A NOVEL AMBIPOLAR FERROELECTRIC TUNNEL FINFET BASED COMPUTING-IN-MEMORY FOR QUANTIZED NEURAL NETWORKS WITH HIGH AREA- AND ENERGY-EFFICIENCY

Runze Han[1,2], Jin Luo[2], Shengjie Cao[2], Qianqian Huang[2,3]* and Ru Huang[1,2,3]*

[1]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

[2]School of Integrated Circuits, Peking University, Beijing 100871, China

[3]Beijing Advanced Innovation Center for Integrated Circuits, Beijing 100871, China

*Corresponding Author's Email: luoj@pku.edu.cn, hqq@pku.edu.cn, ruhuang@pku.edu.cn

ABSTRACT

In this work, for the first time, a novel ambipolar ferroelectric tunnel FET (FeTFET) based computing-in-memory (CIM) scheme is proposed and experimentally demonstrated for quantized neural networks (QNNs) with high area- and energy-efficiency. By leveraging non-monotonic transfer characteristic of ambipolar tunnel FET (TFET) for signed multiplication and utilizing the non-volatile ferroelectric gate modulation for weight storage, the signed CIM cell with weight precision expansion can be implemented by one single FeTFET device based on 14-nm technology node platform. Based on the proposed FeTFET-based CIM design, QNNs for typical pattern recognition tasks are demonstrated with high accuracy and energy efficiency, showing its significant potential for edge AI applications.

INTRODUCTION

The quantized neural networks (QNNs) can guarantee network accuracy with reduced precision requirement of weights and activations, such as binary neural networks (BNNs) with binarized weights and inputs of "±1" (Fig. 1a). QNNs based on the computing-in-memory (CIM) architecture have attracted lots of attention for resource-constrained edge AI applications, due to the advantages of lightweight and high energy efficiency. Recently, the CIM design for QNNs has been proposed based on various memory technologies, such as SRAM [1], RRAM [2], PCRAM [3], ferroelectric FET [4,5], etc. However, the existing designs generally require complementary branches of computing and storage to achieve the non-monotonicity of signed multiplication in QNNs (Fig. 1b), still suffering from high hardware cost and energy consumption. In our previous works, the ambipolar ferroelectric tunnel FET (FeTFET) device has been proposed to implement non-monotonic operation in content-addressable memory (CAM) and encryption computing application [6,7].

In this work, a novel ferroelectric tunnel FinFET (FeFinTFET) based CIM for QNN is proposed and experimentally demonstrated for the first time. By utilizing the ferroelectric gate modulated ambipolarity of band-to-band tunneling (BTBT), the signed multiplication with expandable weight precision is realized by a single FeFinTFET device, without the need of complementary branches. Based on the proposed design, high-accuracy pattern recognition is demonstrated for edge AI with high energy efficiency.

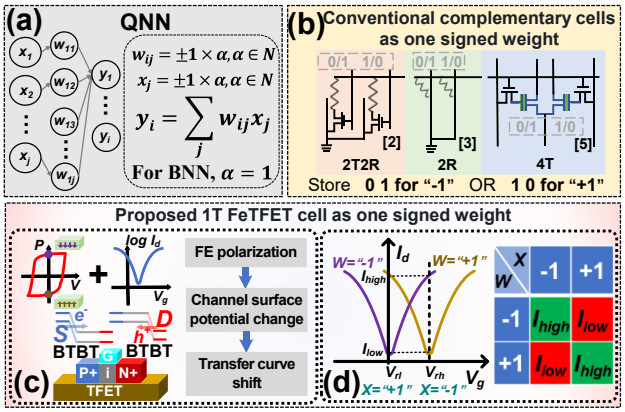

Figure 1: (a) The schematic diagram of a quantized neural network (QNN). (b) Traditional NVM Implementation for QNN. (c) Utilizing the bipolar characteristic curve to perform binary multiplication operations. (d) The schematic diagram of proposed 1T FeTFET cell for signed multiplication.

DESIGN AND EXPERIMENT OF FETFET-BASED CIM CELL FOR QNN

Due to the monotonic response of current to voltage, most non-volatile memories (NVMs) need complementary devices to realize signed multiplication. In our previous works, a three-terminal ferroelectric tunnel FET (FeTFET) device with non-volatile gate modulation and ambipolar transport characteristic has been proposed to achieve CAM and encrypted CIM cell [6,7]. In this work, we further demonstrate signed multiplication operations for QNN on a single FeTFET device and analyze the impact of drain voltage. Non-volatile ferroelectric gate modulation is used to implement weight storage, which can shift the ambipolar transfer characteristics of complementary BTBT behavior (Fig. 1(c)). By applying complementary read voltages, signed operations can be achieved. In detail, when the input is X="-1", the read current is low (I_{low}) for the storage state W="-1", and high (I_{high}) for W="+1". The

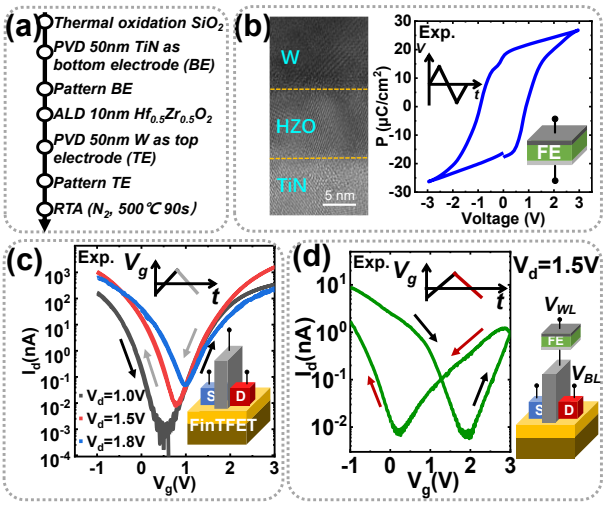

Figure 2: (a) The process flow of the FE layer. (b) P-V loop of the FE layer. (c) FinTFET transfer curve under different V_d. (d) FeFinTFET transfer curve.

opposite behavior occurs when X="+1". The signed multiplication operations can be realized in a single FeTFET without complementary branches of computing and storage (Fig. 1(d)). Furthermore, the multiply accumulate (MAC) operation can be naturally implemented with an FeTFET crossbar array by utilizing Kirchhoff's laws.

The proposed FeTFET is fabricated by connecting a 10nm $Hf_{0.5}Zr_{0.5}O_2$ ferroelectric (FE) layer to the gate of a TFET device. In contrast to our previous study where both top and bottom electrodes are titanium nitride (TiN) [6,7], Tungsten is used as the top electrode in this work (Fig. 2(a, b)). This choice can enhance the formation of the o-phase and steepen the P-V loop by improving tensile stress in the HZO film during annealing process [8], which helps to avoid the read disturbance for computing. The TFET structure is experimentally fabricated based on 14-nm FinFET technology node (Fig. 2(c)). When applying relatively large V_g, the transfer curve of fabricated FeFinTFET shows ferroelectric hysteresis behavior with ambipolarity in Fig. 2(d), due to the nonvolatile FE polarization switching behavior.

FEFINTFET-BASED CIM FOR QNN

FeFinTFET-based MAC of signed binary weights

As shown in Fig. 3(a, b), by applying different writing pulses, the weight (W) data of BNNs can be stored in the FE layer with two distinguishable polarization states, which modulates channel surface potential and leads to a shift in the transfer curve. The minimum current point on the transfer curve aligns with a gate voltage of V_g=0V when in the state W="-1", and shifts to V_g=1.5V for the alternate state, W="+1". The input (X) data is represented by the read voltage (V_{read}). A low V_{read} of 0V (V_{rl}) signifies

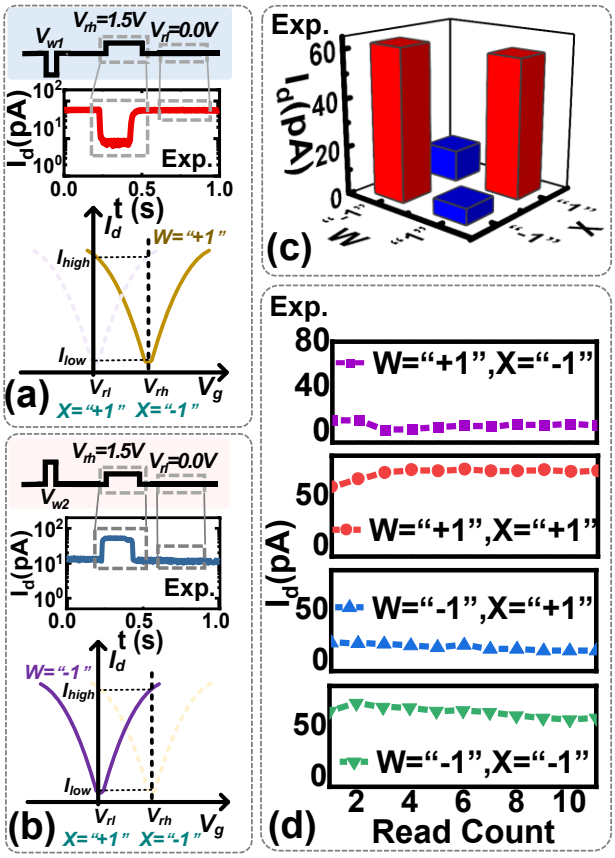

Figure 3: (a) The read currents of two different inputs "X" in the storage state W="+1". (b) The read currents of two different inputs "X" in the storage state W="-1". (c) Histogram of the output for binary multiplication. (d) Read current variation under different read cycles.

X="+1", while a high V_{read} of 1.5V (V_{rh}) indicates X="-1". For the storage state W="-1", a low read current (I_{low}) corresponds to X="+1", outputting "-1", and a high read current (I_{high}) corresponds to X="-1", outputting "+1". The case for W="+1" follows a similar pattern. The signed binary multiplication operation can be achieved by distinguishing the magnitude of the read current (Fig. 3(c)). The current is accumulated in each row of the FeTFET array, realizing the MAC operation of BNNs.

Moreover, the disturbance of read operations on the FeFinTFET storage state is analyzed (Fig. 3(d)). After writing the weight, the readout pulses with V_g=V_{rl} or V_g=V_{rh}, representing X="+1" and X="-1" respectively, are applied to the gate of FeFinTFET. The readout current remains stable under multiple read operations, showing good avoidance of read disturbance.

FeFinTFET-based MAC of signed trinary weights

Utilizing the multilevel ferroelectric polarization states, multibit signed weights can be stored within a single FeFinTFET, enabling in-situ weight precision expansion for QNN with quantized multi-bit weights. As

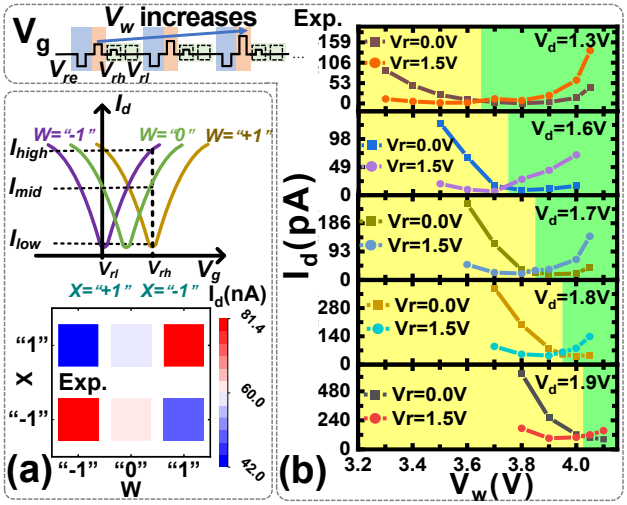

Figure 4: (a) Heat map of the output and schematic diagram for trinary wights multiplication. (b) The output current of two read states varies with the writing pulse amplitude V_w under different drain voltages V_d.

shown in Fig. 4(a), the multiplication of trinary weights and binary inputs for QNN is experimentally demonstrated. For the "W=0" state, the transfer curve is set between the W="+1" and W="-1" states. Under the read voltages of V_{rh} and V_{rl}, the output current corresponding to "W=0" is ideally consistent, which is approximately the average of I_{low} and I_{high}, indicating an output of "0", thus achieving the multiplication with trinary weights.

The effect of drain voltages (V_d) on the multiplication operation is further investigated. As shown in Fig. 4(b), when V_d is relatively low, the output currents for input data of "+1/0/-1" are consistently near I_{low} over a wide range of write pulses. To achieve a higher on-off ratio, a higher write voltage (V_w) is necessary for programming the weights, and the larger V_{rh} is also needed which will lead to read disturbance. When V_d is relatively high, the output currents input data of "+1/0/-1" are notably larger, leading to a smaller ratio of I_{low} to I_{high}. Moreover, as V_d increases, the V_g corresponding to the input data (X) increases due to the dual modulation effect of BTBT dependent on both V_g and V_d. Hence, to achieve a suitable ratio of I_{low} to I_{high} and to employ lower write and read voltages, a moderate V_d value is selected.

EVALUATION OF FEFINTFET-BASED CIM FOR QNN

Based on the proposed FeFinTFET-based CIM (Fig. 5(a)), the QNNs for pattern recognition tasks are demonstrated (Fig. 5(b)). The proposed FeFinTFET-based QNN maintains high accuracy compared with software implementation (Fig. 5(c)). In addition, using our previously established FeTFET model [6], the energy consumption of the FeTFET-based CIM is analyzed with

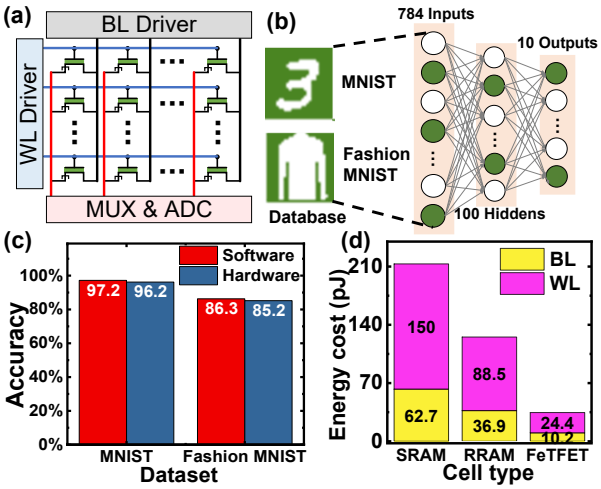

Figure 5: (a) Schematic diagram of FeFinTFET array. (b) The architecture of QNN based on proposed FeFinTFET CIM. (c) The accuracy of pattern recognition based on software and proposed QNN design. (d) Comparison of 1-bit MAC operation energy consumption among different QNN designs.

SPICE circuit simulation of 128×128 array. Benefitting from the reduction of word line (WL) and bit line (BL) capacitance due to the single-device cell design, our approach significantly reduces energy consumption during MAC operation compared with other designs (Fig. 5(d)).

CONCLUSION

This work reports the first experimental demonstration of an ambipolar FeTFET-based CIM for QNNs with high area- and energy-efficiency, in which the signed multiplication with weight precision expansion is realized with only one FeFinFET. Based on the proposed design, QNNs for high-accuracy pattern recognition are demonstrated with high area- and energy efficiency, showcasing its significant potential for edge AI systems.

ACKNOWLEDGEMENTS

This work was supported by NSFC (61927901, 62374009, 62404008), 111 Project (B18001), and China Postdoctoral Science Foundation (2024M750098, GZC20240026).

REFERENCES

[1] H. Valavi et al., IEEE JSSC, pp. 1789–1799, 2019.
[2] X. Sun et al., IEEE DATE, pp. 1423–1428, 2018.
[3] V. Joshi et al., Nat Commun, p. 2473, 2020.
[4] X. Chen et al., IEEE DATE, pp. 1205-1210, 2018.
[5] B. Fu et al., CSTIC, pp. 1–4, 2023.
[6] J. Luo et al., IEDM, pp. 36.5.1-36.5.4, 2022.
[7] J. Luo et al., IEEE VLSI, pp. 226–227, 2022.
[8] Y. Lee et al., IEEE TED, pp. 523-528, 2021.