Non-volatile Inverter with 3D Cylindrical Metal-Ferroelectric-Metal Capacitor Realizing Digitized Voltage Output for Computing-in-Memory

K. Ota[1], J. Okuno[1], T. Yonai[1], R. Ono[1], Y. Shuto[1], M. Sakakibara[1], M. Lederer[2], P. Reinig[2], K. Seidel[2], R. Alcala[3], U. Schroeder[3], T. Mikolajick[3,4], A. Kato[1], and Y. Ueno[1]
[1]Sony Semiconductor Solutions Corporation, Atsugi, Kanagawa, 243-0014, Japan,
[2] Fraunhofer IPMS - Center Nanoelectronics Technologies, Germany, [3]NaMLab gGmbH, Germany, [4]TU Dresden, Germany

***Abstract*— A non-volatile $Hf_{0.5}Zr_{0.5}O_2$(HZO)-based Metal-Ferroelectric-Metal (MFM) inverter enabling AND/XNOR operations for Computation-in-Memory (CiM) is proposed. Owing to the symmetric threshold voltage (V_{th}) shift in the nFET and pFET of the MFM inverter, a digitized output after the multiply operation is successfully demonstrated for the first time with non-volatile memory, enabling the high precision CiM. In addition, the MFM inverter has the advantage of a large memory window (MW) because a large partial programming voltage is structurally applied to the MFM capacitor. Moreover, the MW and the endurance were intensively studied to clarify the key factors for reliable operation. Together with the process compatible FeRAM as a working memory, the MFM inverter has the potential to achieve low-power, high-density, and high precision non-volatile CiM systems.**

Keywords—Computing in Memory, Ferroelectric, $Hf_{0.5}Zr_{0.5}O_2$.

I. INTRODUCTION

Low-power CiM is a promising approach for edge AI applications. Among the various types of memories, FeFET has the potential for high-density and low-power operation owing to its scalability and non-volatility. In addition, HZO-based FeFET technology has attracted much attention thanks to the fast-switching, low voltage operation, and CMOS-compatibility. Nevertheless, many of the reported analog CiM with FeFET suffers from low precision caused by the variability in random polarization switching [1-2]. For example, as shown in Fig.1(a), charge-domain analog CiM, where the charges in capacitors are summed up in the computational line, is promising in terms of precision [2-5]. However, an accurate output voltage after a multiply operation into capacitors is necessary for a high precision so that V_{th} variability in FeFET should be the issue such as with the source-follower voltage output [2]. Another type of high-precision CiM is the digital CiM, which realizes massively parallel computation with no accuracy loss [6-7]. Digital output after multiply operation is essential for digital CiM, as shown in Fig.1(b), where the combination of SRAM and digital circuits was usually used. Thus, the digitized output after a multiply operation with non-volatile memory is promising for both analog and digital CiM. In this work, we propose a novel MFM inverter that enables digitized output for a multiply operation. Compared to SRAM-based CiM, a non-volatile MFM inverter has the potential for low power operation by eliminating standby power consumption. Furthermore, MFM inverter with only a two transistors (Tr.) and 1 MFM structure (2T1MFM) is superior for high density in contrast to 6 Tr. in SRAM.

In addition to realizing CiM with high precision, it is necessary to consider the CiM architecture. As shown in Fig.1(c), CiM with non-volatile memory (NVM) can have the functionality of storing the weight and multiply–accumulate (MAC) operation, while working RAM outside the CiM as input data is still necessary. Since MFM inverter can be fabricated with the same process as FeRAM, both CiM and working memory can be produced on the same chip with the same process. Non-volatility in CiM and working memory enables the fast restart without reloading weights from external ROM when turning on the power, and allow for power reduction by turning off during idle periods of operation.

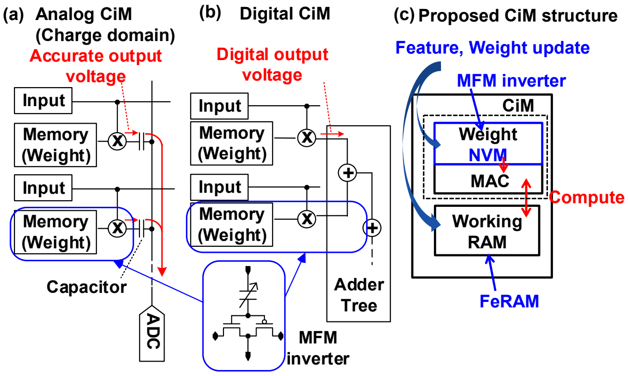

Fig. 1. (a) Charge domain analog CiM and (b) Digital CiM. Accurate or digital output voltage after multiply operation is necessary. (c) Proposed CiM structure with MFM inverter in CiM and with FeRAM as working RAM.

II. MULTIPLY-ACCUMULATE OPERATION WITH MFM INVERTER

Fig.2 shows the fabricated MFM inverter for CiM and 1T1C FeRAM for working RAM. The MFM inverter consists of a CMOS inverter with a 3D cylindrical MFM capacitor connected to the common gate. We have already developed the fabrication process of the 3D cylindrical capacitor for 1T1C FeRAM [8]. Although the 3D cylindrical capacitor on the common gate in the MFM inverter was shorter than that on the source or drain in FeRAM due to the height of the poly-Si gate, both capacitors in MFM inverter and FeRAM were successfully fabricated at the same time. Thus, both CiM and working RAM with the non-volatile memory can be produced on the same chip with the same process. It should be noted that

HZO-based non-volatile SRAM can also be available for working RAM since the same fabrication process is used [9]. Figs.3(a) and 3(b) show the measurement conditions for MFM inverter and I_{out}-V_{in} characteristics of the MFM inverter, respectively. The same amount of V_{th} shift was induced in both the nFET and pFET by applying the program voltage (V_{pgm}) to input gate of the MFM inverter (V_{in}). For comparison, we measured the inverter with MFMs on each gate of the nFET and pFET (2T2MFM), mimicking a conventional FeFET inverter (Fig.3(c)). Owing to the mismatch in the MFMs polarization, an asymmetric V_{th} shift between the nFET and pFET was observed. Moreover, a larger V_{th} shift in the MFM inverter compared with that in the 2T2MFM was achieved. A larger V_{th} shift in the MFM inverter can be explained by the larger partial program voltage applied to the MFM capacitor, which will be discussed in the next section. It should be noted that a symmetric and large V_{th} shift is the key to an accurate digitized output for multiplication operations which is crucial for CiM. Fig.4 shows the transfer characteristics of the MFM inverter. Transfer characteristics were successfully shifted by programming the MFM capacitor. Consequently, the V_{out} equals to V_{IA} in the erased state or to V_{IAb} in the written state.

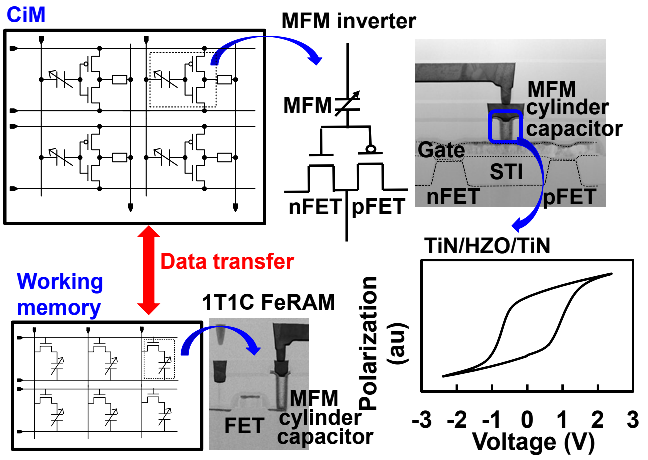

Fig. 2. CiM concepts and TEM images of FeRAM and proposed MFM inverter. The 3D cylindrical MFM capacitor exhibits polarization switching and can be fabricated with the same process as for FeRAM.

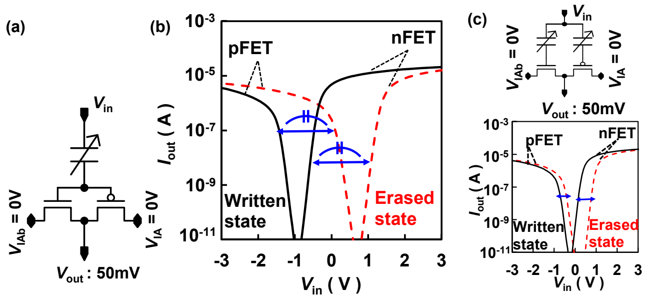

Fig. 3. (a) Schematic of the MFM inverter (b) Measured I_{out} - V_{in} characteristics are in the written and erased states. A large and same V_{th} shift in nFET and pFET was observed. (c) Schematic of the 2T2MFM inverter mimicking conventional FeFETs. Measured I_{out} - V_{in} characteristics show a smaller V_{th} shift than MFM inverter and an asymmetric V_{th} shift between the nFET and pFET due to the mismatch in MFM capacitors.

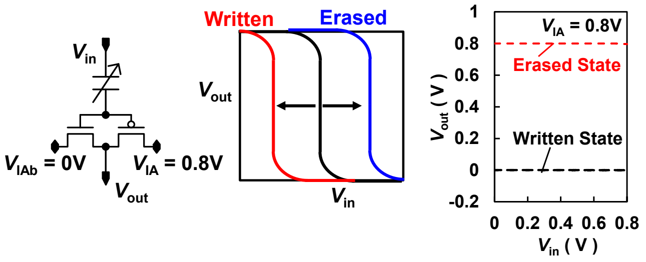

Fig. 4. Transfer characteristic of the MFM inverter was successfully shifted by writing or erasing the MFM capacitor. Consequently, V_{out} was equal to V_{IA}(= -0.8V) in the erased state and V_{IAb}(= 0V) in the written state independent of the input voltage V_{in} in this measurement range.

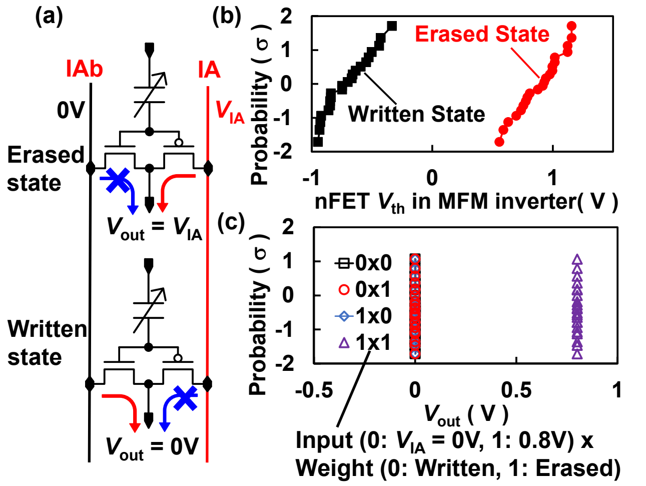

Fig. 5. (a) Schematic of two MFM inverters in the written (top) and erased (bottom) state indicating voltage transfer in the MFM inverter. (b) Measured V_{th} probability plots in the written and erased states. (c) Measured Vout probability plots for each input and weight. The output signal Vout corresponding to the results of an AND operation was successfully digitized.

An AND operation was demonstrated using an MFM inverter, as shown in Fig.5. In the written or erased state, the output voltage (V_{out}) is equal to the voltage applied to IA (V_{IA}) or IAb (V_{IAb}), as shown in Fig.5(a). By applying V_{IA} = 0 or 0.8V as the input and V_{IAb} = 0V, V_{out} is equal to the result of an AND operation. Despite the V_{th} variability in both the written and erased states caused by the random switching in the poly-crystalline ferroelectric layer, V_{out} was successfully digitized, as shown in Figs.5(b) and 5(c). Table 1 summarizes the voltage conditions for writing, erasing, and multiplication operations. In addition to the AND operation, an XNOR operation can be realized by applying V_{IAb} complementary to V_{IA}. MAC operations using the MFM inverter can be achieved as follows. For digital CiM, any of the multiplication results achieved in the MFM inverter can be selected by an additional transistor (Tr.) (Fig.6(a)). The subsequent adder trees then sum these. Compared to conventional digital CiM with SRAM [6,7], a non-volatile MFM inverter has the potential for low power by eliminating the leakage current at the standby. In addition, MFM inverter with only a 2T1MFM structure is superior for high density. For analog CiM, V_{out} from the MFM inverter resulting from the multiplication operation is converted to a charge by an additional capacitor, as shown in Fig.6(b). Subsequently, the accumulated charges from each capacitor are summed on a computation line connected to an ADC, causing the MAC operation to achieve the charge domain analog CiM, which is promising for high precision and low power [2-5].

Table 1. Voltage conditions for program, erase, and multiplication operations. In addition to the AND operation, an XNOR operation can be achieved by additionally applying V_{IAb} complementary to V_{IA}.

Operation	State IN*W	WL	Input (IA)	IAb	Weight (MFM)	Vout
Program	Write	V_{pgm}(7V)	0V	0V	-	-
	Erase	0V	V_{pgm}(7V)	V_{pgm}(7V)	-	-
AND Operation	0·0	V_{read}(0V)	0V	0V	Written	0V
	0·1				Erased	0V
	1·0		V_{IA}(0.8V)		Written	0V
	1·1				Erased	V_{IA}(0.8V)
XNOR Operation	0·0	V_{read}(0V)	0V	V_{IAb}(0.8V)	Written	V_{IA}(0.8V)
	0·1				Erased	0V
	1·0		V_{IA}(0.8V)	0V	Written	0V
	1·1				Erased	V_{IA}(0.8V)
Stand-by	-	V_{read}(0V)	0V	0V	0 or 1	0V

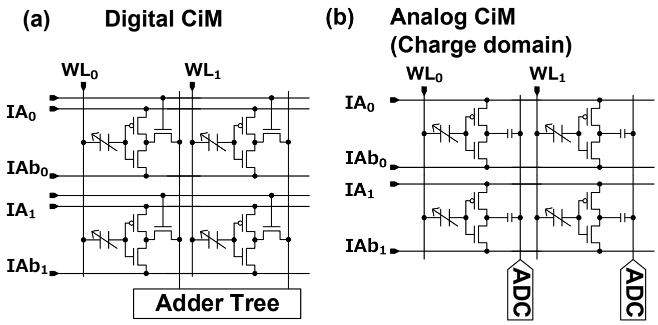

Fig. 6. (a) Schematic of a digital CiM using the MFM inverter. The MAC operation is available with an additional Tr. connected to each Vout and the following adder trees. (b) Schematic of analog CiM with MFM inverters. MAC operation can be achieved using an additional capacitor connected to each Vout and subsquent ADC.

III. Memory window improvement and cycling characteristic

To achieve accurate multiplication, a MW that is larger than the V_{th} variability should be maintained. For the MW study, we measured I_{d}-V_{g} characteristics in nFET with a planar MFM capacitor of 0.22×0.22μm^2 area, as shown in Fig.7. V_{th} was extracted from more than 25 devices after applying the same absolute value (V_{pgm}) of positive and negative voltage pulses to the top electrode of the MFM capacitor. Fig. 8(a) shows the V_{pgm} dependence of the median MW for different Tr. sizes. As V_{pgm} increased, the MW increased, suggesting that V_{pgm} as high as 8V is still insufficient for full polarization switching, particularly with a small-area Tr. In addition, the MW increased with Tr. size. Considering the series connection of the MFM capacitor (C_{MFM}) and gate oxide capacitance (C_{ox}), as shown in Fig.8(b), a larger partial program voltage can be applied to the MFM capacitor with a larger $C_{\mathrm{ox}}/C_{\mathrm{MFM}}$ ratio, resulting in a larger MW. Thus, larger Tr. in proportion to MFM size is beneficial for MW. It should be noted that an enhanced MW was achieved in the MFM inverter, as shown in Fig.3, owing to the large $C_{\mathrm{ox}}/C_{\mathrm{MFM}}$ ratio because the Tr. area is the sum of the nFET and pFET areas. As long as the $C_{\mathrm{ox}}/C_{\mathrm{MFM}}$ ratio is maintained, Tr. can be scaled down while maintaining the MW. Fig.9(a) shows the MW with different gate oxide thickness (T_{ox}). A larger MW was achieved with a thinner gate oxide, owing to the larger $C_{\mathrm{ox}}/C_{\mathrm{MFM}}$ ratio. Fig.9(b) shows the MW with different MFM thickness. The thicker HZO was also effective in increasing the voltage drop on the MFM capacitor by reducing C_{MFM}, whereas a thicker HZO resulted in a larger coercive voltage to induce polarization switching. As a result, a thicker HZO layer had a small impact on the MW.

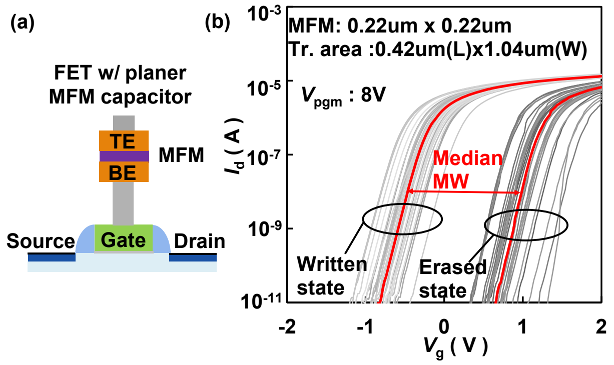

Fig. 7. (a) Schematic of the measured FET with a planar MFM capacitor. (b) I_{d}-V_{g} characteristics of the FET with the planar MFM capacitor. The size of the planar MFM capacitor was 0.22 x 0.22μm^2. More than 25 devices were measured for V_{th} extraction.

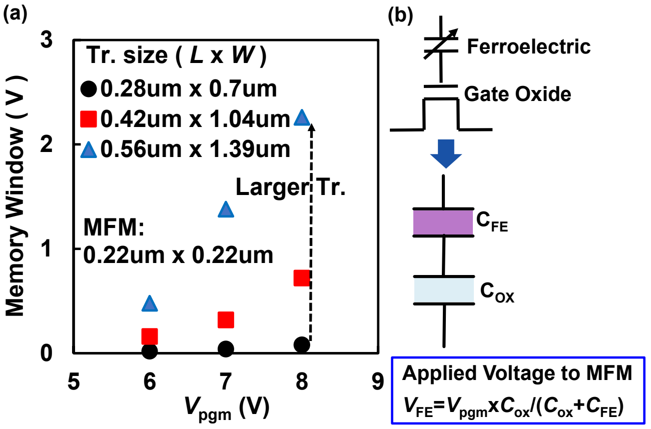

Fig. 8. (a) V_{pgm} dependence of the median MW for different Tr. sizes. The MW increases with the increase in V_{pgm} as well as the Tr. size. (b) Schematic of the FET with the MFM planar capacitor consisting of the series connection of an MFM capacitance (C_{MFM}) and gate oxide capacitance (C_{ox}).

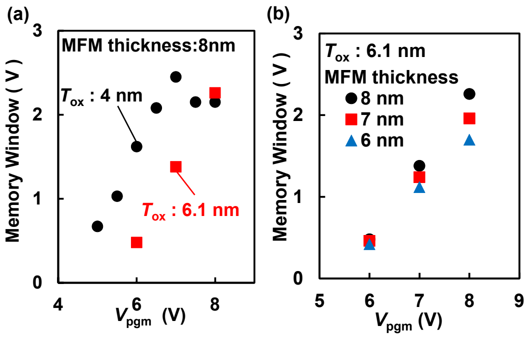

Fig. 9. (a) V_{pgm} dependence of the median MW for different thicknesses of the gate oxide (T_{ox}) of the Tr. A larger MW with the same V_{pgm} was achieved with thinner T_{ox}. (b) Median MW for different MFM thicknesses. A thicker HZO has a slight advantage for achieving a larger MW.

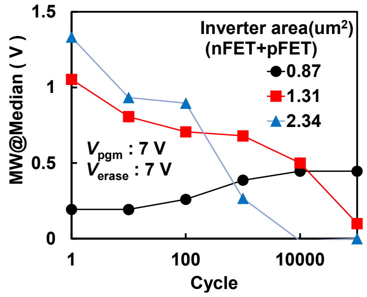

Fig. 10. Cycling characteristics of MFM inverters with different Tr. sizes. As the Tr. size increases, the MW decreases with a smaller number of cycles due to the degradation of the MFM capacitor.

Table 2. Comparison of CiM using various memory types in the literature. The proposed MFM inverter has the advantages of low power, good scaling, and high precision because of its non-volatility, a small foot-print consisting of only 2 Tr., and a digitized output after the multipication operation.

	This work	ISSCC 2024[6]	VLSI 2024[3]	IEDM 2021[10]	VLSI 2021[1]	VLSI 2021[2]
Memory Type	MFM inverter	SRAM	SRAM	IGZO DRAM	FeFET	FeFET
Non-Volatile	Yes	No	No	No	Yes	Yes
Unit cell	2Tr. + 1MFM	6Tr.	6Tr.	4Tr.+1C	1FeFET+ 1R	1FeFET
Cell area on Si Sub	>16F^2	>150F^2	>150F^2	-	>6F^2	>6F^2
MAC scheme	Digital/ Analog (charge)	Digital	Analog (charge)	Analog (charge)	Analog (current)	Analog (charge)
Digitized output	Yes	Yes	Yes	Yes	No	No

※F: Half pitch

The cycling characteristics of MFM inverters were measured with different sizes of Tr., as shown in Fig.10. Cycling above 10^5 was achieved in MFM inverters with smaller size transistors. It should be noted that, as Tr. size increased, the MW closed after a reduced number of cycles in contrast to MW. This result suggests that the degradation of the MFM capacitor was the reason for the observed cycling characteristics since the larger voltage was applied to the MFM capacitor with a larger Tr. sizes. Therefore, development of high quality MFM capacitor that is immune to cycling degradation is necessary in order to obtain better endurance.

IV. Conclusion

A novel MFM inverter for CiM is proposed. 3D capacitors in MFM inverter and FeRAM were fabricated using the same process so that CiM and working memory with the non-volatile memory can be produced on the same chip. Despite the V_{th} variability caused by MFM capacitor polarization, a digitized output was successfully achieved. As summarized in Table 2, the proposed MFM inverter has the advantages of low power, scaling area size, and high precision because of its non-volatility as well as a small foot-print consisting of only 2 Tr., and a digitized output after the multiplication operation.

References

[1] D. Saito, T. Kobayashi, H. Koga, N. Ronchi, K. Banerjee, Y. Shuto, J. Okuno, K. Konishi, L. Di Piazza, A. Mallik, J. Van Houdt, M. Tsukamoto, K. Ohkuri, T. Umebayashi, and T. Ezaki, “Analog In-memory Computing in FeFET-based 1T1R Array for Edge AI Applications” in Symp. VLSI Technol., 2021, JFS2.7.

[2] C. Matsui, K. Toprasertpong, S. Takagi, and K. Takeuchi, “Energy-Efficient Reliable HZO FeFET Computation-in-Memory with Local Multiply & Global Accumulate Array for Source-Follower & Charge-Sharing Voltage Sensing” in Symp. VLSI Technol., 2021, JFS2.8.

[3] H. Wang, R. Liu, R. Dorrance, D. Dasalukunte, N. Gowda, and B. Carlton, “A PVT Robust 8-Bit Signed Analog Compute-In-Memory Accelerator with Integrated Activation Functions for AI Applications” VLSI Circuit, 2024, C20.2.

[4] K. Yoshioka, “An 818-4094 TOPS/W Capacitor-Reconfigured CIM Macro for Unified Acceleration of CNNs and Transformers” in IEEE International Solid-State Circuits Conference (ISSCC), 2024, pp. 574–575.

[5] S. -E. Hsieh, C. -H. Wei, C. -X. Xue, H. -W. Lin, W. -H. Tu, E. -J. Chang, K. -T. Yang, P. -H. Chen, W. -N. Liao, L. L. Low, C. -D. Lee, A. -C. Lu, J. Liang, C. -C. Cheng, and T. -H. Kang, “A 70.85–86.27TOPS/W PVT-Insensitive 8b Word-Wise ACIM with Post-Processing Relaxation” in IEEE International Solid-State Circuits Conference (ISSCC), 2023, pp. 136–137.

[6] H. Fujiwara, H. Mori, W. -C. Zhao, K. Khare, C. -E. Lee, X. Peng, V. Joshi, C. -K. Chuang, S. -H. Hsu, T. Hashizume, T. Naganuma, C. -H. Tien, Y. -Y. Liu, Y. -C. Lai, C. -F. Lee, T. -L. Chou, K. Akarvardar, S. Adham, Y. Wang, Y. -D. Chih, Y. -H. Chen, H. -J. Liao, and T. -Y. J. Chang, “A 3nm 32.5 TOPS/W, 55.0 TOPS/mm^2 and 3.78 Mb/mm^2 Fully Digital Computing-in-Memory Supporting INT12 x INT12 with Parallel MAC Architecture” in IEEE International Solid-State Circuits Conference (ISSCC), 2024, pp. 572–573.

[7] W. Jiang, C. Caron, P. Avasare, M. Pauwels, M. Verhelst, and W. Dehaene, “HUNBN, a 1.77MB Digital In-Memory-Compute SoC for edge applications achieving 126 TOPs/W (4b) at macro level and 24 TOPs/W at SoC level” IEEE 50th European Solid-State Electronics Research Conference (ESSERC), pp. 137-140, 2024.

[8] J. Okuno, T, Kunihiro, T. Yonai, R. Ono, Y. Shuto, R. Alcala, M. Lederer, K, Seidel, T. Mikolajick, U. Schroeder, M. Tsukamoto, and T. Umebayashi, “A highly reliable 1.8 V 1 Mb Hf0.5Zr0.5O2-based 1T1C FeRAM Array with 3-D Capacitors” in IEDM Tech. Dig., 2023, 11-7.

[9] Y. Shuto, J. Okuno, T. Yonai, R. Ono, P. Reinig, M. Lederer, K. Seidel, R. Alcala, T. Mikolajick, U. Schroeder, T. Umebayashi, and K. Akiyama, “HZO-based Nonvolatile SRAM Array with 100% Bit Recall Yield and Sufficient Retention Time at 85°C” in Symp. VLSI Technol., 2024, T2.1.

[10] J. Liu, C. Sun, W. Tang, Z. Zheng, Y. Liu, H, Yang, C. Jiang, K. Ni, X. Gong, and X. Li, “Low-Power and Scalable Retention-Enhanced IGZO TFT eDRAM-Based Charge-Domain Computing”, in IEDM Tech. Dig., 2021, pp. 462–465.