FeMPIM: A FeFET-Based Multifunctional Processing-in-Memory Cell

Aibin Yan, Yu Chen, Zhongyu Gao, Tianming Ni, Zhengfeng Huang, Jie Cui, Patrick Girard, Xiaoqing Wen

To cite this version:

Aibin Yan, Yu Chen, Zhongyu Gao, Tianming Ni, Zhengfeng Huang, et al.. FeMPIM: A FeFET-Based Multifunctional Processing-in-Memory Cell. IEEE Transactions on Circuits and Systems II: Express Briefs, 2024, 71 (4), pp.2299-2303. 10.1109/TCSII.2023.3331267 . lirmm-04737485

HAL Id: lirmm-04737485
https://hal-lirmm.ccsd.cnrs.fr/lirmm-04737485v1

Submitted on 15 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FeMPIM: A FeFET-based Multifunctional Processing-in-Memory Cell

Aibin Yan, Yu Chen, Tianming Ni, Zhengfeng Huang, Jie Cui, Patrick Girard, *Fellow, IEEE,*
and Xiaoqing Wen, *Fellow, IEEE*

***Abstract*—The Von-Neumann memory wall bottleneck that keeps expanding is mainly caused by the frequent data transfer between the main memory and the processor. The Processing in-Memory (PiM) capabilities of emerging nonvolatile devices have the potential to partially alleviate the memory wall problem. In this paper, we use the ferroelectric field-effect transistor (FeFET), one of the emerging nonvolatile devices, to design a multifunctional processing in-memory cell, namely FeMPIM. It can perform multiple logic operations in computing mode as well as content searching in ternary content-addressable memory (TCAM) mode. Simulation results demonstrate the multifunctional capability of the proposed FeMPIM as well as its moderate overhead when compared with the complementary metal-oxide-semiconductor (CMOS) based and the existing FeFET-based devices.**

***Index Terms*—Ferroelectric field-effect transistor, ternary content-addressable memory, processing-in-memory**

I. Introduction

WITH the performance of processors and memories ever-increasing, the separation of calculation and storage in the conventional Von-Neumann architecture constrains the performance of computing systems, which is known as the memory wall [1]. The memory wall problem limits the data transfer between the processor and memory, and it causes extra latency and power overhead [2]. To solve the memory wall problem, researchers proposed PiM that can process the computational tasks in the memory. By reducing computational tasks in the processor, PiM can alleviate the memory wall problem.

PiM is realized with traditional transistors or emerging non-volatile devices. Researchers have proposed single-function PiM cells [3-5] and multifunctional PiM cell arrays [6-8]. However, these designs have some disadvantages, such as negative voltage supply, low storage density, and complexity of control circuits. In this paper, we propose a novel FeFET-based multifunctional PiM cell. Our major contributions are summarized as follows.

(1) We propose a FeFET-based multifunctional PiM cell that can provide storage, logic calculation, and TCAM functions.

(2) We design control signals for the proposed FeMPIM and consider its implementation in an array. In the array, FeMPIM requires only sense amplifiers (SAs) to provide results for different functions.

(3) We compare with several existing similar designs. The comparison results show that the proposed design has a high memory density and moderate power latency overhead.

II. Backgrounds

A. The Ferroelectric Field-Effect Transistor

The FeFET is a novel non-volatile device. Figure 1 (a) shows the structure of the FeFET device. The FeFET contains a ferroelectric layer composed of ferroelectric material. The ferroelectric material provides non-volatility, allowing the FeFET to retain its state after power loss.

The polarization of the FeFET is an important parameter to indicate its state. The FeFET switches its polarization according to the positive and negative gate-source voltages (Vgs). Figure 1 (b) shows the polarization-voltage characteristic of the FeFET. Here, we define the positive polarization as state "1" and the negative polarization as state "0". The FeFET has low resistance in positive polarization and high resistance in negative polarization.

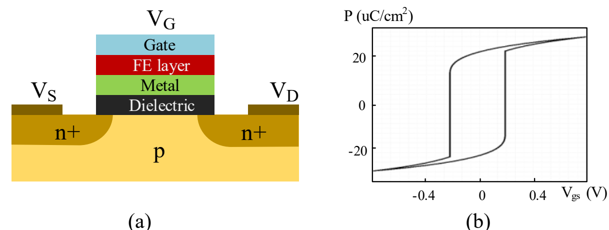

Fig. 1. The FeFET. (a) The device-level structure of FeFET. (b) The p-v characteristic of FeFET.

Aibin Yan is with School of Computer Science and Technology, Anhui University, and School of Microelectronics, Hefei University of Technology, Hefei 230601, China (email: abyan@mail.ustc.edu.cn).

Yu Chen and Jie Cui are with School of Computer Science and Technology, Anhui University, Hefei 230601, China. (email: chenyu_ao@163.com, cuijie@mail.ustc.edu.cn).

Tianming Ni is with School of Integrated Circuits, Anhui Polytechnic University, Wuhu 241000, China. (email: timmyni126@ 126.com).

Zhengfeng Huang is with School of Microelectronics, Hefei University of Technology, Hefei 230009, China. (email: huangzhengfeng@139.com).

Patrick Girard is with the Laboratory of Informatics, Robotics and Microelectronics of Montpellier, University of Montpellier / CNRS, Montpellier 34095, France (email: girard@lirmm.fr).

Xiaoqing Wen is with Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology, Fukuoka 8208502, Japan (email: wen@cse.kyutech.ac.jp).

B. Existing FeFET-based PiM Designs

FeFET-based Single-function designs have been proposed [3-5]. However, their single-function and additional circuits limit their applications. Recently, researchers proposed FeFET-based multifunctional cell array designs [6-8]. These designs combine extra cells to achieve additional functions (see Fig. 2).

Figure 2 (a) shows the structure of the attention-in-memory (AiM) array [6]. An AiM cell is located in the dotted wireframe. The AiM array can provide storage and TCAM functions. Figure 2 (b) shows the structure of the FeFET-based memory architecture (FeMAT) [7]. A FeMAT cell is located in the dotted wireframe. The FeMAT implements storage, logic calculation, and TCAM functions. In addition, several designs using multiple FeFET devices have been proposed [9-11]. The multiple FeFET design increases the cell's storage capacity.

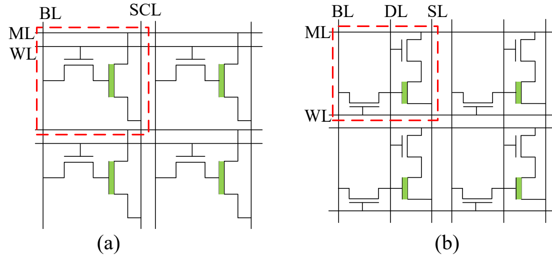

Fig. 2. Schematic of existing multifunctional FeFET designs. (a) Attention-in-memory array [6]. (b) FeFET-based memory architecture array [7].

III. Proposed FeFET-Based Multifunctional Processing In-Memory Cell

A. FeMPIM Cell

Figure 3 shows the schematic of the proposed FeFET-based multifunctional PiM cell, namely FeMPIM. The FeMPIM consists of two FeFET devices (i.e., FeFET1 and FeFET2) and two CMOS transistors. The source of FeFET1 is connected to ML2 through a transistor controlled by WL2. The source of FeFET2 is connected to ML1 through a transistor controlled by WL1.

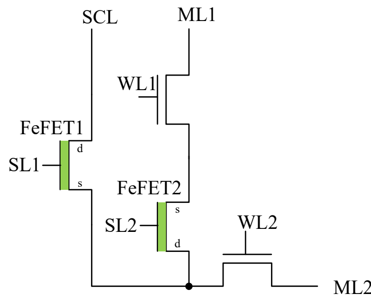

Fig. 3. Schematic of the proposed FeMPIM cell.

B. FeMPIM Array Structure

Figure 4 shows the array structure of the proposed FeMPIM. In the array, the selected SLs, WLs, and MLs are used to determine the address and input data. The unselected lines are kept at the opposite level of the selected lines to maintain the state of the other cells. SAs for read operations are connected to MLs. Due to the separation of reading and writing, SAs and MLs do not generate data conflicts. As described below, the proposed FeMPIM provide different functions based on SCLs and MLs. Compared with the other array designs, the proposed FeMPIM array avoids negative voltages and the calculation results can be read out directly by the SAs. Using negative voltages in the array affects the other cells on the same line. Eliminating these effects also needs negative voltages. Performing calculations outside the array will make the array less parallel.

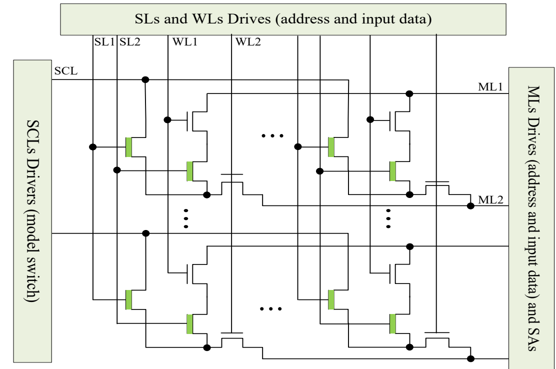

Fig. 4. Array structure of the proposed FeMPIM.

C. Memory Mode

In memory mode, the proposed FeMPIM can write and read the states of FeFET1 and FeFET2, independently. The write voltage (Vw) is a voltage above the threshold of Vgs. We apply voltages to the gate and source to generate Vgs and design the control signal according to this method.

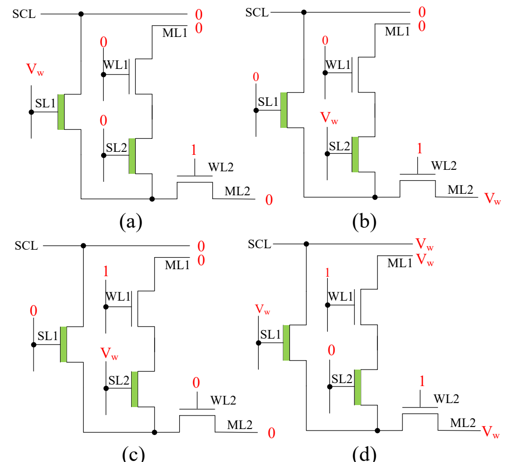

Fig. 5. Write operation in memory mode. (a) Write "1" to FeFET1. (b) Write "0" to FeFET1. (c) Write "1" to FeFET2. (d) Write "0" to FeFET2.

Write Operation: Figure 5 shows the write operation of the proposed FeMPIM. Figure 5 (a)-(b) shows the operation when writing "1" and "0" to FeFET1. During the write

operation, WL1 is low and WL2 is high. FeFET1 receives the appropriate Vgs by SL1 and ML2. SL2 is used to prevent FeFET2 to be affected. Figure 5 (c)-(d) shows the write operation of FeFET2 that is based on the same principle as the write operation of FeFET1. In the array, the other cells can be maintained in their original states by turning off WLs and applying Vw at the other terminals of FeFET devices.

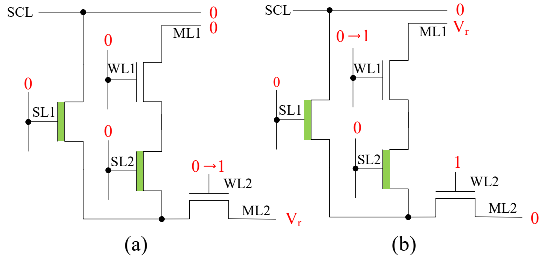

Fig. 6. Read operation in memory mode. (a) Voltage read for FeFET1. (b) Voltage read for FeFET2.

Read Operation: Figure 6 shows the read operation of the proposed FeMPIM. Figure 6 (a) shows the read operation of FeFET1. When FeFET1 is read, SL1, SL2, SCL, WL1, and ML1 are low. Firstly, WL2 is low and ML2 is pre-charged to Vr. Then, WL2 switches to high. At this time, if the state of FeFET1 is “1”, the voltage of ML2 drops quickly; if the state of FeFET1 is “0”, ML2 is maintained. Finally, the FeFET1 is determined by detecting the voltage of ML2. Figure 6 (b) shows the read operation of FeFET2. When reading FeFET2, ML1 is pre-charged to Vr, and the other operations are similar to reading FeFET1. The FeFET2 is determined by ML1.

D. Computing Mode

Logic computations using non-volatile devices are characterized by low power consumption, non-destructive reads, and high flexibility compared to logic gates in CMOS technology. Fig. 7 shows the OR and AND operations of the proposed FeMPIM in computing mode.

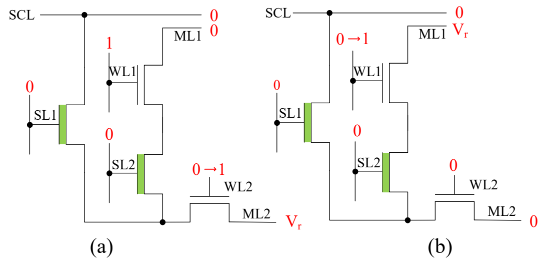

Fig. 7. Computing mode. (a) OR operation. (b) AND operation.

The states of FeFET1 and FeFET2 indicate the inputs to the OR and AND operations. Fig. 7 (a) shows the OR operation. WL1 holds high and performs a read operation on FeFET1.The result is “0” only when the FeFET1 and FeFET2 are “0”. In the other cases, the result is “1”. Fig. 7 (b) shows the AND operation. WL2 holds low and performs a read operation on FeFET2. The overall resistance of FeFET1 and FeFET2 is low only when the states of FeFET1 and FeFET2 are “1”. Since FeFET1 and FeFET2 are connected in series, the result of the read operation is “0” when FeFET1 or FeFET2 is “0”.

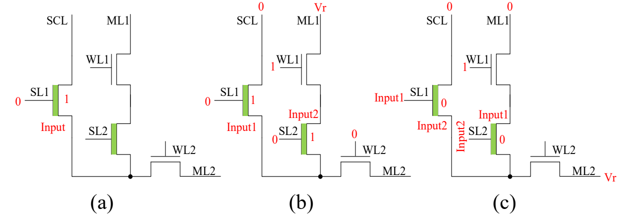

Fig. 8. NOT, NOR, and XOR operations. (a) NOT operation. (b) NOR operation. (c) XOR operation.

Figure 8 shows the NOT, NOR, and XOR operations. The NOT operation requires a FeFET device with the state “1”. As shown in Fig. 8 (a), its gate is low and its source is the input. When the input is high, the state of the FeFET changes to “0”; when the input is low, the state remains “1”. As shown in Fig. 8 (b), the NOR operation requires NOT operations on two inputs and then performs an AND operation. The XOR operation requires two FeFET devices with the state “0”. The input method is shown in Fig. 8 (c). When the two inputs are the same, the states of FeFET devices remain “0”; when the inputs are different, one state of the FeFET device changes to “1”. Then, the XOR operation performs an OR operation.

E. TCAM Mode

Figure 9 shows the match operation of the proposed FeMPIM in TCAM mode. Before the match operation, FeFET1 and FeFET2 have stored a pair of complementary states (i.e., S and S’). SCL and ML1 are a pair of complementary inputs Y and Y’ (the voltage of high input is Vr). WL1 and WL2 are high. ML2 is pre-charged to Vr. We can determine the match result between S and Y by the voltage at ML2. If S and Y are the same, the low input is connected to a FeFET device with state “0”. Therefore, the voltage on ML2 will not drop. If S and Y are reversed, the low input is connected to a FeFET device with state “1”, and the voltage of ML2 drops rapidly.

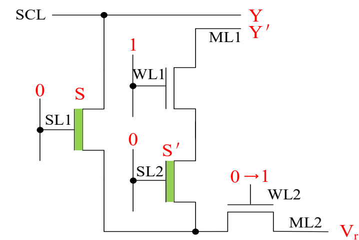

Fig. 9. The match operation of the proposed FeMPIM in TCAM mode**.**

In particular, in the TCAM’s “don’t care” mode, the states

of S and S’ are both “0”. In this case, the cell necessarily provides the result that S and Y are matched. The “don’t care” mode allows the data to be fuzzily found and returns the address with partially identical data. This will significantly increase the effectiveness and range of the match operation.

F. Simulations

Figure 10 shows the timing simulation result of the proposed FeMPIM in memory mode. In the simulation, the proposed FeMPIM performs write and read operations on FeFET1 and FeFET2, respectively. Figure 11 shows the timing simulation result of the proposed FeMPIM in TCAM mode. In the pre-charge phase, ML2 is pre-charged and FeFET devices switch polarization. In the match phase, if the inputs (SCL and ML1) are matched with FeFET devices, ML2 holds; conversely, ML2 drops.

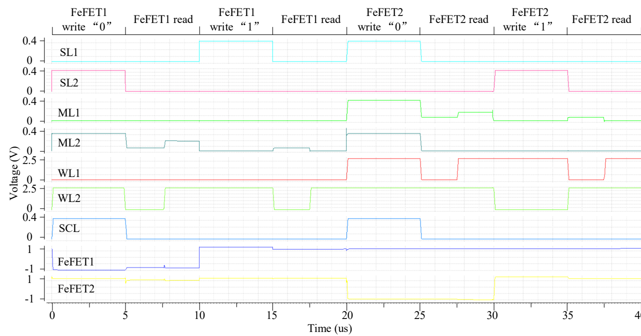

Fig. 10. Timing simulation result of the proposed FeMPIM in memory mode.

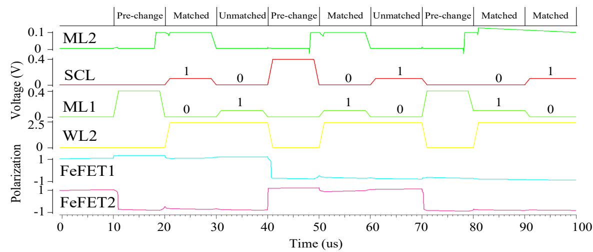

Fig. 11. Timing simulation result of the proposed FeMPIM in TCAM mode.

IV. Comparison and Evaluation

We measured the latency and energy overhead of FeMPIM in Cadence Virtuoso. We used the FeFET model calibrated in [12-13] and TSMC 65nm CMOS transistors for simulations. In the FeFET model, the kinetic coefficient (ρ) was set to 0.25 and the ferroelectric layer thickness (TFE) was set to 6 nm. ρ influences the polarization switching time and TFE influences the threshold voltage [12]. For the simulations, VDD was set to 2.5V, Vw was set to 0.4V, and Vr was set to 0.1V.

We compared existing designs using FeFET devices. The comparisons include memory density, power and latency. The density is the bit-cell memory density that indicates the number of bits stored in the same area. A higher density results in a higher utilization of the area. The power indicates the transient power when performing different operations and it is calculated by summing the power consumption from the supply voltage. The latency indicates the differential time between the changes of two adjacent effective voltages.

Table I shows density, power and latency comparisons in memory mode for alternative designs. Through sharing transistors, the proposed FeMPIM has a high density. The negative voltage requires a charge pump, which generates additional power consumption and area overhead. Therefore, we divide the solutions in Table I into two groups based on the use of negative voltages. The write power of the proposed FeMPIM is moderate among the designs without the negative voltages. In terms of read power overhead, the FeMPIM reduces the number of CMOS transistors and the leakage current. Therefore, the proposed design has a low read power overhead. The read latency of AiM and TI-FeFET is low, but their read power overhead significantly exceeds the other designs. Moreover, the proposed FeMPIM has large write latency and moderate read latency compared to the other designs.

For computing mode, the existing designs read out the values outside the memory array and use gate circuits to calculate the results of the values. The proposed FeMPIM performs calculations in the memory array and reads out the result of the calculation. Compared to the existing designs, the proposed FeMPIM has high parallelism in calculation mode but this can increase the complexity of control signals.

TABLE I
Density, Power and Latency Comparisons in Memory Mode for Alternative Designs

Designs	Negative Voltage*	Memory density[#] (bit/um²)	Read Power (nW)	Read Latency (ps)	Write Power (nW)	Write Latency (ps)
AiM [6]	Yes	4.46	1193	0.69	0.13	9.30e2
FeMAT [7]	Yes	2.98	0.72	21.87	0.28	1.03e3
Re-FeMAT [8]	Yes	2.98	1.20	17.70	0.20	9.70e2
TI-FeFET [10]	Yes	2.23	532	2.29	0.01	9.20e2
NV-LiM [9]	No	2.98	0.75	16.66	0.36	2.76e3
FeHM [11]	No	2.23	3.20	5.97	1.91	3.60e3
This work	No	4.46	0.32	9.17	1.06	3.84e3

[*] The negative voltage column indicates that the design uses the negative voltage or not. Non-using of negative voltage is better.

[#] Density indicates the number of bits stored in the same area. A higher density results in a higher utilization of the area.

Table II shows the density, power and latency comparisons in TCAM mode for alternative designs. Similar to the memory mode, the proposed FeMPIM in TCAM mode has a high memory density. The power indicates the average power overhead during a matching operation. We divide the comparison of power into cases of matched and unmatched. The power overhead of AiM and FeHM are

significantly different between matched and unmatched cases. In addition, because the state of the FeFET device needs to modify in TCAM mode, the match operations of AiM and FeHM are destructive. In contrast, the power overhead of NV-LiM, FeMAT, Re-FeMAT, and the proposed FeMPIM are similar and stable because they use additional CMOS transistors to reduce the leakage current. The proposed FeMPIM uses fewer CMOS transistors, avoids destructive match, and thus it has lower leakage current and moderate overhead in TCAM mode.

TABLE II
DENSITY, POWER AND LATENCY COMPARISONS IN TCAM MODE FOR ALTERNATIVE DESIGNS

Designs	Negative Voltage	Density (bit/um^2)	Power (nW)		Latency (ps)
			Matched	Unmatched	
AiM [6]	Yes	2.23	760.40	0.12	0.59
FeMAT [7]	Yes	1.49	0.62	0.52	1.38
Re-FeMAT [8]	Yes	1.49	1.10	0.98	29.09
Ref. [14] *	No	0.56	2.13	1.27	3.94
NV-LiM [9]	No	1.49	0.72	0.78	16.66
FeHM [11]	No	2.23	7.55	0.53	8.29
This work	No	2.23	1.16	1.19	8.74

* Ref. [14] only has TCAM function so that it is not compared in Table I. TI-FeFET has no TCAM function so that it is not compared in Table II.

V. CONCLUSIONS

In this paper, we have proposed a novel multifunctional FeFET-based cell. The proposed design can perform multiple functions, such as storage, logic calculation, and TCAM, in a single cell. In addition, the design provides independent read/write, preventing the use of negative voltages. Simulation results based on a calibrated FeFET model as well as a TSMC 65nm CMOS library has demonstrated the multiple correct functions and a moderate overhead for the proposed design compared with similar designs.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 61974001, 62274052 and 62174001, the Open Project of the State Key Laboratory of Computing Institute of Chinese Academy of Sciences under Grant CARCHA202101, the NSFC-JSPS Exchange Program under Grant 62111540164, the Outstanding Young Talent Support Program Key Project of Anhui Provincial Universities under Grant gxyqZD2022005, and the Distinguished Young Scholar Fund of Anhui Province under Grant 2022AH020014.

REFERENCES

[1] Sebastian, M. Gallo, R. Khaddam-Aljameh, et al, "Memory devices and applications for in-memory computing," Nature nanotechnology, vol. 15, no. 7, pp. 529-544, 2020.

[2] S. Han, X. Liu, H. Mao, et al, “EIE: Efficient Inference Engine on Compressed Deep Neural Network,” ACM/IEEE Annual International Symposium on Computer Architecture, pp. 243–254, 2016.

[3] 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.

[4] S. George, K. Ma, A. Aziz, et al, “Nonvolatile memory design based on ferroelectric FETs,” ACM/IEEE Design Automation Conference, pp. 1–6, 2016.

[5] D. Reis, M. Niemier, and X. S. Hu, “Computing in Memory with FeFETs,” International Symposium on Low Power Electronics and Design, pp. 1-6, 2018.

[6] D. Reis, A. Laguna, M. Niemier, et al, “Attention-in-Memory for Few-Shot Learning with Configurable Ferroelectric FET Arrays,” Asia and South Pacific Design Automation Conference, pp. 49–54, 2021.

[7] X. Zhang, X. Chen and Y. Han, “FeMAT: Exploring In-Memory Processing in Multifunctional FeFET-Based Memory Array,” IEEE International Conference on Computer Design, pp. 541-549, 2019.

[8] X. Zhang, R. Liu, T. Song, et al, “Re-FeMAT: A Reconfigurable Multifunctional FeFET-based Memory Architecture,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, early access, pp. 1-14, 2022.

[9] X. Yin, X. Chen, M. Niemier, et al, “Ferroelectric FETs-Based Nonvolatile Logic-in-Memory Circuits,” IEEE Transactions on Very Large-Scale Integration (VLSI) Systems, vol. 27, no. 1, pp. 159-172, 2019.

[10]E. Breyer, H. Mulaosmanovic, S. Slesazeck, et al, “Flexible Memory, Bit-Passing and Mixed Logic/Memory Operation of two Intercoupled FeFET Arrays,” IEEE International Symposium on Circuits and System, pp. 1-5, 2020.

[11]Marchand, I. O’Connor, M. Cantan, et al, “A FeFET-Based Hybrid Memory Accessible by Content and by Address,” IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, vol. 8, no. 1, pp. 19-26, 2022.

[12]Aziz, S. Ghosh, S. Datta, et al, “Physics-Based Circuit Compatible SPICE Model for Ferroelectric Transistors,” IEEE Electron Device Letters, vol. 37, no. 6, pp. 805–808, 2016.

[13]T. K. Song, "Landau-Khalatnikov simulations for ferroelectric switching in ferroelectric random access memory application," Journal of the Korean Physical Society, vol. 46, no. 1, pp. 1-5, 2005.

[14]K. Pagiamtzis and A. Sheikholeslami, “Content-addressable memory circuits and architectures: a tutorial and survey,” IEEE Journal of Solid-State Circuits, vol. 41, no. 3, pp. 712-727, 2006.