Compact Ferroelectric Programmable Majority Gate for Compute-in-Memory Applications

Shan Deng[1*], Mahdi Benkhelifa[2*], Simon Thomann[2*], Zubair Faris[1], Zijian Zhao[1], Tzu-Jung Huang[1], Yixin Xu[3], Vijaykrishnan Narayanan[3], Kai Ni[1†], and Hussam Amrouch[2†]
[*]Equal contribution; [†]Email: kai.ni@rit.edu, amrouch@iti.uni-stuttgart.de
[1]Rochester Institute of Technology; [2]University of Stuttgart; [3]Pennsylvania State University

2022 IEEE International Electron Devices Meeting (IEDM) | 978-1-6654-8959-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IEDM45625.2022.10019400

***Abstract*— In this work, a compact and novel ferroelectric (FE) programmable majority gate is proposed and its novel application in Binary Neural Network (BNNs) is investigated. We demonstrate: i) by integrating *N* metal-ferroelectric-metal (MFM) capacitors on the gate of a transistor (1T-*N*-MFM structure), a nonvolatile and programmable majority (MAJ) gate that performs MAJ of AND between the gate input and polarization is realized; ii) validation the functionality of our 3-input MAJ of AND gate through comprehensive theoretical and experimental investigations; iii) a compact implementation of 3-input MAJ of XNOR gate that leverages only five of our 3-input MAJ of AND gates connected in parallel; iv) application of MAJ of XNOR gates to replace the XNOR gates and the first layer of the adder tree in the BNNs for up to 21x area saving on top of eliminating the energy-hungry memory accesses due to the compute-in-memory nature.**

I. Introduction

To accommodate the ever-growing deep learning models on chip and accelerate model execution through compute-in-memory architectures, there is a growing interest to exploit dense memory as the technology platform. Multi-gate transistor structures come naturally as promising candidates due to their high density and vertical 3D integration capability. One distinguished example is the workhorse of solid-state drive, NAND flash memory (Fig.1(a)), where multiple gates control different sections of the shared channel. The storage medium, such as floating gate, charge storage layer or FE, is local to each gate. Aside from being mass storage, it also finds applications in matrix-vector multiplication [1] and pattern matching [2] (Fig.1(a)) exploiting its NAND logic.

Another variant of multi-gate transistor has a shared floating gate (Fig.1(b)), which is no longer local to each gate, but instead shared among all the gates. This enables a fundamentally different functionality compared to the NAND structure. The channel conductivity is determined by the floating gate potential, which is a majority function of all the gates when more than two gates are present. Leveraging this structure, a 3-input majority gate that performs majority of FE polarization has been proposed [3]. When only two gates are used, this structure also enables non-destructive read of FE capacitor memory [4] and artificial FE neuron implementation for neuromorphic computing [5].

Various forms of majority logic function have been proposed previously using the multi-gate transistor with shared floating gate (Fig.2). For example, the 1T-*N*-MIM (metal-insulator-metal) structure (Fig.2(a)) can perform majority of the *N* gate inputs as they jointly determine the floating gate potential, and hence the channel current [6]. Since the insulator layer is a normal dielectric, the 1T-*N*-MIM structure is volatile and not programmable. The 1T-*N*-MFM (metal-ferroelectric-metal)-1MIM structure (Fig.2(b)) replaces the *N* MIM capacitors in the previous structure with FE capacitors and adds 1 MIM for read out [3]. It performs majority operation over *N* MFM polarizations, which is therefore nonvolatile. However, since the majority function is performed on the polarization, the function form is fixed, and not programmable and gate latency is limited by the polarization programming speed.

In this work, we propose a novel programmable and nonvolatile majority function using an 1T-*N*-MFM structure (Fig.2(c)). This transistor realizes MAJ of AND logic since it takes the majority over the *N* AND logic outputs between the gate inputs and the polarizations. By configuring the *N* MFM polarizations, the majority function over *N* gate inputs can be programmed. This structure is therefore highly versatile and find wide applications. As an example, we show that using MAJ of AND gates, a MAJ of XNOR logic gate can be constructed, which enables compute-in-memory acceleration of BNNs (Fig.2(d)) by replacing the XNORs and complex adder tree in BNN CMOS implementation with remarkable area saving and latency and energy reductions by eliminating memory access.

II. Theoretical and Experimental Verification of Majority of AND Operation

Fig.3(a) shows that AND operation between the gate input *G* and polarization *P* is naturally implemented in an 1T-1MFM structure. It is realized by encoding the low threshold voltage (V_{TH}) (polarization down) as *P*="1" and high-V_{TH} (polarization up) as *P*="0", and the read voltage V_G below low-V_{TH} as *G*="0" and V_G between low-V_{TH} and high-V_{TH} as *G*="1". In this case, a high drain current (I_D) is *only* possible when a high read bias is applied to the low-V_{TH} state. Following the same principle, in an 1T-*N*-MFM structure, the AND function is performed on each local MFM gate. The output of the AND operation in each local gate determines its contribution to the floating gate potential, majority of which determines the channel current. Fig.3(b) illustrates an example of 3-input programmable majority gate in which: only when the majority of the AND (MAJ of AND) outputs are bit "1", the channel is inverted and has a high current, hence achieving the proposed function.

To verify the functionality of MAJ of AND in 1T-*N*-MFM structure, we performed both theoretical investigation through TCAD simulations and experimental validation. First a baseline TCAD model of 14 nm FDSOI logic transistor is built (Fig.4(a)) and is used to calibrate the model transport and

electrostatic parameters using the reported I_D-V_G characteristics [7] (Fig.4(b)). Building on this, a 3-input programmable majority gate is built by inserting the MFM with a 10nm thick $Hf_{0.5}Zr_{0.5}O_2$ into the gate stack (Fig.4(c)). Using the ferroelectric Preisach model, calibrated with experimentally-measured Q_{FE}-V_{FE} curves [8], the transistor shows a counter-clockwise I_D-V_G curve, indicating ferroelectric switching.

Theoretical validation: The MAJ of AND operation is validated in TCAD. Different configurations of polarization are first written into the gates and then I_D-V_G curves are obtained by sweeping all the three gates (Fig.5(a)), two gates with the third gate fixed at logic "0" (Fig.5(b)), and one gate with the rest two gates fixed at logic "0" (Fig.5(c)). From those curves, read gate voltages to encode gate input logic "0" and "1" can be defined such that a high I_D is obtained when the majority of AND results between gate input and the polarization are bit "1". The electron density map for fixed input "111" (Fig.5(d)) and fixed polarization configuration "DDD", i.e., "111", (Fig.5(e)) confirms the majority operation. Fig.6 shows I_D corresponding to all combinations of gate input and polarization configuration, which is high only when the MAJ of AND gate outputs bit "1".

Experimental validation: To verify the TCAD simulation results, we fabricated 3-gate MFM capacitors on a heavily-doped p-type silicon wafer (Fig.7(a)). Bottom tungsten electrode is sputtered on the thermally grown SiO_2 and then covered by 10nm $Hf_{0.5}Zr_{0.5}O_2$ deposited through atomic layer deposition. Then top tungsten electrode is sputtered, followed by dry etching to open via on the bottom electrode. The final structure goes through a rapid thermal annealing at 600°C for 60s in N_2 atmosphere. A top-view SEM of 3-gate MFM is shown in Fig.7(b). To verify the functionality of 3-input majority of AND gate, we connect the bottom electrode of the 3-gate MFM capacitor with the gate of a discrete transistor, as shown in Fig.7(c). The Q_{FE}-V_{FE} hysteresis loops of the 3 MFM capacitors (Fig.7(d)) are almost the same. DC sweep of the 1T-*3*-MFM structure is shown in Fig.7(e) for the cases of sweeping 1, 2, and 3 gates simultaneously. Interestingly, it shows that the larger number of gates swept at the same time, the smaller the hysteresis window. The phenomenon that the memory window reduces with the number of simultaneous gate sweep is also observed in pulse measurement (Fig.8(a) and (b)). This is because the number of MFM gates effectively increases its area, A_{MFM}, and increases its capacitance. As the MFM is forming a capacitor divider with the transistor gate, the MFM voltage drop, V_{FE}, decreases (Fig.8(c)), and therefore the memory window reduces (Fig.8(d)).

To achieve a larger memory window for MAJ of AND gates demonstration, we propose to use the sequential write (Fig.8(e)) , where the programming of MFM gates is performed serially. Fig.8(f) shows that after sequential write of each gate, similar memory window is obtained by simultaneous sweeping of all gates during memory read. Fig.9(a), (b), and (c) show the I_D-V_G curves for different configurations of polarization by sweeping all the three gates, two gates with the third gate fixed at logic "0", and one gate with the rest two gates fixed at logic "0", respectively. By choosing a read bias of 0.5V/1.3V to encode the gate input "0"/"1", respectively, the I_D will be high *only* when the majority of AND outputs between the gate inputs and polarizations are bit "1", hence validating the logic functionality experimentally.

III. Implementation of MAJ of XNOR Gate and Application for Binary Neural Network

With the single transistor capable of implementing MAJ of AND logics, it is possible to construct a compact 3-input MAJ of XNOR gate. Fig.10(a) shows that to implement a 3-input MAJ of XNOR gate, it is necessary to connect five 3-input MAJ of AND gates in parallel, by following the logic relationship between the two logic gates (Fig.10(b)). The OR relationship in Fig.10(b) can be conveniently implemented by parallel connection of the MAJ of AND transistors. Fig.10(c) shows the simulated I_D using TCAD under different combinations of gate input and the polarization configurations, which is consistent with the logic truth table shown in Fig.10(d). This confirms the correct logic function of the 3-input MAJ of XNOR gate.

With a compact implementation of 3-input MAJ of XNOR gate, it can be used for the acceleration of BNN. Though BNN greatly simplifies the multiplication into simple XNOR gate when restricting the input and weight to be binary. However, for BNN with a large model size (Fig.2(d)), the adder tree exponentially grows and becomes a bottleneck. To address the issue, we approximate the 1-bit full adder (used in the entire first layer of the adder tree) with a MAJ gate (Fig.11(a)), as in [11], which together with the XNOR gate can be replaced with a compact MAJ of XNOR gate using our FE multi-gate structure. In this way, significant area saving (up to 21x reduction) is achieved as shown in the number of transistors needed to implement a 3-input XNOR and accumulation computation (Fig.11(b)). This area saving comes with only 2% accuracy loss when evaluating the Fashion MNIST dataset using the VGG-based BNN (Fig.11(c)). In addition, the energy-latency plot for the execution of compact gate shows that even excluding the memory access, our FE MAJ of XNOR approach shows a comparable energy-delay product as the CMOS implementation. All these results demonstrate great promise of our approach. Note that, the proposed device structure can also benefit from the vertical 3D structure and maximize its density.

IV. Conclusions

In this work, we propose a compact and novel ferroelectric programmable majority gate that can accelerate BNN inference. Through comprehensive theoretical and experimental investigations, the logic functionality of majority of AND operation between the gate input and polarization configuration is validated. Building on this, we implemented and validated the functionality of the majority of XNOR gate and shown significant area saving and memory access elimination when replacing the XNOR and adder tree in CMOS implementations of binary neural network. There the programmable majority gate is highly promising for compute-in-memory applications.

Acknowledgement: Authors thank O. Prakahs, A. Mema, A. Dave, F. Frustaci, M. Yayla for their help in BNN implementation/evaluation and S. Chatterjee for TCAD simulation. This work was primarily supported by U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Energy Frontier Research Centers program under Award Number DE-SC0021118 and partially supported by Army Research Office under Grant Number W911NF-21-1-0341 and NSF 2008365 and NSF 2132918.

REFERENCES: [1] P. Wang et al., IEEE TVLSI 2019; [2] F. Wang et al., IEEE EDL 2020; [3] J. Hwang et al., IEEE EDL 2022; [4] S. Ogasawara et al., Jpn. J. Appl. Phys. 2002; [5] G. Lee et al., IEEE EDL 2022; [6] T. Shibata et al., IEEE TED 1992; [7] Q. Liu et al., IEDM 2013; [8] K. Ni et al., IEDM 2021; [9] K. Ni et al., IEDM 2018; [10] R. Seyedramin et al., ICFPT 2019.

Compact Ferroelectric Programmable Majority Gate for Computing Applications

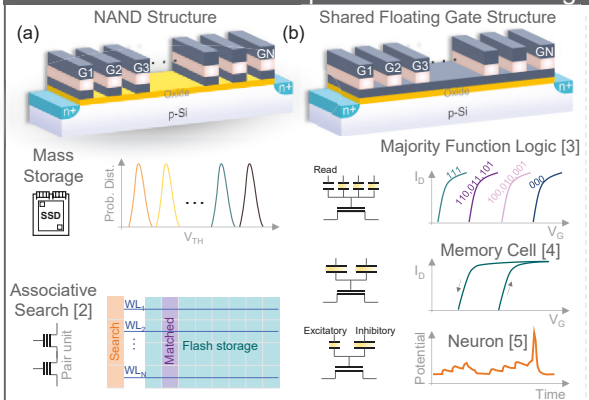

Fig.1 Applications of multi-gate memory string. (a) NAND structure and its applications. (b) Multi-gate structure with shared floating gate finds applications in majority logic, memory, and neuromorphic computing. This work proposes a new function: programmable majority gate.

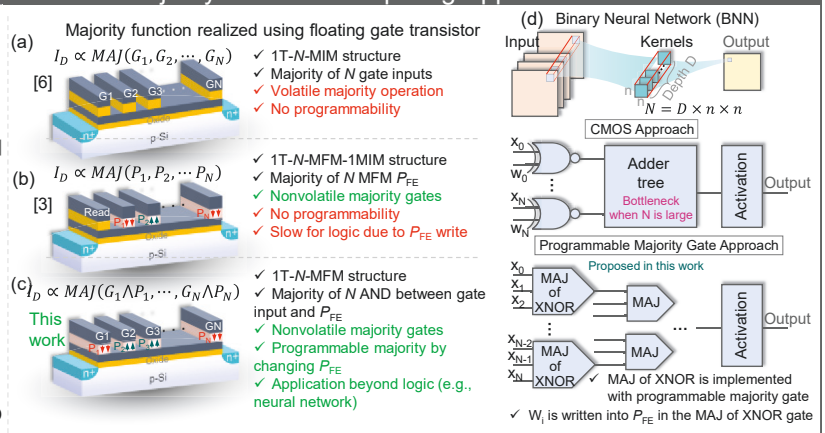

Fig.2. Previously proposed majority function using floating gate performs majority either on (a) the input or (b) the polarization. (c) We are demonstrating majority over the AND of input and polarization, making it a programmable majority gate. Such a functionality finds wide applications, e.g., (d) replacing XNOR & adder tree in BNN.

Operation Principles of Ferroelectric Programmable Majority of AND Gate

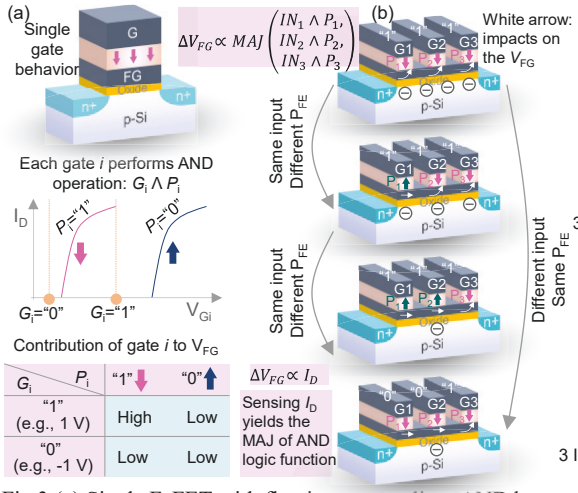

Fig.3.(a) Single FeFET with floating gate realizes AND between input G and polarization P. (b) 3-input structure implements AND function at each local gate, but also jointly determining the floating gate potential through a MAJ operation.

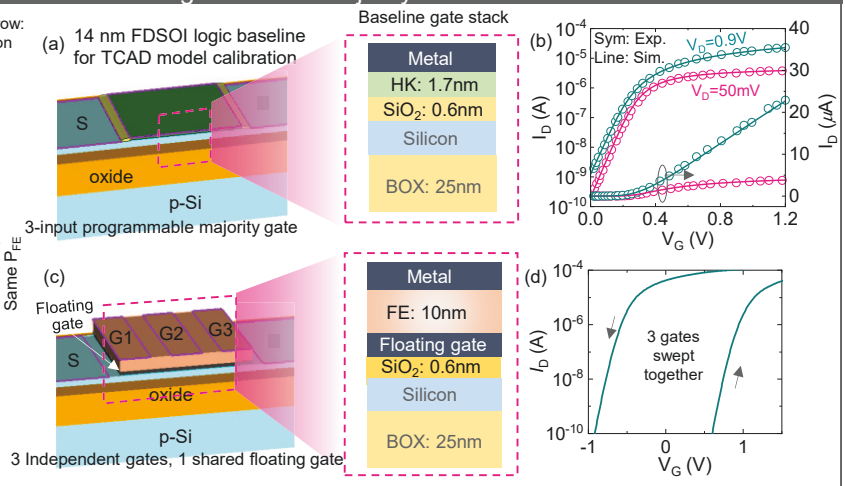

Fig.4. TCAD modeling setup for theoretical investigation. (a) 14nm FDSOI baseline transistor structure, which is used to (b) calibrate the TCAD model parameters with experimentally measured I_D-V_G curves. Using the model, (c) a 3-input majority gate is built and (d) simulated I_D-V_G curves using the standard ferroelectric Preisach model.

Verification of the Functionality of Majority of AND Gates in TCAD

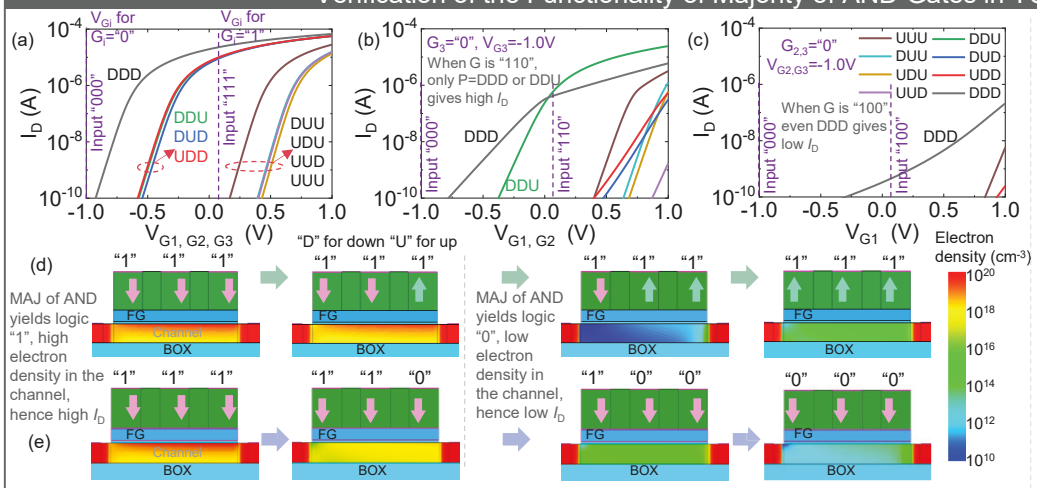

Fig.5. TCAD simulation results of (a) "111", (b) "110" , (c) "100" inputs under different polarization states. It shows that only when the AND results between the input G and polarization P give ≥2 bits of logic "1" can the I_D be high, otherwise, it is low. Electron density map for (d) fixed "111" input and varying polarization, (e) fixed polarization ("DDD", i.e., "111") and varying input. High electron density only occurs when majority of polarization is down in (d) and majority of gate input is "1" in (e), confirming the function of MAJ of AND.

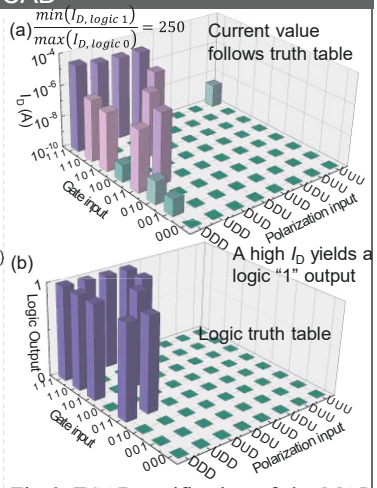

Fig.6. TCAD verification of the MAJ of AND function. (a) I_D under different configurations of input G and polarization P follows (b) the truth table, confirming the functionality.

Experimental Verification of Operation of Majority of AND Gate Using Multi-Gate FeFETs

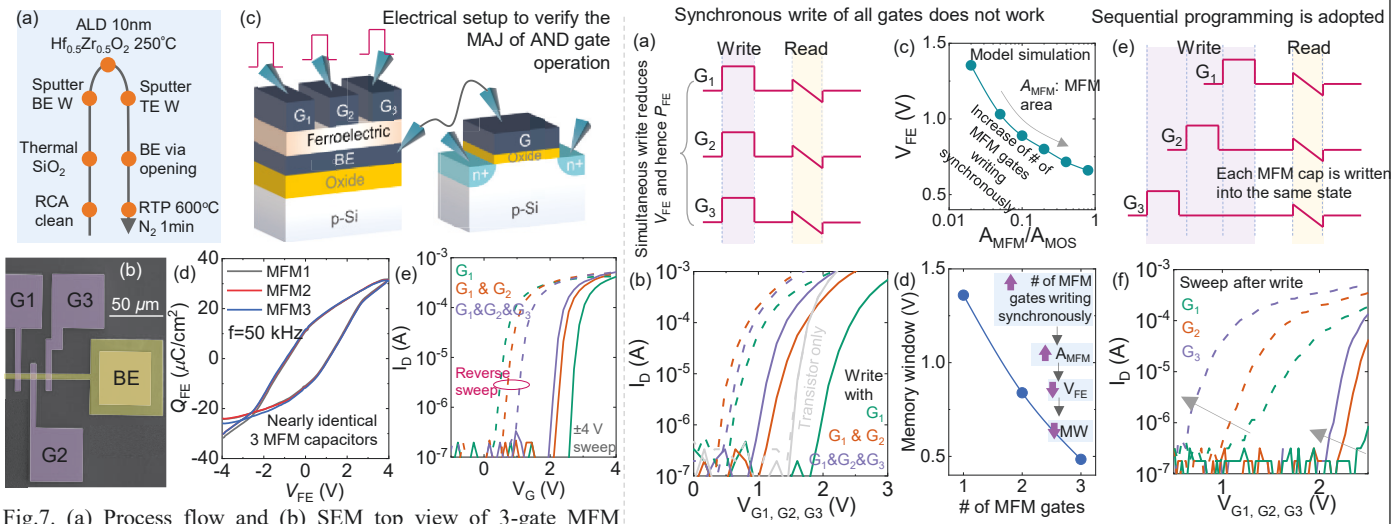

Fig.7. (a) Process flow and (b) SEM top view of 3-gate MFM capacitors. (c) Electrical setup for functionality verification of 3-input MAJ of AND gate. (d) 3 MFM capacitors have uniform Q_{FE}-V_{FE} hysteresis loops. (e) With increasing number of gates jointly swept during ±4 V DC sweep, the hysteresis window reduces.

Fig.8. Waveforms/I_D-V_G curves after programming for (a)/(b) synchronous write and (e)/(f) sequential write, respectively. Writing more MFMs synchronously (c) increases A_{MFM} and reduces V_{FE}, and hence (d) resulting in a smaller MW. Sequential write is a solution to this challenge.

Majority of XNOR Gate Operations and Application for Binary Neural Network Accelerator

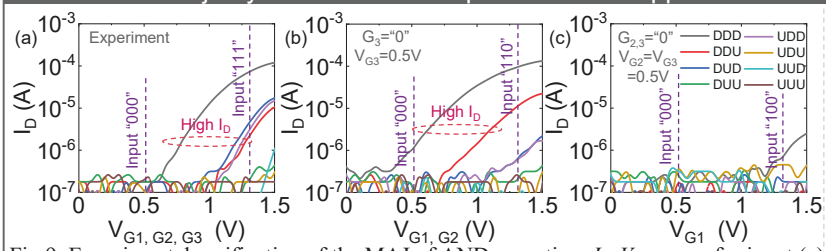

Fig.9. Experimental verification of the MAJ of AND operation. I_D-V_G curves for input (a) "111", (b) "110", and (c) "100" show successful MAJ of AND operation. Only when ≥2 gates storing bit "1" ("D") receive input bit "1", the I_D is high, confirming functionality.

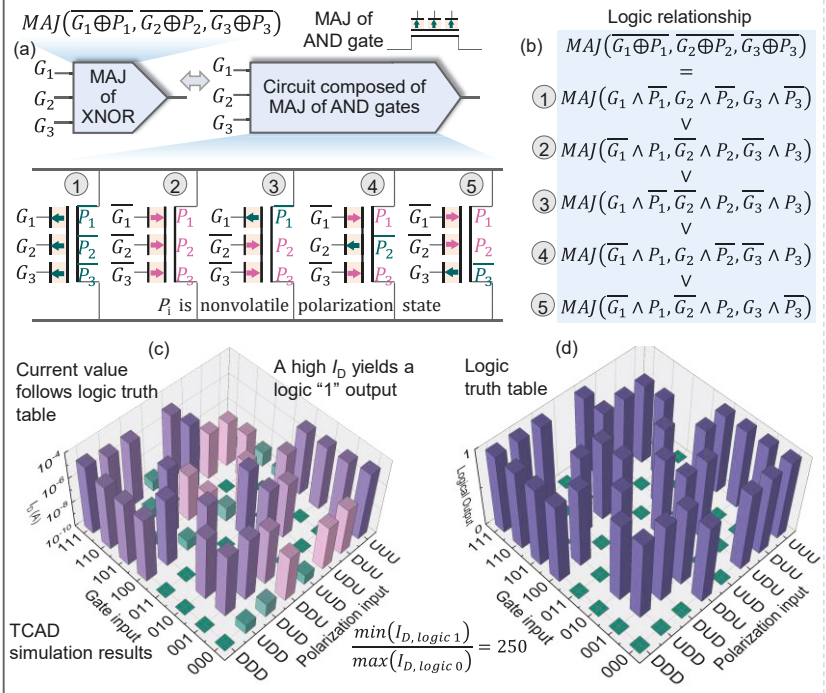

Fig.10. Construction of MAJ of XNOR gate from the basic MAJ of AND gates and verification. (a) Using 5 MAJ of AND gates, a 3-input MAJ of XNOR gates can be realized, by following (b) the logic relationship between the two. (c) TCAD simulated I_D under different configuration confirms (d) correct logic function.

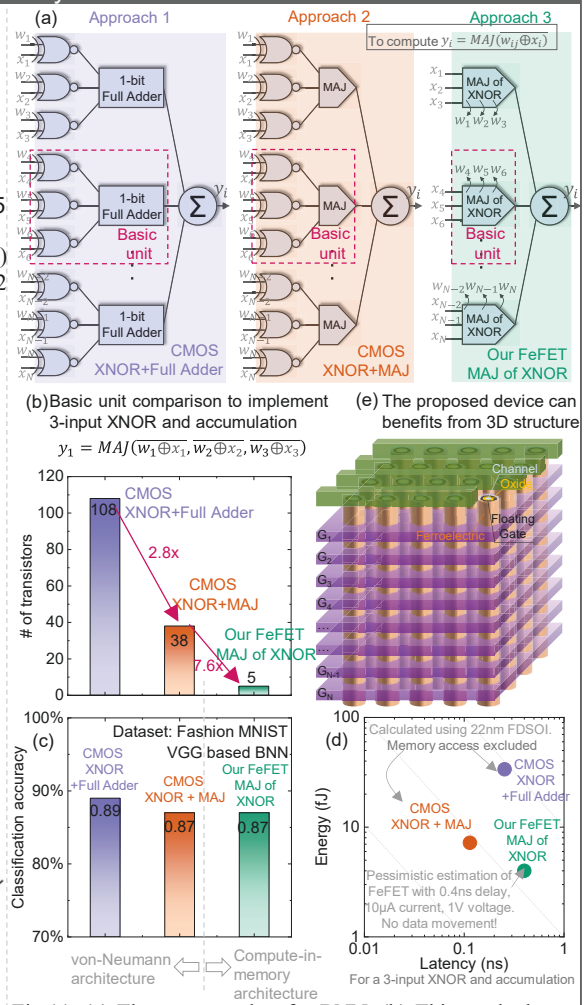

Fig.11. (a) Three approaches for BNN. (b) This work shows a much compact design, while (c) only 2% accuracy loss. (d) FeFET MAJ of XNOR shows comparable energy-delay product as CMOS, even after excluding memory access. (e) The proposed device can scale up by going to 3D.