Received 25 November 2019; revised 30 December 2019; accepted 5 January 2020. Date of publication 8 January 2020; date of current version 22 January 2020.
The review of this article was arranged by Editor M. Liu.

Digital Object Identifier 10.1109/JEDS.2020.2964820

An Enhanced Floating Gate Memory for the Online Training of Analog Neural Networks

LURONG GAN, CHEN WANG, LIN CHEN, HAO ZHU, QINGQING SUN, AND DAVID WEI ZHANG

State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai 200433, China

CORRESPONDING AUTHOR: L. CHEN (e-mail: linchen@fudan.edu.cn)

This work was supported in part by NSFC under Grant 61704030 and Grant 61522404, in part by the Shanghai Rising-Star Program under Grant 19QA1400600, in part by the Program of Shanghai Subject Chief Scientist under Grant 18XD1402800, and in part by the Support Plans for the Youth Top-Notch Talents of China.

ABSTRACT Floating gate (FG) memory has long erasing time, which limits its application as an electronic synapse in online training. This paper proposes a novel enhanced floating gate memory (EFM) by TCAD simulation. Here, three other structures are simulated just for comparison. The simulation results show that the erasing speed is about 34ns while the other three need the time over 1.8ms, which makes the operation speed of long-term potentiation (LTP) more symmetrical to long-term depression (LTD). In addition, both LTP and LTD are approximately linear in the simulation results. The speed, linearity, and symmetry of weight update are the keys to online training of analog neural networks. These excellent performances indicated a potential application of EFM in analog neuro-inspired computing.

INDEX TERMS Neural network, FG memory, U-shaped channel, erasing speed, online training.

I. INTRODUCTION

Recent years have seen increased research attention being directed towards artificial neural networks, but there are significant challenges for on-chip memory capacity, off-chip memory access, and online learning capability [1]. In a large-scale neural network, when a large amount of data needed to compute in the training or testing process, accessing the off-chip memory can cause substantial energy consumption and latency. Therefore, high integration density on-chip memory is needed which means the synaptic memory cell should be well scaling down and/or store a multi-bit thereby reducing or eliminating the off-chip memory. Recently, industrial grade 180 nm [2], [3] and 55 nm [4] NOR flash memories have been designed as electronic synapses. The FG memory element is appropriate as adjustable conductances in a pseudocrossbar fashion, and the accuracy is better than 1% [4]. Thereby the memory blocks can be modified accurately independently of each device. It can go back to 1980s that the nonvolatile FG memory is widely used as a synaptic weight storage cell in analog neural networks [5]–[7]. Recently, a prototype mixed-signal, 28×28-binary-input, 10-ouput, 3-layer neuromorphic network based on embedded nonvolatile floating-gate cell arrays redesigned from a commercial 180-nm NOR flash memory have designed, fabricated, and tested for MNIST image classification [8]. To achieve similar fidelity in the same task, the time delay and energy dissipation (per one pattern classification) were at least three orders of magnitude better than the 28-nm IBM TrueNorth chip, which were below to 1μs and 20nJ respectively [9]. But the high erasing voltage and the long operation time (100μs∼1ms) restrict the online training capability [1].

We report a novel FG memory structure with Sentaurus TCAD tools. Simulation results show that EFM can achieve nanosecond erasing speed, and the more symmetric LTP/LTD properties with a lower operation voltage can be implemented. So, it can be better used in online training of analog neural networks. Additionally, the U-shaped channel of EFM also reduces the short-channel effects [10]–[12].

II. DEVICE STRUCTURE AND ELECTRICAL CHARACTERISTIC

Fig. 1(a) shows the configuration of EFM. The left region of the *p*-doped FG downwards into the substrate. The channel changes from horizontal to partly U-shaped along the edge of the FG between source (S) and drain (D). There is an *n*-doped well controlled by two parts—the downward extending control gate (CG) and FG. HfO_2 with a thickness

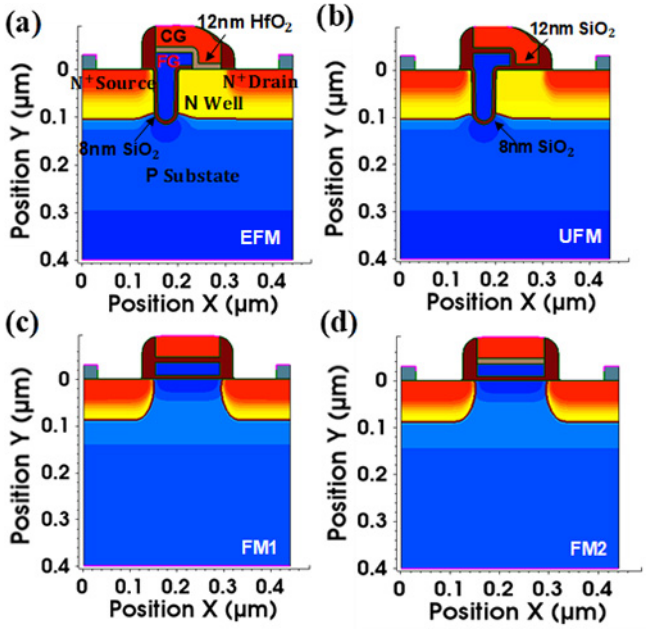

FIGURE 1. **The devices structures of (a) EFM, (b) UFM, (c) FM1 and (d) FM2.**

of 12nm is used as a medium layer under CG while 8nm SiO_2 is used as a medium layer under FG.

A U-shaped floating gate memory (UFM), which replaces the HfO_2 layer under CG in EFM with SiO_2 (Fig. 1(b)), an FG memory with traditional straightened FG (FM1) in Fig. 1(c) and another memory which replacing the SiO_2 layer under the CG in FM1 with HfO_2 (FM2) in Fig. 1(d) are just used as contrasts to EFM under the condition that other dimensions and doping concentration are the same.

Fig. 2 shows the process flow of EFM. First, the light n^--doped region is formed by the implantation of arsenic ions. After that, a U-typed channel is formed by reactive ion etching and the threshold voltage adjustment injection is then completed by using BF_2. Next, an 8-nm FG oxide layer is grown and the first boron-doped poly-Si layer is then deposited. Then a chemico-mechanical polishing (CMP) process is used to determine the FG height and smooth the FG surface. Next, the FG pattern is formed by photolithography and anisotropic etching process, by which the extra poly-Si and SiO_2 on both sides are etched. Subsequently, a 12-nm HfO_2 is grown as the inter-dielectric between FG and CG. Then, an n^+-doped poly-Si is deposited and the CG pattern is then formed by photolithography. After the spacer processes, the n^+-doped source and drain regions are formed by self-aligned arsenic ions implantation. Finally, the formation of S/D electrode is carried out. In order to monitor the changes of FG potential and charge, we define a virtual contact at FG. Due to the compatibility with CMOS processing and the maturity of device manufacturing technology, EFM cells can be nicely embedded into CMOS logic process. In the whole process of device preparation, four masks are used to define U-shaped channel, FG, CG and S/D contact respectively. The reference values of doping concentration distribution are as follows: the doping of substrate is p type and the concentration is $2.5 \times 10^{17} cm^{-3}$. The contact area between source/drain and electrode is n type with high doping concentration of $5 \times 10^{20} cm^{-3}$. The doping concentration of n-doped well extending from drain is $1.5 \times 10^{18} cm^{-3}$ while the doping concentration of p-doped FG is $8.5 \times 10^{17} cm^{-3}$. The main tunneling models used in the simulation are shown in Table 1. In order to better monitor the leakage and obtain more accurate retention characteristics, nonlocal tunneling is added to both the oxide/FG polysilicon interface and the oxide/silicon interface.

Fig. 3 depicts the variation of FG potential in EFM, UFM, FM1, and FM2 under different V_{cg} after 40-ns erasing operation intuitively. Under the same conditions, the potential change of EFM is several orders of magnitude larger than that of the other three devices. For example, at $V_{cg} = -12$V, the potential change of EFM is about 1V but the other three are less than 9e-5V.

As shown in Fig. 4, When the FG potential changes about 1V at $V_{cg} = -12$V, the time needed for EFM is about 34ns while the other three need more than 1.8ms. In the design of semiconductor devices, the balance between speed and power consumption is an important topic. We can speed up erasing operation by increasing V_{cg}, but the corresponding power consumption will increase ($P \propto V^2 f$, where P is power dissipation, V is voltage and f is frequency). If the change of FG potential is about 0.17V which can already make a distinction between state 0 and state 1, the erasing voltage of EFM can be decreased to -10V while the speed is within 40ns. Therefore, EFM can contribute to low power design.

III. PHYSICAL MECHANISM ANALYSIS

For EFM, the programming operation relies on channel hot electron injection (CHE), while erasing operation rely on Fowler-Nordheim (F-N) tunneling. Firstly, the U-shaped channel roughly doubles the F-N tunneling area. Then, the HfO_2, a high-k material, can increase the coupling capacitance between CG and FG by increasing the permittivity ($C \propto \varepsilon_r$, where C is coupling capacitance and ε_r is permittivity). Meanwhile, the downward extending CG can directly control the n-doped channel through HfO_2, thus greatly enhancing F-N tunneling. In fact, if the downward extending CG and the n-doped well are added to S side, the erasing speed will be further accelerated, but during the programming operation, the electrons accelerate from S to D side along the channel, and this introduction will reduce the electron concentration in S region, which will reduce the CHE speed.

As shown in Fig. 5(a) and Fig. 5(b), during the erasing operation, a highly negative bias is added to CG ($V_{cg} = -10$V), which drops a high potential on FG oxide layer, making the barrier thickness of FG oxide layer narrow down, a large number of electron tunnel from the p-typed polysilicon valence band to the conduction band of FG oxide layer and then flow rapidly to the silicon conduction band.

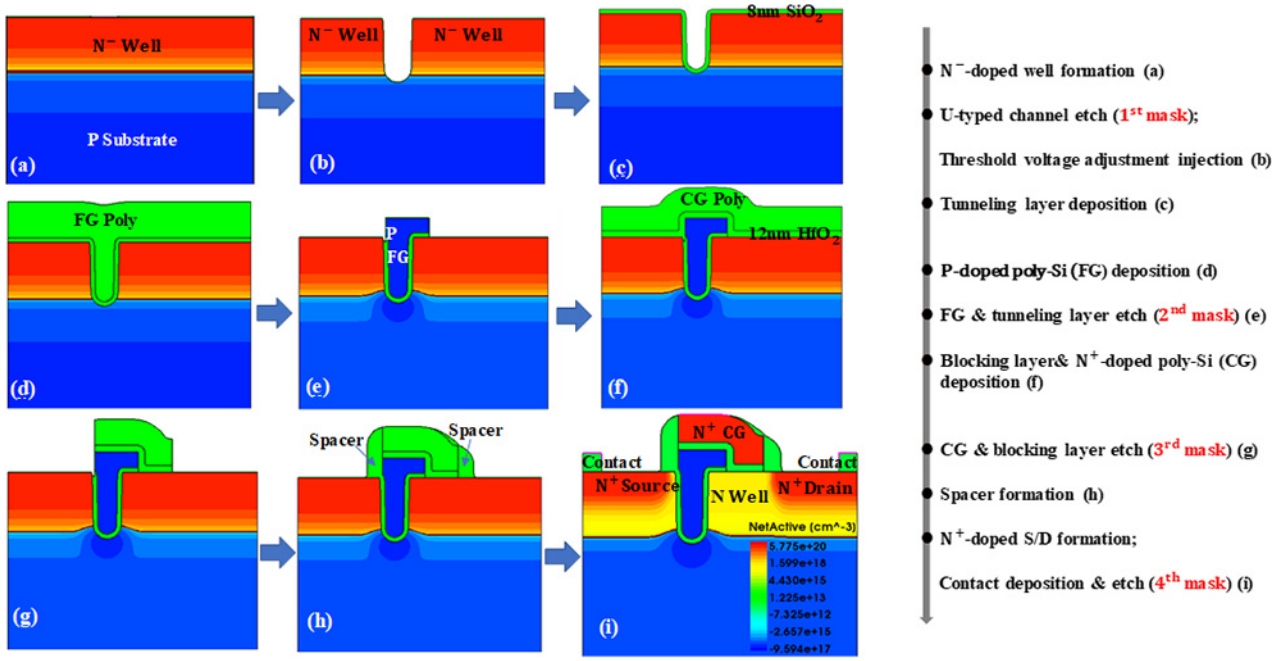

FIGURE 2. Schematic cross-sectional views of the key fabrication process steps of EFM.

TABLE 1. Main physical models selection.

Interface	Physical mechanism	Model selection
P^+/N^+ junction	Band2Band tunneling	Band2Band (E2)
Oxide/FG polysilicon	Nonlocal tunneling	eBarrierTunneling hBarrierTunneling
Oxide/silicon	Nonlocal tunneling & hot carrier injection	eBarrierTunneling hBarrierTunneling eLucky Injection

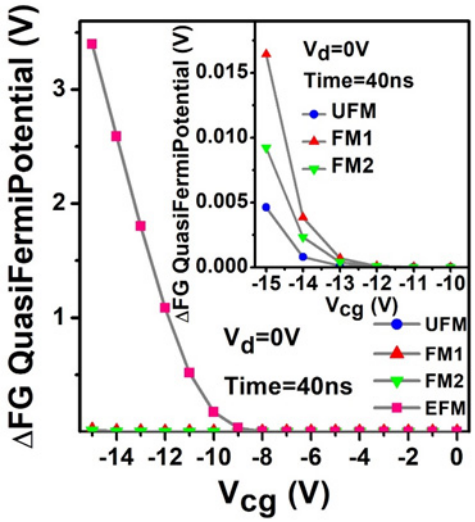

FIGURE 3. FG potential shift in EFM, UFM, FM1, and FM2 as a function of V_{cg} after 40-ns erasing operation.

The moving velocity of electrons in the conduction band of silicon oxide is very high, which can reach $10^7 cm/s$ [13]. Furtherly, these electrons will move into the silicon valence band and recombine with holes. This process can be equivalent to that the holes in the valence band of the silicon tunnel to the valence of FG oxide layer through the F-N tunneling, and then flow to the valence band of FG polysilicon.

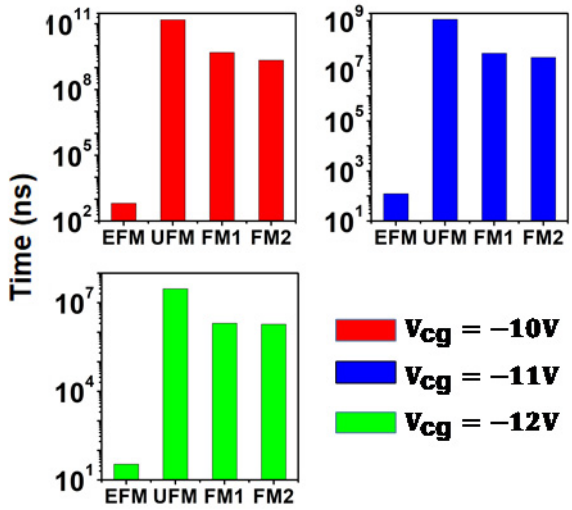

FIGURE 4. Comparison of the time needed for FG potential to change about 1V under different V_{cg} of EFM, UFM, FM1, and FM2 respectively ($V_d = 0V$).

The black triangles and squares in Fig. 5(a) and Fig. 5(b) represent the rate at which electrons are generated or disappear due to tunneling. The electron barrier tunneling rate at the interface of FG polysilicon and FG oxide layer is actually negative (electronic vanishing), there we take the absolute value for a better presentation. At another interface

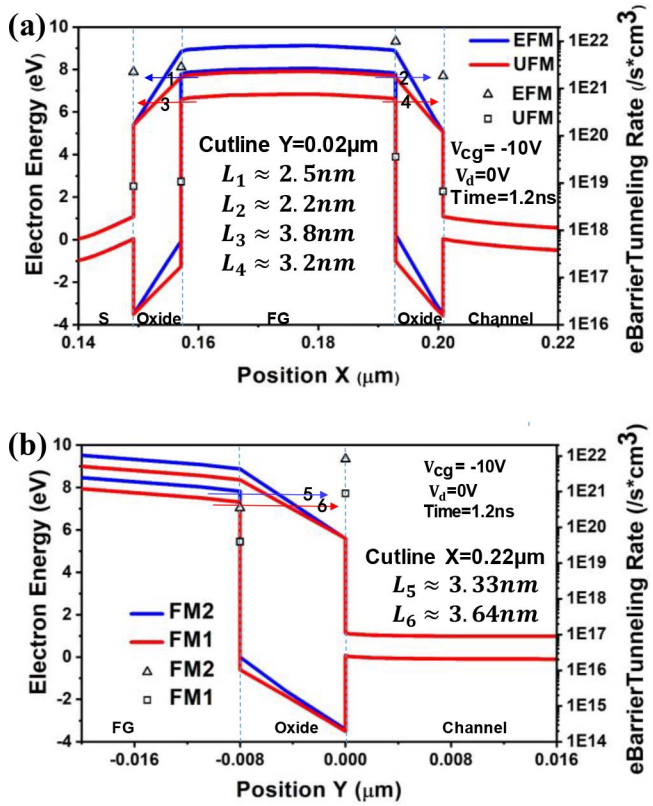

FIGURE 5. **Energy band diagrams (blue and red lines) and electron barrier tunneling rate (black triangles and squares) after 1.2-ns bias applied along (a) the cutline Y=0.02μm in Fig. 1(a) and Fig. 1(b). (b) the cutline $X = 0.22\mu m$ in Fig. 1(c) and Fig. 1(d).**

of FG oxide layer, electrons and holes are combined, and we can interpret the electron barrier tunneling rate here as the recombination rate of electrons and holes. The tunneling rate (T_t) is expressed as (1).

$$T_t \approx \frac{16E(U_0 - E)}{U_0^2} \exp\left(-2\sqrt{\frac{2m^*(U_0 - E)}{\hbar^2}}W\right) \quad (1)$$

where U_0 represents the height of the barrier, W is the width of the barrier, E is the energy of the tunneling electron, m^* is the effective mass of the carrier, and $\hbar = h/2\pi$, where h is the Planck constant. After obtaining the tunneling probability, the number of carriers existing in the starting area A is multiplied by the holes number of the target area B, and the tunneling current density (J_t) is obtained as (2).

$$J_t = \frac{qm^*}{2\pi^2\hbar^3} \int F_A N_A T_t (1 - F_B) N_B dE \quad (2)$$

where q is the charge quantity of a single electron, F_A, F_B, N_A and N_B represent the Fermi-Dirac distribution function and the density of states in the corresponding region respectively. The total tunneling current (I_t) is expressed as (3), where S_t is the total tunneling area.

$$I_t = J_t S_t \quad (3)$$

Therefore, the qualitative analysis of the tunneling current during the erasing operation is related to the four macroscopic variables: the number of electrons at the interface of FG oxide layer and FG polysilicon, the number of holes at the interface of FG oxide layer and the silicon, the tunneling rate and the total tunneling area.

We cut two lines along Y=0.02μm in Fig. 1(a) and Fig. 1(b) respectively and cut the other two lines along X=0.22μm in Fig. 1(c) and Fig. 1(d) respectively. Then, the electron energy band along these lines and the electron barrier tunneling rate at the two interfaces are represented in Fig. 5(a) and Fig. 5(b). 1∼6 are the corresponding shortest tunneling paths, and L_1∼L_6 are the distance of these shortest tunneling paths. There we do a brief qualitative analysis. As can be seen from Fig. 5(a) and Fig. 5(b), under the same bias and time condition, the tunneling distance L_2 of EFM is the shortest (about 58% of FM1). The tunneling rate is exponentially dependent on the tunneling distance, so as shown in Fig. 5(a) and Fig. 5(b), the electron tunneling rate at the interface of FG and FG oxide layer of the EFM is also the largest. In addition, we can find that whether it is the U-channel device or the horizontal channel device, the use of HfO_2 will shorten the tunneling distance and improve the tunneling rate at both two interfaces.

Looking separately, UFM with U-shaped channel can also reduce the tunneling distance (L_4 is shorter than L_5 and L_6), but the introduction of U-shaped channel reduces the holes concentration outside FG oxide layer, so the erasing speed of UFM has not been improved. As can be seen from the recombination rate of electrons and holes at the interface of channel and FG oxide layer in Fig. 5, EFM is higher than FM1 while UFM is the smallest. EFM can rapidly gather a large number of holes on the n-doped well directly through the coupling of HfO_2 during the erasing operation to maintain a high recombination rate for the tunneling electrons, thus maintaining the speed advantage. For FM2 with HfO_2 blocking layer, because the voltage applied to CG is coupled to the channel through 12-nm HfO_2, 27-nm FG and 8-nm SiO_2, three capacitors in series reduce the coupling capacitance between CG and the channel. But for EFM, the downward extending CG can directly control the channel on the right side of the FG via HfO_2 coupling. Therefore, under the same voltage applied to CG, FM2 has much less control over the channel than EFM, which results in a much slower speed than EFM.

In summary, there are three key factors: the U-shaped channel, the downward extending CG and the high-k layer under CG. It is the combination of these three factors that accelerates the erasing speed from millisecond level to nanosecond level. From the above analysis, we can see that it is impossible to achieve such a huge improve in the device performance for only one of the three reasons, which is also proved in the simulation results of Fig. 3 and Fig. 4. The combination of these three key factors is the magic of EFM.

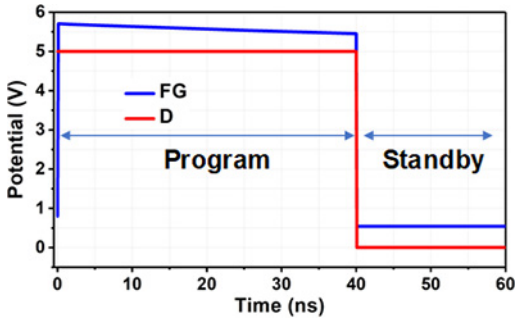

FIGURE 6. Potential of FG and D during programming and standby operation.

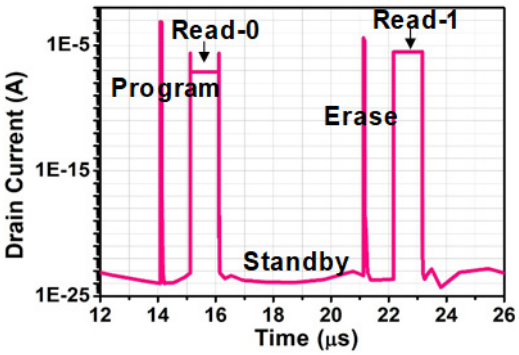

FIGURE 7. The I_d change curve of EFM with time in a transient simulation using the operation voltage in Table 2.

IV. MEMORY FUNCTIONS AND CORRESPONDING WEIGHT UPDATE CHARACTERISTICS

The dynamic range refers to the I_{on}/I_{off} ratio between the maximum current and the minimum current. The larger the dynamic range is, the stronger the mapping ability between the weights in the algorithm of neural networks and the current in the device, because we usually need to normalize the weights in the algorithms within a range (e.g., 0 to 1). For example, a I_{on}/I_{off} ratio of 1000 signifies that the minimum weight can be expressed as 0.001. In order to better simulate the biosynapses and realize the diverse applications based on neuro-inspired computing, the essential requirements for memory are more current states and higher resolution (certain noise tolerance) between different current states, thus providing finer weights representations to meet more accurate computational requirements [14].

Because the FG is p-type, the internal potential of FG is higher than D when no voltage is applied. When $V_{cg} = V_d = 5$V, the V_{cg} coupled to FG via HfO_2 will superimpose the initial potential of FG. As shown in Fig. 6, a reasonable CHE can be achieved when the FG potential is higher than D under the bias voltage of $V_{cg} = V_d = 5$V. If we use n-type FG, we need to increase V_{cg}, and as long as the channel is turned on and the FG potential is higher than D, we can also achieve successful programming. But choosing the same V_{cg} and V_d is more convenient for the specific voltage generation.

TABLE 2. Operation voltage and time of EFM.

	Program	Erase	Read 0/1	Standby after program/erase/read 0/read 1
V_{cg} (V)	5	-10	1.5	0
V_d (V)	5	0	1.1	0
PWHH of Fig. 6 / 7	40ns	40ns	1μs/1μs	1μs/1μs/5μs/5μs
PWHH of Fig. 10 / 11	1ns	1ns	2ns/2ns	1ns/1ns/1ns/1ns

Note: $V_s = V_{sub} = 0$V; PWHH represents peak width at half height.

Fig. 7 is a sequence diagram of memory operations. After 40-ns programming operation, the FG potential is decreased about 0.25V. The change of FG potential can cause a change in threshold voltage due to the capacitive coupling effect. During the subsequent reading operation, a small D current (I_d) of about 7.09e-8A can be measured. Then, an erasing operation is continued for 40ns, the FG potential is increased about 0.25V and a large I_d of about 2.9e-6A is read. The I_{on}/I_{off} ratio is over 40, indicating that state "1" and state "0" are successfully written. According to the specifications and design of the sense amplifier in the readout circuit, one can operate the memory with an appropriate voltage/time. When we set the erasing voltage V_{cg} to -10.4V, the programming voltage of V_{cg} and V_d both to 6V, and the operation time to 50ns, the switching ratio can over 1e6. The large switching ratio memory can be implemented, which can provide enough current states (for example, up to several hundreds of states) that are suitable for neuro-inspired computing. When we set the erasing voltage V_{cg} to -11V, the programming voltage of V_{cg} and V_d both to 6V, and the operation time to 10ns, the switching ratio is about 26. The ultrahigh-speed memory operation can be realized. In addition, due to the internal gain of FG memory, the arrays composed of EFM have an advantage in terms of the necessary gain and energy consumption of the peripheral circuitry for the analog application [8]. And as a gain cell, the read operation of EFM is nondestructive.

In order to verify the ability of EFM with different oxide thickness to hold state "1" under the same conditions, we first erase EFM with oxide thickness of 6nm, 7nm, and 8nm at the same voltage ($V_{cg} = -10$V and $V_d = 0$V). Over different times, they get the same change in FG charge. As we can see from Fig. 8(a)-(c), when the net change of FG charge is about 4e-14C, the time needed for the EFM with 6-nm FG oxide layer is 9.6ns, 7-nm FG oxide layer is 77.5ns, and 8-nm FG oxide layer is 545ns. Fig. 8(d)-(f) exhibit the ability to hold the charge of 4e-14C without a power supply of the three EFMs respectively. The simulation results indicate that the thinner the FG oxide layer is, the faster the erasing will be, but the corresponding retention performance will be reduced. 8nm is enough to hold the charge of 4e-14C because the electric leakage is negligible for a decade.

Fig. 9(a) displays the net changes of FG charge in EFM with three different thicknesses of tunneling oxide layer during the 40-ns programming operation ($V_{cg} = 5$V and $V_d = 5$V). Fig. 9(b) shows the

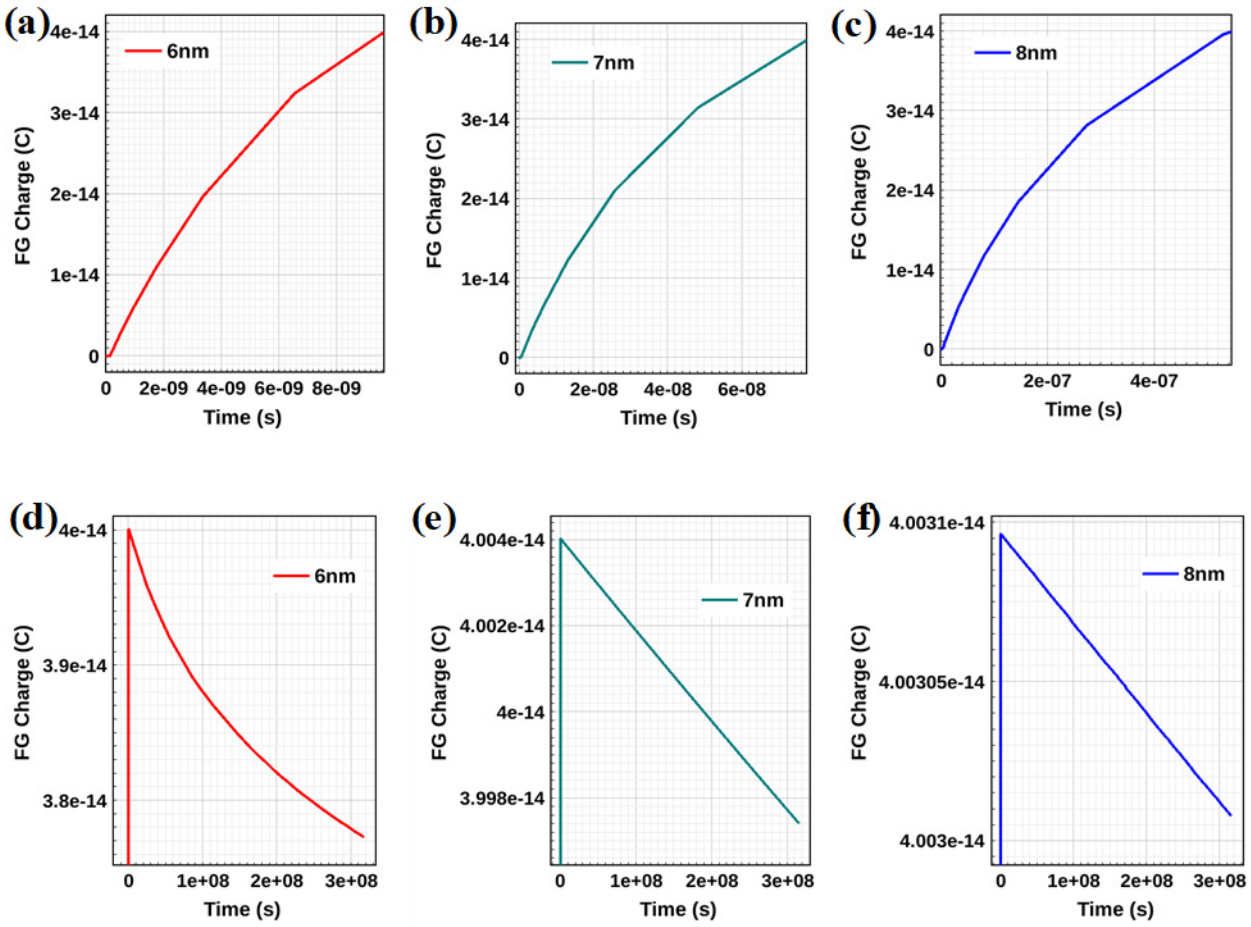

FIGURE 8. The ability of tunneling oxide layers of different thicknesses to hold state "1" at room temperature with power free.

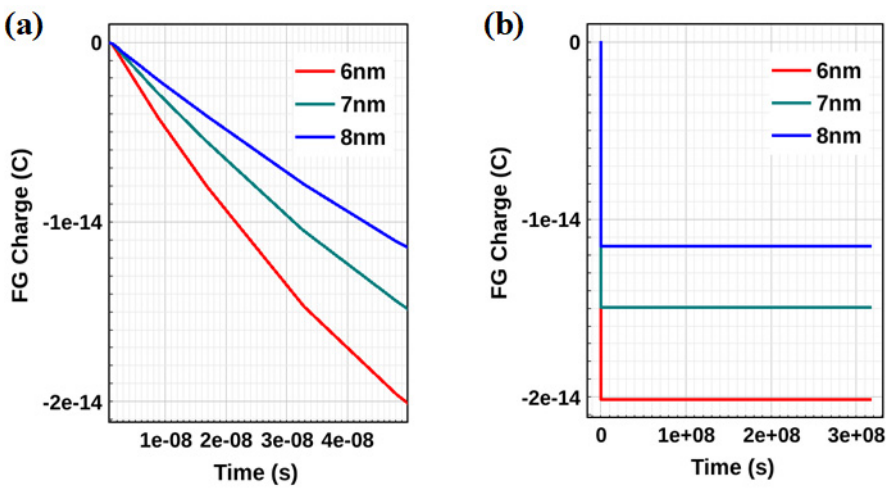

FIGURE 9. The ability of tunneling oxide layers of different thicknesses to hold state "0" at room temperature with power free.

corresponding retention characteristics after the 40-ns programming operation. All three thicknesses exhibit excellent retention performance when holding a low FG potential. Therefore, in the timing operation of programming/erasing, the programming operation is first performed to make FG potential become low, which can not only further reduce the erasing voltage to enhance the erasing effect but also improve the retention performance.

In the online training of artificial analog neural networks, the weight update nonlinearity/asymmetry is a critical issue, which can cause the learning accuracy loss [15]–[17].

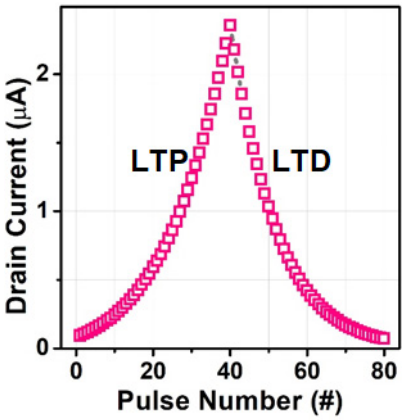

FIGURE 10. Weight-update (LTP/LTD) characteristics of EFM measured by using a series of identical pulses in Table 2.

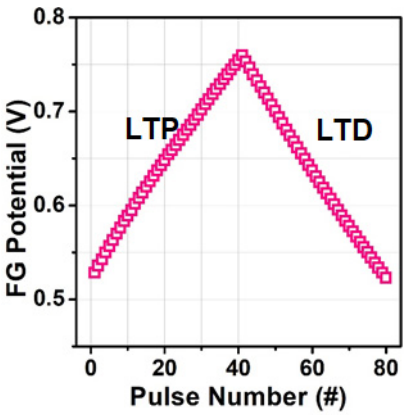

FIGURE 11. The FG potential variation of EFM corresponding to the weight update operation in Fig. 9.

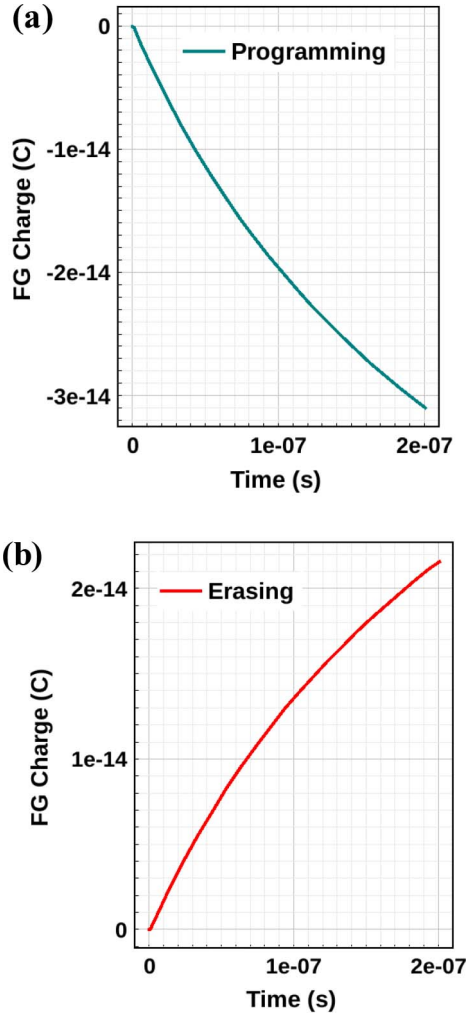

FIGURE 12. The net change of FG charge during (a) 200-ns programming operation, (b) 200-ns erasing operation using the operation voltage in Table 2.

It demands a continuous and smooth read current tuning, while for offline training, the nonlinearity can be handled with the iterative programming and/or erasing operation by write-verity technique [1]. For EFM, the weight update behavior (LTP/LTD) is shown in Fig. 10. The 40ns operation time of the erasing operation and the programming operation in Fig. 7 are divided into 40 independent voltage-pulses and the PWHH is 1ns. When the 40 identical erasing voltage-pulses are applied to EFM, I_d is approximately linear increase with the increase of the pulses number. When the 40 identical programming voltage-pulses are applied to EFM, I_d approximately linear decrease with the increase of the pulses number. Since the erasing speed of EFM can reach the nanosecond level to well match the programming speed, thereby a very symmetrical LTP/LTD weight update can be gotten. The corresponding variation of FG potential with the increase of pulse number during LTP/LTD operation is depicted in Fig. 11.

FG memory elements are suitable for adjustable conductance in the form of pseudo-crossbar with an accuracy of more than 1% [4]. Vector matrix operation is the most time-consuming step in neuro-inspired learning algorithm, but using a crossbar array structure to realize this weighted sum operation can greatly accelerate the neuromorphic computing [18]. Crossbar arrays consist of vertical rows and columns, and memory devices are located at the intersection of those rows and columns. When weights are updated, the programming/erasing voltage can be applied from both ends (rows and columns) to the devices, and the weights in the neural network are mapped to the current of FG memory devices. This operation can be performed in the array in parallel. And the weighted sum operation can also work in parallel: the read voltage is applied to all rows, and the current is read, thus generating the weighted sum of currents in each column.

As the time increases to 200 ns, we can see from Fig. 12 that the charge change in FG still presents a good approximate linear change. As shown in Fig. 13, the charge of FG tends to saturate when the time increases to 10μs. The reported artificial synaptic devices have less than 100 modulated conductance statements [19], [20] which is not beneficial to realize the continuous modulation of

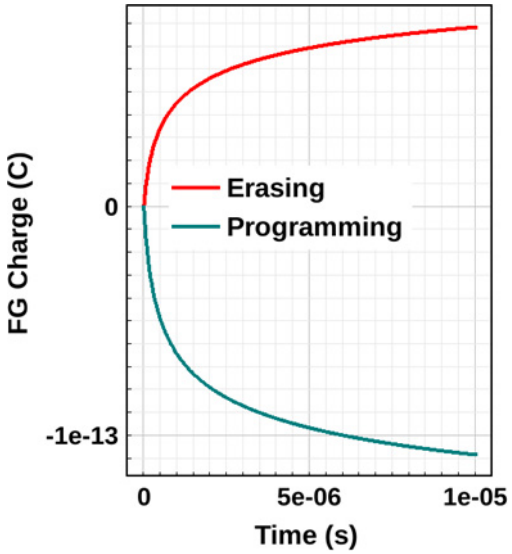

FIGURE 13. Saturation characteristics of EFM during 10-μs programming/erasing operation using the operation voltage in Table 2.

weights in neural networks. But for EFM, it will be easy to extract hundreds of current states.

V. CONCLUSION

Due to the great improvement of the erasing speed based on F-N tunneling model to nanosecond level, EFM can achieve a more symmetrical operation for programming and erasing, and show superior neural synaptic performance in LTP/LTD simulation. These performances enable it to be well applied to the online learning of analog neural networks. These simulation results show that EFM, a multi-bit memory, has the potential to save a lot of resources to store the weights and intermediate data during online training, and ensure a high calculation accuracy in neuro-inspired computing systems. A limitation of EFM is that, like all forms of FG memory, the operation voltages are large. Although our research has reduced the erasing voltage to a certain extent, further exploration is needed in the future.

REFERENCES

[1] S. Yu, "Neuro-inspired computing with emerging nonvolatile memory," *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, Feb. 2018.
[2] F. M. Bayat, X. Guo, M. Klachko, N. Do, K. K. Likharev, and D. B. Strukov, "Model-based high-precision tuning of NOR flash memory cells for analog computing applications," in *Proc. 74th Annu. Device Res. Conf. (DRC)*, 2016, pp. 1–2.
[3] F. M. Bayat, X. Guo, H. A. Om'mani, N. Do, K. K. Likharev, and D. B. Strukov, "Redesigning commercial floating-gate memory for analog computing applications," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Lisbon, Portugal, 2015, pp. 1921–1924.
[4] X. Guo *et al.*, "Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Austin, TX, USA, 2017, pp. 1–4.
[5] F. Faggin, G. S. Lynch, and J. S. Sukonick, "Brain emulation circuit with reduced confusion," U.S. Patent 4 773 024, Sep. 20, 1988.
[6] M. Holler, S. Tam, H. Castro, and R. Benson, "An electrically trainable artificial neural network (ETANN) with 10240 'floating gate' synapses," in *Proc. Int. Joint Conf. Neural Netw.*, vol. 2. Washington, DC, USA, 1989, pp. 191–196.
[7] A. Kramer, V. Hu, C. K. Sin, B. R. K. Gupta, R. Chu, and P. K. Ko, "EEPROM device as a reconfigurable analog element for neural networks," in *Int. Tech. Dig. Electron Devices Meeting (IEDM)*, 1989, pp. 259–262.
[8] X. Guo *et al.*, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in *Proc. Int. Electron Devices Meeting (IEDM)*, 2017, pp. 6.5.1–6.5.4.
[9] S. K. Esser, R. Appuswamy, P. A. Merolla, J. V. Arthur, and D. S. Modha, "Backpropagation for energy-efficient neuromorphic computing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1117–1125.
[10] A. Heinrich and S. Loth, "A logical use for atoms," *Science*, vol. 332, no. 6033, pp. 1039–1040, 2011.
[11] S. Y. Jiang *et al.*, "A semi-floating gate transistor with enhanced embedded tunneling field-effect transistor," *IEEE Electron Device Lett.*, vol. 39, no. 10, pp. 1497–1499, Oct. 2018.
[12] W. Wang *et al.*, "Design of U-shape channel tunnel FETs with SiGe source regions," *IEEE Trans. Electron Devices*, vol. 61, no. 1, pp. 193–197, Jan. 2014.
[13] R. C. Hughes, "High field electronic properties of SiO_2," *Solid-State Electron.* vol. 21, no. 1, pp. 251–258, 1978.
[14] T. Y. Wang *et al.*, "Flexible electronic synapses for face recognition application with multimodulated conductance states," *ACS Appl. Mater. Interfaces*, vol. 10, no. 43, pp. 37345–37352, 2018.
[15] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3498–3507, Nov. 2015.
[16] P.-Y. Chen *et al.*, "Mitigating effects of non-ideal synaptic device characteristics for on-chip learning," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Austin, TX, USA, 2015, pp. 194–199.
[17] H. Kosaka, T. Shibata, H. Ishii, and T. Ohmi, "An excellent weight-updating-linearity EEPROM synapse memory cell for self-learning neuron-MOS neural networks," *IEEE Trans. Electron Devices*, vol. 42, no. 1, pp. 135–143, Jan. 1995.
[18] L. G. Gao *et al.*, "Fully parallel write/read in resistive synaptic array for accelerating on-chip learning," *Nanotechnology*, vol. 26, no. 45, 2015, Art. no. 455204.
[19] Y. Park and J. S. Lee, "Artificial synapses with short-and long-term memory for spiking neural networks based on renewable materials," *ACS Nano*, vol. 11, no. 9, pp. 8962–8969, 2017.
[20] M. K. Kim and J. S. Lee, "Short-term plasticity and long-term potentiation in artificial biosynapses with diffusive dynamics," *ACS Nano*, vol. 12, no. 2, pp. 1680–1687, 2018.