol. 5, no. 3es, p. 2, 1999.

[3] S. A. McKee et al., “Reflections on the memory wall.,” in *Conf. Computing Frontiers*, 2004, p. 162.

[4] Y. Huai et al., “Spin-Transfer Torque MRAM (STT-MRAM): Challenges and prospects,” *AAPPS bulletin*, vol. 18, no. 6, pp. 33– 40, 2008.

[5] H.-S. P. Wong, S. Raoux, et al., “Phase Change Memory,” *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2201–2227, 2010.

[6] H. Akinaga and H. Shima, “Resistive Random Access Memory (ReRAM) based on Metal Oxides,” *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2237–2251, 2010.

[7] M.-F. Chang, C.-C. Lin, et al., “17.5 a 3t1r Nonvolatile TCAM Using MLC ReRAM with sub-1ns Search Time,” in *2015 IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, IEEE, 2015, pp. 1–3.

[8] J. Cheon, I. Lee, et al., “Non-resistance Metric Based Read Scheme for Multi-Level PCRAM in 25nm Technology,” in *2015 IEEE Custom Integrated Circuits Conference (CICC), IEEE,* 2015, pp. 1–4.

[9] L. Xia, P. Gu, et al., “Technological Exploration of RRAM Crossbar Array for Matrix-Vector Multiplication,” *Journal of Computer Science and Technology*, vol. 31, no. 1, pp. 3–19, 2016.

[10] S. H. Jo, T. Kumar, et al., “3d-Stackable Crossbar Resistive Memory Based on Field Assisted Superlinear Threshold (FAST) Selector,” in *2014 IEEE International Electron Devices Meeting, IEEE,* 2014, pp. 6–7.

[11] M. Hu, H. Li, et al., “Memristor Crossbar-based Neuromorphic Computing System: A case study,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 10, pp. 1864–1878, 2014.

[12] A. I. Khan, C. W. Yeung, et al., “Ferroelectric Negative capacitance MOSFET: Capacitance tuning & antiferroelectric operation,” in *2011 International Electron Devices Meeting, IEEE,* 2011, pp. 11–3.

[13] Z. Krivokapic, U. Rana, et al., “14nm Ferroelectric FinFET Technology with Steep Subthreshold Slope for Ultra Low Power Applications,” in *2017 IEEE International Electron Devices Meeting (IEDM), IEEE,* 2017, pp. 15–1.

[14] A. Sharma and K. Roy, “1t Non-Volatile Memory Design Using Sub-10nm Ferroelectric FETs,” *IEEE Electron Device Letters,* vol. 39, no. 3, pp. 359–362, 2018.

[15] S. Dünkel, M. Trentzsch, et al., “A Fefet Based Super-Lowpower Ultra-Fast Embedded NVM Technology For 22nm FDSOI and Beyond,” in 2*017 IEEE International Electron Devices Meeting (IEDM), IEEE,* 2017, pp. 19–7.

[16] K. Chatterjee, et al., “Self-Aligned, Gate Last, FDSOI, Ferroelectric Gate Memory Device with 5.5-nm Hf0.8Zr0.2O2, High Endurance and Breakdown Recovery”, *IEEE Electron Device Lett.* vol. 38, no. 10, pp. 1379–1382, 2017.

[17] S. Slesazeck, U. Schroeder and T. Mikolajick, “Embedding Hafnium Oxide Based FeFETs in the Memory Landscape,” in *2018 International Conference on IC Design & Technology (ICICDT),* Otranto, 2018, pp. 121-124.

[18] M. Trentzsch, S. Flachowsky, et al., “A 28nm HKMG Super Low Power Embedded NVM Technology Based on Ferroelectric FETs,” in *2016 IEEE International Electron Devices Meeting (IEDM), IEEE,* 2016, pp. 11–5.

[19] J. Muller, T. S. Boscke, U. Schroder, R. Hoffmann, T. Mikolajick and L. Frey, "Nanosecond Polarization Switching and Long Retention in a Novel MFIS-FET Based on Ferroelectric HfO2," in *IEEE Electron Device Letters*, vol. 33, no. 2, pp. 185-187, Feb. 2012.

[20] S. Mueller, S. R. Summerfelt, J. Muller, U. Schroeder and T. Mikolajick, "Ten-Nanometer Ferroelectric HfO2 Films for Next-Generation FRAM Capacitors," in *IEEE Electron Device Letters*, vol. 33, no. 9, pp. 1300-1302, Sept. 2012.

[21] B. Zeng et al., "Compatibility of HfN Metal Gate Electrodes With Hf0.5Zr0.5O2Ferroelectric Thin Films for Ferroelectric Field-Effect Transistors," in *IEEE Electron Device Letters*, vol. 39, no. 10, pp. 1508-1511, Oct. 2018.

[22] F. B. Yahya, M. M. Mansour, J. Tschanz and M. M. Khellah, “Designing Low-VTh STT-RAM for Write Energy Reduction in Scaled Technologies,” in *Sixteenth International Symposium on Quality Electronic Design*, Santa Clara, CA, 2015, pp. 5-9.

[23] W. Khwa et al., “7.3 A Resistance-Drift Compensation Scheme to Reduce MLC PCM Raw BER by Over 100× for Storage-Class Memory Applications,” *2016 IEEE International Solid-State Circuits Conference (ISSCC),* San Francisco, CA, 2016, pp. 134-135.

[24] W. Kim et al., “ALD-based Confined PCM with a Metallic Liner toward Unlimited Endurance,” *2016 IEEE International Electron Devices Meeting (IEDM),* San Francisco, CA, 2016, pp. 4.2.1-4.2.4.

[25] K. Ni, X. Li, et al., “Write Disturb in Ferroelectric FETs and its Implication for 1T-FeFET and Memory Arrays,” *IEEE Electron Device Letters*, vol. 39, no. 11, pp. 1656–1659, 2018.

[26] O. Auciello, C. A. P. de Araujo, et al., “Review of the Science and Technology for Low- and High-Density Nonvolatile Ferroelectric Memories,” in *Emerging Non-Volatile Memories,* Springer, 2014, pp. 3–35.

[27] S.-Y. Wu, “A New Ferroelectric Memory Device, Metal-Ferroelectric-Semiconductor Transistor,” *IEEE Transactions on Electron Devices,* vol. 21, no. 8, pp. 499–504, 1974.

[28] J. Müller, E. Yurchuk, et al., “Ferroelectricity in HFO2 Enables Nonvolatile Data Storage in 28 nm HKMG,” in *2012 Symposium on VLSI Technology (VLSIT)*, IEEE, 2012, pp. 25–26.

[29] A. Chen, “A Review of Emerging Non-volatile Memory (NVM) Technologies and Applications,” *Solid-State Electronics,* vol. 125, pp. 25–38, 2016.

[30] S. Salahuddin, K. Ni, et al., “The Era of Hyper-Scaling in Electronics,” *Nature Electronics*, vol. 1, no. 8, pp. 442–450, 2018.

[31] S. George, K. Ma, et al, "Nonvolatile Memory Design Based on Ferroelectric FETs," in *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, Austin, TX, 2016, pp. 1-6.

[32] S. George et al., “Symmetric 2-D-Memory Access to Multidimensional Data,” *IEEE Transactions on Very Large Scale Integration Systems*, vol. 26, no. 6, pp. 1040–1050, 2018.

[33] X. Li et al, “Enabling Energy-Efficient Nonvolatile Computing with Negative Capacitance FET,” *IEEE Transactions on Electron Devices,* vol. 64, no. 8, pp. 3452- 3458, August 2017.

[34] X. Li and L. Lai, “Nonvolatile Memory and Computing Using Emerging Ferroelectric Transistors,” in *2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2018, pp. 750–755.

[35] X. Li, J. Wu, et al., “Design of 2T/cell and 3T/cell Nonvolatile Memories with Emerging Ferroelectric FETs,” *IEEE Design & Test,* vol. 36, no. 3, pp. 39–45, 2019.

[36] C. Xu, D. Niu, et al., “Understanding the Trade-offs in Multi-level Cell ReRAM Memory Design,” in *2013 50th ACM/EDAC/IEEE Design Automation Conference (DAC)*, IEEE, 2013, pp. 1–6.

[37] M.-F. Chang, C.-C. Lin, et al., “17.5 a 3T1R Nonvolatile TCAM Using MLC ReRAM with sub-1ns Search Time,” in *2015 IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers, IEEE,* 2015, pp. 1–3.

[38] M. A. Qureshi, H. Kim, et al., “A Restore-Free Mode for MLC STT-RAM Caches,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 27, no. 6, pp. 1465–1469, 2019.

[39] H. Luo, L. Shi, et al., “Energy, Latency, and Lifetime Improvements in MLC NVM with Enhanced WOM Code,” in *2018 23rd Asia and South Pacific Design Automation Conference (ASPDAC), IEEE,* 2018, pp. 554–559.

[40] M. Jerry, P.-Y. Chen, et al., “Ferroelectric FET analog Synapse for Acceleration of Deep Neural Network Training,” in *2017 IEEE International Electron Devices Meeting (IEDM)*, pp. 6–2.

[41] I. Yoon, M. Chang, et al., “A FerroFET Based In-Memory Processor for Solving Distributed and Iterative Optimizations via Least-Squares Method,” *IEEE Journal on Exploratory SolidState Computational Devices and Circuits,* 2019.

[42] X. Yin, K. Ni, et al., “An Ultra-Dense 2FeFET TCAM Design Based on a Multi-Domain FeFET Model,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2018.

[43] X. Sun, P. Wang, et al., “Exploiting Hybrid Precision for Training and Inference: A 2T-FeFET Based Analog Synaptic Weight Cell,” in *2018 IEEE International Electron Devices Meeting (IEDM),* 2018, pp. 3–1.

[44] J. Wu, H. Zhong, et al., “A 3T/cell Practical Embedded Nonvolatile Memory Supporting Symmetric Read and Write Access Based on Ferroelectric FETs,” in *Proceedings of the 56th Annual Design Automation Conference 2019, ACM*, 2019, p. 82.

[45] K. Ni, J. Smith, et al., “Soc Logic Compatible Multi-bit FemFET Weight Cell for Neuromorphic Applications,” in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018, pp. 13–2.