Demonstration of Multiply-Accumulate Operation With 28 nm FeFET Crossbar Array

Sourav De, *Member, IEEE*, Franz Müller, Nellie Laleni, Maximilian Lederer, Yannick Raffel, Shaown Mojumder, Alptekin Vardar, Sukhrob Abdulazhanov, Tarek Ali, Stefan Dünkel, Sven Beyer, Konrad Seidel, and Thomas Kämpfe

Abstract—This letter reports a linear multiply-accumulate (MAC) operation conducted on a crossbar memory array based on 28nm high-k metal gate (HKMG) Complementary Metal Oxide Semiconductor (CMOS) and ferroelectric field effect transistor (FeFET). The fabrication is conducted at *GlobalFoundries* with their standard 28nm technology. The crossbar arrays show a 100% yield in MAC operation on a 300mm wafer. The arrays were divided into 8×8 segments. The FeFET crossbar arrays were fabricated with access transistors, current-limiter transistors and a current-mode analog-to-digital converter (ADC) on the same wafer. Finally, the data retention characteristics reveal excellent data retention characteristics up to 5×10^4 seconds, which makes this memory array suitable for carrying out MAC operations in inference engine applications.

Index Terms—Hafnium oxide, HfO_2, FeFET, ferroelectric memory, memory array.

I. INTRODUCTION

RECENT progress in researching hafnium oxide (HfO_2) based ferroelectric (Fe) materials has manifested its potential as emerging non-volatile memories. The compatibility of the FeFET process with the CMOS process and the demonstration of 28nm HKMG FeFET [1], [2], [3], [4], ferroelectric fin-field effect transistors (Fe-finFET) [5], [6], and ferroelectric thin film transistors (Fe-TFT) [7], [8] have paved the way for high-density FeFET based memory array. Amidst many developments, the device-to-device in the drain current (ΔI_d^{D2D}) of the FeFETs, especially in the low-threshold voltage (LVT) state, creates a significant bottleneck in the wafer-
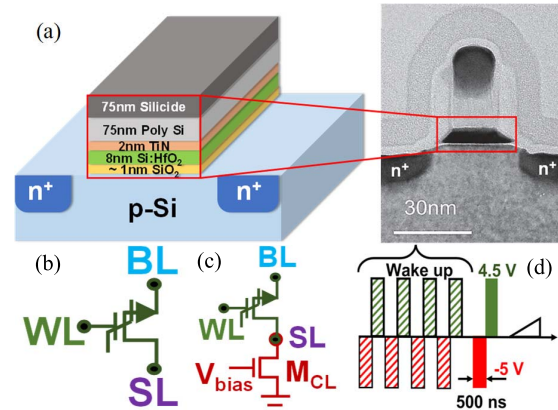
Fig. 1. (a) Schematic representation with TEM image of the FeFET cell demonstrates the details of the gate stack and device dimension. Schematic of (b). 1F and (c). 1F-1T cells. (d). The waveform applied at the W.L. of the cells for *WRITE* operation.

scale operation of the FeFET-based memory array [9], [10], [11], [12], [13], [14], [15]. The impact of ΔI_d^{D2D} increases with scaling and the memory array size [3], [16]. Although multi-level FeFETs have been demonstrated recently [15], [17], [18], they were limited to a standalone device, and in some cases, the reported channel current necessitates very complex and expensive read circuits. Previously, an effective reduction of ΔI_d^{D2D} on a 300mm wafer was reported by shunting an external resistor at the drain terminal of the FeFETs [16]. However, shunting an external resistor makes the macro-design complex.

In this work, we focus on FeFET-array operation on 300mm wafers. The impact of ΔI_d^{D2D} was controlled effectively by cascoding a current-limiting transistor with the source line (S.L.) of the FeFET crossbar arrays. The FeFET array, current limiting transistor, access transistors, and ADCs were fabricated on the same wafer. The memory array was divided into 8×8 segments to prevent further accumulation of errors over bit line current (I_{BL}) and to reduce the impact of voltage swing across word lines (W.L.), bit lines (B.L.), and select lines (S.L.). The retention characteristics demonstrate excellent stability of the I_{BL} over 5×10^4 seconds, which proves the aptness of the memory array for being deployed for MAC operation in an inference engine. The B.L.s of the segments were connected to the input of the ADC. Although the ADC's nonlinearity was reduced by optimizing the reference voltage, the ADC output differed from the ideal value due to the capacitive coupling from B.L.s [16].
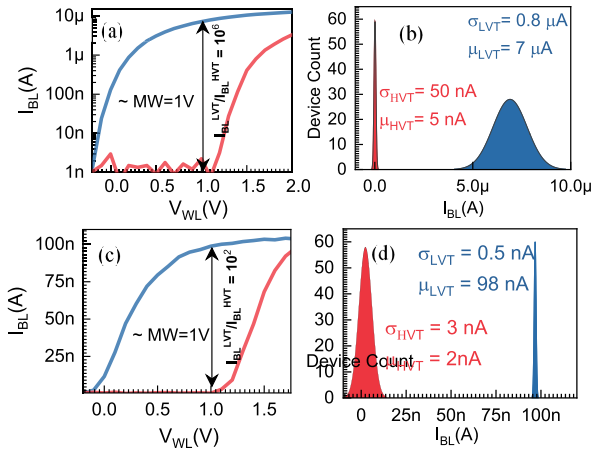
Fig. 2.   (a) The transfer characteristics of a 1F cell. (b). The distribution of I_{BL}^{LVT} and I_{BL}^{HVT} shows high variation in I_{BL}^{LVT}, which overshadows the advantage of high ratio of I_{BL}^{LVT} and I_{BL}^{HVT}. (c). The transfer characteristics demonstrates reduction in the ratio of I_{BL}^{LVT} and I_{BL}^{HVT} for 1F-1T cells. (d). However, the variation in the I_{BL}^{LVT} has also been improved significantly, which is essential for carrying out MAC operations.

II. EXPERIMENTS AND RESULTS

A. Device Characterization

The crossbar arrays and memory cell test structures were fabricated on 300mm wafers at *GlobalFoundries* with the 28nm HKMG technology [1], [2], [3]. 8nm thick silicon doped HfO_2 (Si:HfO_2) was used as the ferroelectric layer with a 1nm interfacial layer of silicon dioxide (SiO_2). Fig.1(a) shows a single cell's schematic representation with the details of the metal-ferroelectric-insulator-semiconductor (MFIS) gate stack and corresponding transmission electron microscopic (TEM) image, denoting the device dimensions. The experiments began by characterizing FeFET cells in two splits. The first one, shown in fig.1(b) is standalone FeFETs (1F), and the second one, shown in fig.1(c), is FeFETs connected with a current-limiting transistor (M_{CL}) fabricated on the same wafer (1F-1T). 1F and 1F-1T were programmed and erased by 500ns pulses of amplitudes 4.5V and -5V at the gate terminal or word line (W.L.), and they were subjected to 50 wake-up cycling. The wake-up cycling, endurance ($\sim10^5$ cycles) and retention of 1F structures are reported in the previous work, depicting the characteristics of FeFET devices without current limiter [1], [2], [3]. The wake-up cycling was unnecessary for 1F-1T structures, as the current limiter prevented drain current drift. However, the retention and endurance characteristics remained the same for both cases. The B.L. and S.L. were held at 0V during the *WRITE* operation. After programming, the devices were allowed to de-trap for a minimum of 300ms. A details analysis of the *WRITE*-to-*READ* delay on similar devices is shown elsewhere [19]. *READ* operation was conducted by a voltage ramp with a step size of 100mV at W.L. while keeping bit line (B.L.), source line (S.L.) and bulk biased at 100mV, 0V and 0V, respectively. *READ* operation takes $\sim$1ms to complete. The MCL's gate bias voltage (Vbias) in 1F-1T cell was 1.8V and 0.3V during *WRITE* and *READ* operations, respectively.

Fig.2 shows the transfer characteristics and the variability statistics of 1F and 1F-1T cells. We observe a significant

decrease in the drain/BL current ratio for LVT-state (I_{BL}^{LVT}) to HVT-state (I_{BL}^{HVT}) for 1F-1T cells (fig.2 (c)), compared to 1F cells (fig.2 (a)). The advantage of high I_{BL}^{LVT} ($\sim7\mu$A) to I_{BL}^{HVT} ($\sim$5nA) ratio in 1F cells was overshadowed by the high variation in I_{BL}^{LVT}, which is 0.8μA (fig.2b).

Although the mean value of I_{BL}^{LVT} in 1F-1T cells was $\sim$100nA (fig.2c), variation was less than 3nA (fig.2d). This low variation makes the 1F-1T cells perfect for MAC operation, which is highly susceptible to the variation of the I_{BL}^{LVT} and gets significantly affected by an increasing number of activated memory cells.

B. Array Characterization

The characterization of AND-type arrays followed the characterization of single devices. The layout, die photo, and schematic representation of the 8×8 segments is shown in fig. 3(a). We have used only one current limiting transistor, denoted by M_{CL}, attached to the S.L. of the segment for array-level operation. The gate terminals of FeFETs are connected row-wise in the W.L. The drains and sources are connected column-wise in B.L. and S.L. The FeFETs in the memory array (green colour) are fabricated with other non-FeFET transistors in the same 300 mm wafer at *GlobalFoundries* with their HKMG-28nm technology.

The arrays are characterized using a PXI-Express system from National Instruments. The contacts of the memory array were controlled by the pin parametric measurement unit (PPMU) of NI PXIe-6570 and the source measure unit (SMU) of NI PXIe-4143. The W.L.s were selected by a custom switch matrix, which connects the contact pads of the array via a probe card. The *WRITE* operation was conducted row-wise with a 4.5V pulse of 500ns. The memory array was block-wise erased by applying a 5V pulse of 40μs in bulk while applying 0V at W.L., B.L. and S.L. Fig. 3(b) show the voltage scheme used for *WRITE* operation on a row of the crossbar while keeping the other rows inhibited. Fig. 3(c) shows the biasing scheme for cell-wise *READ* operation. The S.L. was connected to the drain terminal of M_{CL}. All the B.L.s were biased to 0.1V by applying the same at V_{BL}^{PRG}. The column of the target cell accessed the S.L. and B.L. via N_{SL} and N_{BL} transistors, which were turned on by applying 1.8V at V_{SL} and V_{BL}. The current output from N_{BL} was the connection to the input of the ADCs. The inhibit transistors of the same columns were deactivated by applying 0V at V_{INH} during the active read operation. Inhibit mode on B.L. and S.L. was used to deactivate the undesired column during a *READ* operation. The inhibit operation on B.L. and S.L. was facilitated by turning on M_{BL} and M_{SL} by applying 1.8V at V_{INH} and turning off N_{BL} and N_{SL} by applying 0V at V_{BL} and V_{SL}. Further, the inhibit operation on W.L. by applying -0.3V enabled cell-wise *READ* operation. The operating voltages for the array operation are shown in Table I.

Fig.4a demonstrates the MAC operation from a single column of the crossbar segment. Initially, the complete array segment was block-wise erased. The erase operation was followed by row-wise *WRITE* and cell-wise *READ* operations. The leakage current from the bulk with all transistors in the erased state was lower than 1nA. We can also observe negligible leakage current during the MAC operation as well. Fig.4b
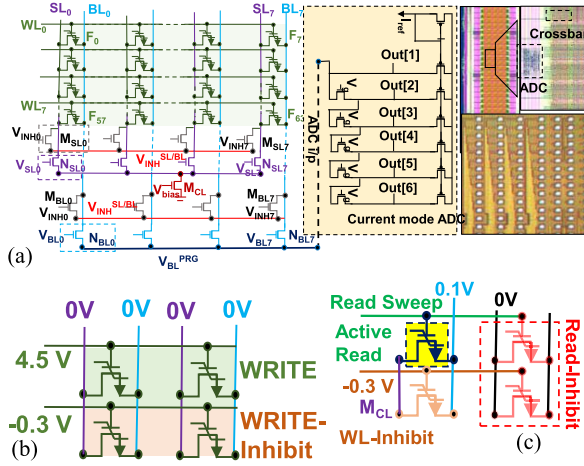
Fig. 3. (a). Schematic representation of the ADC embedded memory array with layout and optical image. (b). The operating voltages along BL, SL and WL for conducting the *WRITE* operation. (c). The applied voltages along BL, SL and WL for bit-wise READ operation while keeping other cells inactive.

TABLE I
OPERATING VOLTAGES

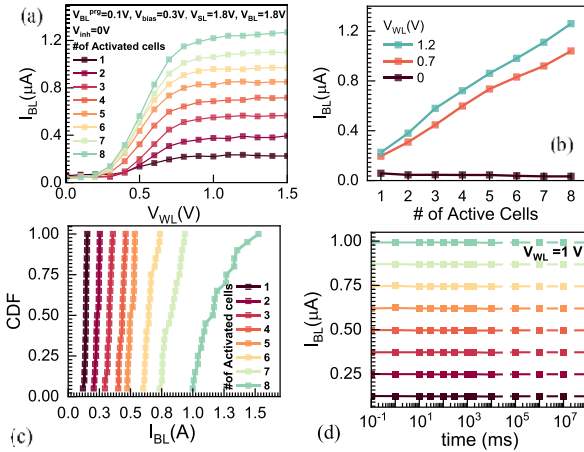|	V_{WL}	V_{BL}	V_{SL}	V_{INH}	$V_{INH}^{SL/BL}$	V_{bias}
Write	4.5V	1.8V	1.8V	0V	Floating	1.8V
Read	0 to1.5V	1.8V	1.8V	0V	Floating	0.3V
Inhibit W.L.	-0.3V	Not Relevant				
Inhibit BL/SL	Not Used	0V	0V	1.8V	0V	0.3V

Fig. 4. (a) MAC Operation shows I_{BL} measured w.r.t V_{WL} from an 8×8 segment of the crossbar array. (b) The current-limiter embedded at the end of each tile generates linear MAC operation with high precision. (c) The CDF plot of I_{BL} variation shows stable array level operation. (d). The data retention characteristics shows stable data retention measured up 5×10^4 seconds and extrapolated retention up to 10 years.

shows highly linear and V_{WL}-independent MAC operation. For statistical modelling, the MAC operation was performed over 20 segments across 300mm wafers. Fig.4c shows stable MAC operation over 20 different segments from the crossbar array. Finally, the retention characteristics of the MAC operation were measured. Finally, the retention characteristics have been measured for each activation level at room temperature. Fig.4d shows stable data retention capability up to 5×10^4 seconds for each activation level.
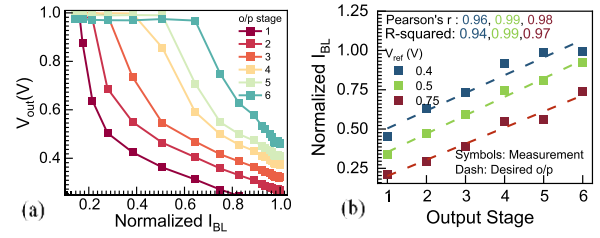

Fig. 5. (a) Output Voltage vs Input current of thermometer code ADC. (b). Linearity in ADC plays important role in neuromorphic applications. Comparison between actual o/p from the desired one shows only a small deviation.

TABLE II
BENCHMARKING

	[22] VLSI, IBM	[21] Nature Samsung	[19]Nature	This Work
Device	PCM	MRAM	RRAM	FeFET
Technology	14 nm	28 nm	130 nm	28 nm
Write Latency	130 ns	NA	1µs	500 ns
Sensing Mode	Voltage	Time domain	Voltage	Current
ADC Technology	Shift and Add CCO	TDC Readout	Voltage Mode	Current mode 3b
WRITE style	Close-loop	NA	Close-loop	One shot
Write Voltage	>2.5V	1.5V	3.1V	4.5V
Cell Type	8T-4R	2T-2R	1T-1R	1T
Periphery	Yes	Yes	Yes	Yes

C. Analog to Digital Conversion

The output of the B.L.s was connected to the thermometer code current-mode ADC. The ADCs consist of stacked PMOS-based current mirrors, and the NMOS transistors forward the measured current to the next stage. During *READ* and *WRITE* operations, the V_g in the ADC was biased at 0.6V and 0.312V, respectively. The ADC's reference current (Iref) was set at 100nA. While the input current is smaller than 100nA, the corresponding current mirror maintains high output voltage. The output drops sharply upon the increase of the input current beyond I_{ref}. Fig.5a represents the transfer characteristics of the ADC for six different output stages. The input of the ADC is the I_{BL} of the FeFET array. The linearity of ADC operation has been shown in Fig.5b, where it is shown that a proper choice of reference-bias voltage minimizes the nonlinearity in ADC operation. Pearson's coefficient corroborates that the final measured output from the ADCs marginally differs from the desired values.

III. CONCLUSION

MAC operation on 28nm HKMG FeFET-based memory array with ADC has been demonstrated. Adding a current-limiting transistor at the end of S.L. reduces the impact drain current variation. The output of the MAC operation shows high linearity and data retention capability of up to 5×10^4 seconds for each level of cell activation. Further, the linearity of the ADC output was increased by optimizing the reference voltage value of the ADC. This experimental demonstration of the ADC-embedded FeFET array paves the way for FeFET-based CIM-macro. Finally, we have benchmarked the performance of this memory array with other state-of-art memory arrays with different emerging-nonvolatile memories.

REFERENCES

[1] M. Trentzsch, S. Flachowsky, R. Richter, J. Paul, B. Reimer, D. Utess, S. Jansen, H. Mulaosmanovic, S. Müller, S. Slesazeck, J. Ocker, M. Noack, J. Müller, P. Polakowski, J. Schreiter, S. Beyer, T. Mikolajick, and B. Rice, "A 28 nm HKMG super low power embedded NVM technology based on ferroelectric FETs," in *IEDM Tech. Dig.*, Dec. 2016, p. 11, doi: 10.1109/IEDM.2016.7838397.

[2] J. Müller, E. Yurchuk, T. Schlosser, J. Paul, R. Hoffmann, S. Müller, D. Martin, S. Slesazeck, P. Polakowski, J. Sundqvist, M. Czernohorsky, K. Seidel, P. Kucher, R. Boschke, M. Trentzsch, K. Gebauer, U. Schroder, and T. Mikolajick, "Ferroelectricity in HfO$_2$ enables nonvolatile data storage in 28 nm HKMG," in *Proc. Symp. VLSI Technol. (VLSIT)*, Jun. 2012, pp. 25–26, doi: 10.1109/VLSIT.2012.6242443.

[3] S. Beyer, S. Dunkel, M. Trentzsch, J. Müller, A. Hellmich, D. Utess, J. Paul, D. Kleimaier, J. Pellerin, S. Müller, J. Ocker, A. Benoist, H. Zhou, M. Mennenga, M. Schuster, F. Tassan, M. Noack, A. Pourkeramati, F. Müller, M. Lederer, T. Ali, R. Hoffmann, T. Kampfe, K. Seidel, H. Mulaosmanovic, E. T. Breyer, T. Mikolajick, and S. Slesazeck, "FeFET: A versatile CMOS compatible device with game-changing potential," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2020, pp. 1–4, doi: 10.1109/IMW48823.2020.9108150.

[4] S. De, F. Müller, H.-H. Le, M. Lederer, Y. Raffel, T. Ali, D. Lu, and T. Kämpfe, "READ-optimized 28 nm HKMG multibit FeFET synapses for inference-engine applications," *IEEE J. Electron Devices Soc.*, vol. 10, pp. 637–641, 2022, doi: 10.1109/JEDS.2022.3195119.

[5] S. De, D. D. Lu, H.-H. Le, S. Mazumder, Y.-J. Lee, W.-C. Tseng, B.-H. Qiu, M. A. Baig, P. J. Sung, C.-J. Su, C.-T. Wu, W.-F. Wu, W.-K. Yeh, Y.-H. Wang, and Yeong-Her, "Ultra-low power robust 3bit/cell Hf$_{0.5}$Zr$_{0.5}$O$_2$ ferroelectric FinFET with high endurance for advanced computing-in-memory technology," in *Proc. Symp. VLSI Technol.*, 2021, pp. 1–2.

[6] S. De, M. A. Baig, B.-H. Qiu, H.-H. Le, Y.-J. Lee, and D. Lu, "Neuromorphic computing with Fe-FinFETs in the presence of variation," in *Proc. Int. Symp. VLSI Technol., Syst. Appl. (VLSI-TSA)*, Apr. 2022, pp. 1–2, doi: 10.1109/VLSI-TSA54299.2022.9771015.

[7] C. Sun, K. Han, S. Samanta, Q. Kong, J. Zhang, H. Xu, X. Wang, A. Kumar, C. Wang, Z. Zheng, X. Yin, K. Ni, and X. Gong, "Highly scaled InGaZnO ferroelectric field-effect transistors and ternary content-addressable memory," *IEEE Trans. Electron Devices*, vol. 69, no. 9, pp. 5262–5269, Sep. 2022, doi: 10.1109/TED.2022.3188582.

[8] S. Dutta, H. Ye, A. A. Khandker, S. G. Kirtania, A. Khanna, K. Ni, and S. Datta, "Logic compatible high-performance ferroelectric transistor memory," *IEEE Electron Device Lett.*, vol. 43, no. 3, pp. 382–385, Mar. 2022, doi: 10.1109/LED.2022.3148669.

[9] E. Yurchuk, J. Müller, J. Paul, T. Schlosser, D. Martin, R. Hoffmann, S. Müeller, S. Slesazeck, U. Schroeder, R. Boschke, R. van Bentum, and T. Mikolajick, "Impact of scaling on the performance of HfO$_2$-based ferroelectric field effect transistors," *IEEE Trans. Electron Devices*, vol. 61, no. 11, pp. 3699–3706, Nov. 2014, doi: 10.1109/TED.2014.2354833.

[10] K. Ni, W. Chakraborty, J. Smith, B. Grisafe, and S. Datta, "Fundamental understanding and control of device-to-device variation in deeply scaled ferroelectric FETs," in *Proc. Symp. VLSI Technol.*, Jun. 2019, pp. T40–T41, doi: 10.23919/VLSIT.2019.8776497.

[11] G. Choe and S. Yu, "Variability study of ferroelectric field-effect transistors towards 7 nm technology node," *IEEE J. Electron Devices Soc.*, vol. 9, pp. 1131–1136, 2021, doi: 10.1109/JEDS.2021.3100290.

[12] M. Lederer, F. Müller, A. Varanasi, R. Olivo, K. Mertens, D. Lehninger, Y. Raffel, R. Hoffmann, T. Ali, K. Seidel, T. Kämpfe, and L. M. Eng, "Influence of microstructure on the variability and current percolation paths in ferroelectric hafnium oxide based neuromorphic FeFET synapses," in *Proc. Silicon Nanoelectronics Workshop (SNW)*, 2021, pp. 1–2.

[13] Y.-S. Liu and P. Su, "Superior immunity to trapped-charge induced variability in 2D FeFET NVMs," in *Proc. Silicon Nanoelectronics Workshop (SNW)*, 2021, pp. 1–2.

[14] B. Zeng, M. Liao, Q. Peng, W. Xiao, J. Liao, S. Zheng, and Y. Zhou, "2-bit/cell operation of Hf$_{0.5}$Zr$_{0.5}$O$_2$ based FeFET memory devices for NAND applications," *IEEE J. Electron Devices Soc.*, vol. 7, pp. 551–556, 2019, doi: 10.1109/JEDS.2019.2913426.

[15] S. De, M. A. Baig, F. Müller, H.-H. Le, M. Lederer, T. Kämpfe, T. Ali, P.-J. Sung, C.-J. Su, Y.-J. Lee, and D. D. Lu, "Random and systematic variation in nanoscale Hf$_{0.5}$Zr$_{0.5}$O$_2$ ferroelectric FinFETs: Physical origin and neuromorphic circuit implications," *Frontiers Nanotechnol.*, vol. 3, Jan. 2022, Art. no. 826232, doi: 10.3389/fnano.2021.826232.

[16] T. Soliman, F. Müller, T. Kirchner, T. Hoffmann, H. Ganem, E. Karimov, T. Ali, M. Lederer, C. Sudarshan, T. Kämpfe, A. Guntoro, and N. Wehn, "Ultra-low power flexible precision FeFET based analog in-memory computing," in *IEDM Tech. Dig.*, 2020, p. 29, doi: 10.1109/IEDM13553.2020.9372124.

[17] K. A. Aabrar, S. G. Kirtania, A. Lu, A. Khanna, W. Chakraborty, M. S. Jose, S. Yu, and S. Datta, "A thousand state superlattice(SL) FEFET analog weight cell," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technology Circuits)*, Jun. 2022, pp. 242–243, doi: 10.1109/VLSITechnologyandCir46769.2022.9830333.

[18] K. A. Aabrar, J. Gomez, S. G. Kirtania, M. S. Jose, Y. Luo, P. G. Ravikumar, P. V. Ravindran, H. Ye, S. Banerjee, S. Dutta, A. I. Khan, S. Yu, and S. Datta, "BEOL compatible superlattice FerroFET-based high precision analog weight cell with superior linearity and symmetry," in *IEDM Tech. Dig.*, Dec. 2021, p. 19, doi: 10.1109/IEDM19574.2021.9720713.

[19] W. Wan, R. Kubendran, C. Schaefer, S. B. Eryilmaz, W. Zhang, D. Wu, S. Deiss, P. Raina, H. Qian, B. Gao, S. Joshi, H. Wu, H.-S.-P. Wong, and G. Cauwenberghs, "A compute-in-memory chip based on resistive random-access memory," *Nature*, vol. 608, no. 7923, pp. 504–512, Aug. 2022, doi: 10.1038/s41586-022-04992-8.

[20] D. Kleimaier, H. Mulaosmanovic, S. Dunkel, S. Beyer, S. Soss, S. Slesazeck, and T. Mikolajick, "Demonstration of a p-type ferroelectric FET with immediate read-after-write capability," *IEEE Electron Device Lett.*, vol. 42, no. 12, pp. 1774–1777, Dec. 2021, doi: 10.1109/LED.2021.3118645.

[21] S. Jung, H. Lee, S. Myung, H. Kim, S. K. Yoon, S.-W. Kwon, Y. Ju, M. Kim, W. Yi, S. Han, B. Kwon, B. Seo, K. Lee, G.-H. Koh, K. Lee, Y. Song, C. Choi, D. Ham, and S. J. Kim, "A crossbar array of magnetoresistive memory devices for in-memory computing," *Nature*, vol. 601, no. 7892, pp. 211–216, Jan. 2022, doi: 10.1038/s41586-021-04196-6.

[22] R. Khaddam-Aljameh, M. Stanisavljevic, J. Fornt Mas, G. Karunaratne, M. Braendli, F. Liu, A. Singh, S. M. Müller, U. Egger, A. Petropoulos, T. Antonakopoulos, K. Brew, S. Choi, I. Ok, F. L. Lie, N. Saulnier, V. Chan, I. Ahsan, V. Narayanan, S. R. Nandakumar, M. Le Gallo, P. A. Francese, A. Sebastian, and E. Eleftheriou, "HERMES core— A 14 nm CMOS and PCM-based in-memory compute core using an array of 300ps/LSB linearized CCO-based ADCs and local digital processing," in *Proc. Symp. VLSI Circuits*, Jun. 2021, pp. 1–2, doi: 10.23919/VLSICircuits52068.2021.9492362.