A 129.83 TOPS/W Area Efficient Digital SOT/STT MRAM-Based Computing-In-Memory for Advanced Edge AI Chips

Lu Lu, Aarthy Mani, and Anh Tuan Do
Institute of Microelectronics, Agency for Science, Technology and Research, Singapore
Lu_Lu@ime.a-star.edu.sg and doat@ime.a-star.edu.sg

***Abstract*—This paper proposes a spin-orbit torque (SOT) magnetoresistive random access memory (MRAM)-based digital computing in memory (CIM) structure for advanced CIM edge AI chips. To avoid frequent data reloading and reduce the area overhead caused by the large transistor in the write path, 11 transistors and 4 shared heavy metal (HM) SOT MRAM bitcell is proposed. It contains 4b weight in a single cell in a compact area able to hold a large capacity with reduced latency. Compared to the previous SRAM+NOR digital CIM design, the SOT/STT MRAM CIM designs occupy only 40% and 30% bitcell area respectively. Additionally, SOT MRAM has a 1.16% leakage current of SRAM at the TT corner at room temperature. The proposed design is verified using statistic simulations in 28nm technology. It achieves 129.83 TOPS/W at 1b/1b/8b precision.**

Keywords—Computing in memory, SOT, STT, MRAM, neural network.

I. Introduction

As cloud computing is reaching the peak of its run, edge computing is becoming the future of data transformation. By moving the computing power to the edge of the internet of things (IoT) devices and sensors, it benefits from lower latency, increased data bandwidth, and reduced power consumption on heavy data transportation [1]. However, the capacity of computing in memory (CIM) restricts the performance of advanced edge-based chips. Large-capacity memory devices can reduce the reload frequency to move the data from the main memory (e.g., DRAM and large on-chip cache) to CIM and avoid the extra memory buffer to keep the weight (Fig. 1) [2]. It comes with lower energy efficiency, less sensing margin, and degraded precision. To improve computing accuracy, digital CIM is a preferred option because of its robustness to process variations and environmental noise. Bitcell of digital SRAM-CIM usually comprises an SRAM cell with a logic gate, which is volatile storage and costs a significant area overhead and large leakage current.

Magnetoresistive random access memory (MRAM) receives considerable attention as a promising future memory due to its non-volatility, fast speed, CMOS compatibility, and high endurance [3]. Each memory cell uses a magnetic tunnel junction (MTJ) device to store data. Spin-transfer-torque (STT) MRAM is expected in a wide range of applications with its high scalability (1 transistor and 1 MTJ) (Fig. 2(a)). However, lowering the critical switching current leads to read disturb failure, and the shared path also reduces device endurance. To address these limitations, spin-orbit torque (SOT)-based MTJ switching

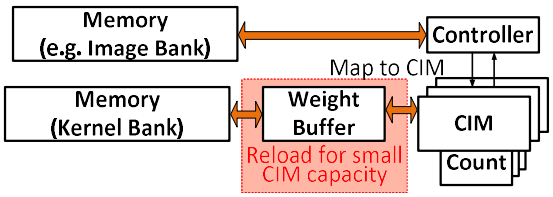

Fig. 1. Challenge of the CIM

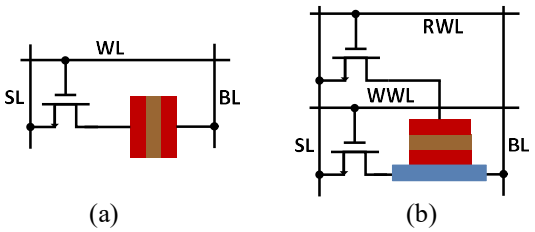

Fig. 2. MRAM structure for (a) STT, and (b) SOT

mechanism has been introduced. By decoupling the read-and-write path with heavy metal (HM), SOT MRAM offers better stability, shorter switching time, lower write current, and larger endurance. On the other side, the 3-terminal concept increases the number of transistors in a single cell and thus increases layout area (2 transistors, 1 MTJ, and 1 HM) as shown in Fig. 2(b). To compensate for the area overhead, recent studies proposed several area-efficient SOT MRAM structures [4-6]. Karim et al. stacked two MTJs and two HMs head-to-head and used three transistors to select the read/write paths [4]. In [5], Yeo et al. shared BLs between two adjacent cells, reducing the number of metal lines within one column to scale down the area. Chao et al. introduced a shared HM between multiple MTJs, by regulating the width of the HM to control the current density [6].

The nature of resistive device-based memory benefits the integration of analog CIM structures. The multiply-and-accumulate (MAC) operation is correlated to the resistance value in both voltage and current modes. For instance, MRAM computes matrix-vector multiplication (MVM) with its reconfigurable logic while SRAM needs an extra assist circuit. The challenges of implementing MRAM in the CIM structure are 1) limited $R_{parallel}$ (R_p) and $R_{antiparallel}$ (R_{ap}) ratio (around 2 or less), 2) the large current caused by the low resistance value (lower than 20k) causing power inefficiency, 3) large write transistor to support the switching current. To fit the MRAM into the CIM role, one previous work built multilevel cells and increased the resistance value by stacking three levels of SOT and STT MTJs [7]. Geng's team used the model compression techniques to reduce the power consumption in architecture from the software level

[8]. Chao et al. adopted a self-reference method to read out data with small resistance differences, which senses the data in sequence and then compares the results [9]. These techniques are still working on analog CIM with sensing amplifiers to read data out.

This paper proposes a SOT/STT MRAM-based digital CIM cell. To address the above issues, we first adopt shared HM between MTJs to reduce the area and expand the weight capacity. Then, by regulating the current on the read path we reduce power consumption. Finally, by converting the analog resistance weight into digital output data, the proposed MRAM-based digital CIM does not lose accuracy by increasing the capacity of memory.

II. Proposed Digital MRAM-Based CIM Bitcell

SOT MRAM provides long endurance by sacrificing the compact area. To provide sufficient switching current the transistor size of the write path is usually larger than the one on the read path. The key point in building an area efficient bitcell is reducing the transistor size on the write path. We adopt a method that fabricates multiple MTJs on the same HM with non-uniform widths [6] by placing two larger transistors (M0 and M1) on both BL/SL sides to control the write path. As shown in Fig. 3(a), the proposed digital MRAM-CIM bitcell contains 11 transistors, 4 MTJs, and 1 HM (11T4-in-1M). The four MTJs (MTJ0 to MTJ3) have individual transistors (M2 to M5) to select the weight 0-3. 1 transmission gate, 1 inverter, and 1 PMOS are shared with 4 MTJs to convert the resistance difference into digital data. Only one read transistor can be activated at one time. To regulate the current during the computing operation, the length of the PMOS M8 is increased to 2.6 times of the minimum length. The transmission gate employs the minimum width and enlarged length to reduce the leakage current. When the input is 0, all the signals are connected to GND. Only M7 and M8 are on to force the OUT to be 0. When the input is 1, RY will be tied high and RYB is asserted to GND, the voltage on BLI (V_{BLI}) will be formed to a high or low range depending on the MTJ resistance (R_p represents weight 1 and R_{ap} represents weight 0). The V_{BLI} will connect with V_{BLO} through the transmission gate M6 and M7 (V_{BLI} = V_{BLO}). The size of the inverter is optimized to obtain a proper switching threshold voltage and can distinguish between the high and low V_{BLO} with R_p and R_{ap} (Section II Paragraph 4). The computing output in a cell is shown in the table at the top-right of Fig. 3(a).

As mentioned in section I, the SOT MRAM has better performance but a larger area. For applications requiring a more compact area with relaxed endurance and speed criteria, STT MRAM is a more suitable candidate to replace SOT MRAM. As shown in Fig. 3(b), 4 STT MRAMs are controlled by their access transistors M0-M3. The write and read path are shared. The PMOS M6 is connected to BL instead of VDD and controlled by signal WL. The M6 is on in both write and read operations. The M6, transmission gate, and inverter are shared with 4 STT MRAM with the same mechanism as the SOT MRAM-based digital CIM bitcell.

Fig. 3(c) shows the layout of the SOT MRAM in 28nm CMOS technology. Since the aggressive rule is not applied, the layout size of a single bitcell is 1.77 µm X 0.565 µm. The layout is rotated due to the requirement of vertical poly. The control signals RY/RYB, four RWLs, and one WWL are running horizontally, and the SL and BL are placed vertically. The two red-dotted rectangles highlight the regions of M0 and M1, the size of the write transistor is X4 of the minimum widths. In the layout of STT MRAM bitcell, the M0 and M1 can be omitted, and the overall size of the layout is reduced to 1.34 µm X 0.565 µm. Fig. 3(d) shows the area comparison of the proposed MRAM-based CIM bitcell with state-of-the-art. [10] and [11] are SRAM-based digital CIM structures of (6T SRAM + 12T XNOR) and (6T SRAM + 4T NOR), the area depends on the logic gate for bitwise multiplies. As the equivalent area for the same weight storage capacity, the proposed 4-in-1 SOT MRAM and STT MRAM CIM structures obtain 319 F^2 and 241.4 F^2, respectively, which are 40% and 30% percent of the baseline 10T SRAM + NOR. RRAM-based CIM structure in [12] achieves the smallest area of 200.5 F^2. The compact size not

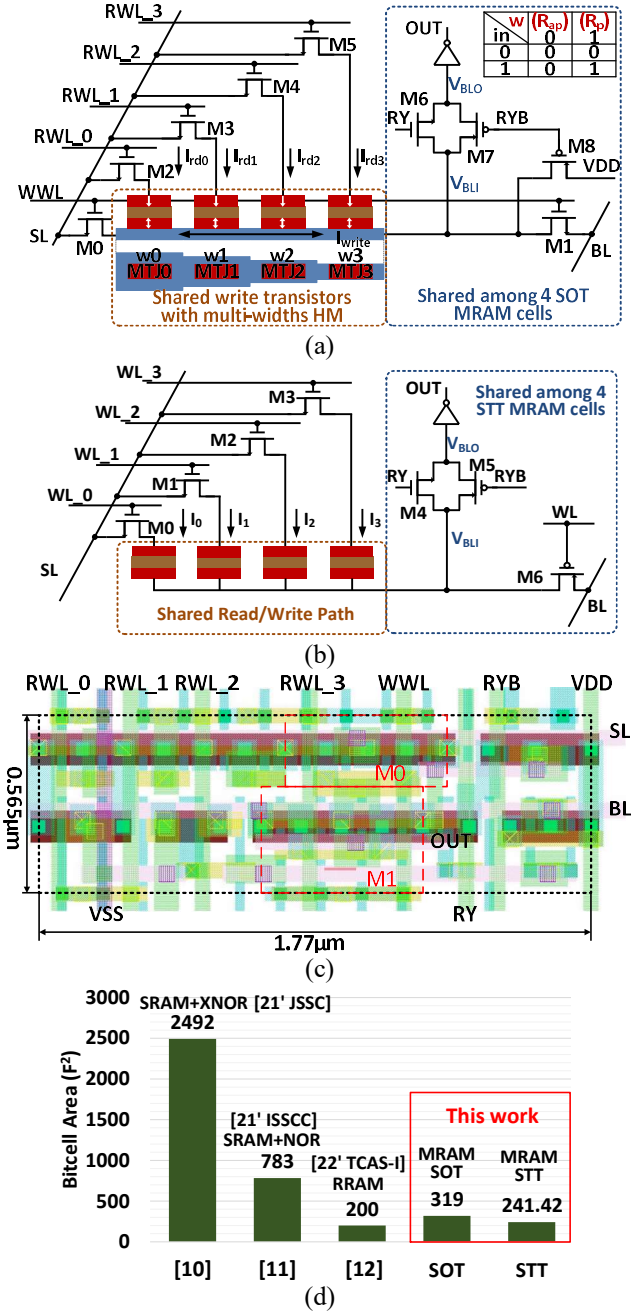

Fig. 3. Proposed area-efficient digital 4-in-1 MRAM-CIM cell structure (a) schematic of SOT MRAM; (b) schematic of STT MRAM; (c) layout of SOT MRAM; (d) Area comparison with state-of-the-arts.

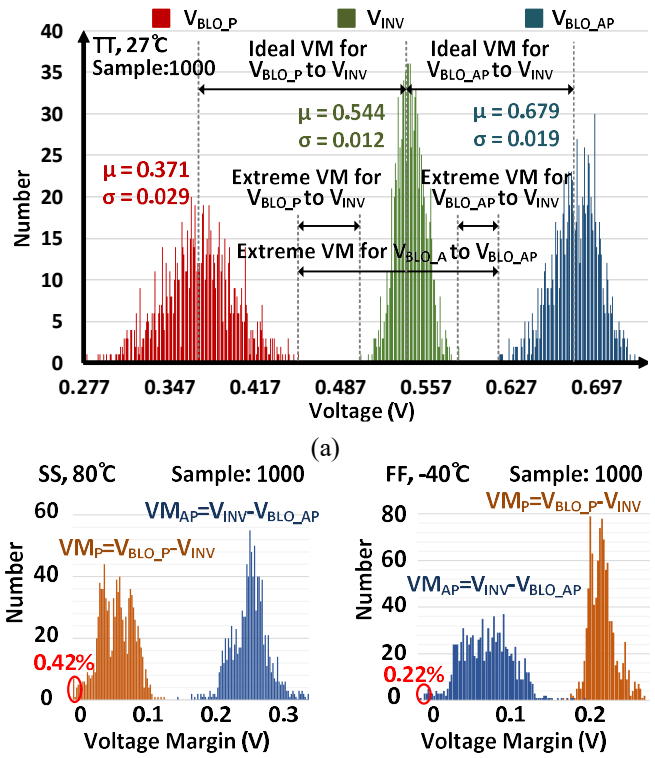

Fig. 4. Voltage margin between the V_{BLO} and switching threshold voltage V_{inv}: (a) in TT corner and 27°C; (b) in SS corner and 80°C; (c) in FF corner and 80°C.

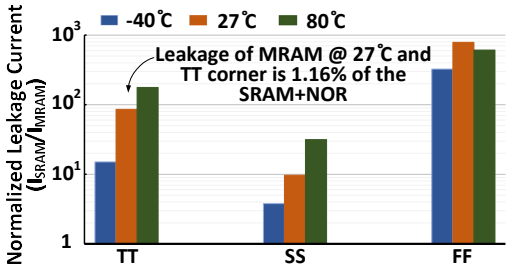

Fig. 5. Normalized leakage current of SRAM+NOR [11] by the proposed SOT MRAM-based CIM bitcell in varying process corners and temperatures from -40°C to 80°C.

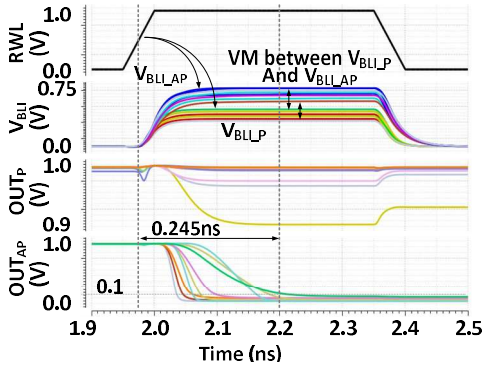

Fig. 6. Access time of the proposed SOT MRAM based CIM bitcell in SS, TT, and FF corners and a temperature range from -40°C to 80°C.

only reduces the cost of manufacturing but also reduces the parasitic parameters on the metal line on the layout, improving the latency of the macro chip.

The challenge of the MRAM device is reading out the R_p and R_{ap} due to the limited resistance on/off ratio and small resistance value. The sensing amplifier and the reference generator designs require high performances in the analog CIM, which causes a significant area and energy overhead compared to SRAM and RRAM-based designs. The proposed digital CIM structure converts the resistance to digital output at the cell level to avoid accuracy loss during the computing operation. To verify the reliability of the proposed digital CIM cell, we have run the Monte Carlo simulation with a sample number of 1000 in SS, TT, and FF corners under -40°C, 27°C, and 80°C. For the MTJ resistance, we adopt the value of R_p = 9 kΩ and R_{ap} = 19 kΩ [4]. Fig. 4 (a) depicts the voltage distribution of V_{BLO} when $R_{MTJ} = R_p$ (red) and $R_{MTJ} = R_{ap}$ (blue) under TT corner and room temperature. The green bars indicate the switching threshold voltage of the inverter (V_{inv}). The gaps indicate the extreme voltage margins between V_{inv} and V_{BL0_P}/V_{BL0_AP}. It shows a sufficient margin in the typical condition. The proposed structure acquires 100% positive margins in most cases except the extreme conditions. Fig. 4(b) and (c) show the distribution of the voltage margin in the worst cases SS corner 80°C and FF corner -40°C. The VM_p is the margin between the V_{BL0_P} and V_{inv} in orange lines and VM_{ap} is the margin between the V_{BL0_AP} and V_{inv} in blue lines. In these extreme conditions, 0.42% and 0.22% negative margins are observed in SS/80°C and FF/-40°C, respectively. The reliable operating temperature range with sufficient margin is -30-60°C.

Leakage is one of the key parameters in memory design. The advantage of NVM over SRAM is that it does not need to keep the power on to hold the data. However, with every shutdown, the power ramp-up will consume energy to charge all the internal nodes. The MRAM-based CIM bitcell has a comparable number of transistors in a typical SRAM+NOR structure [11] (11T vs. 10T). Considering the leakage during the idle time, SRAM still produces a significantly larger leakage compared to MRAM due to the latch-up structure. Fig. 5(b) exhibits the normalized leakage current of SRAM+NOR [11] by the proposed SOT MRAM in process TT, SS, and FF corners and varying temperatures. Both structures are simulated in 28nm technology. We increase the lengths in transistors to regulate the current and reduce the leakage while maintaining the necessary width to provide sufficient switching current. Assuming all the MRAM resistances are R_p. In the typical environment (TT/27°C), the leakage of MRAM-based CIM is 1.16% of the SRAM+NOR structure. Especially in the FF corner, this work achieves 0.12%-0.3% leakage current of [11].

Fig. 6 shows the waveform of the computing operation in the unit cell in varying process corners and temperatures. Due to the small resistance of the MRAM device, the access time (defined as the duration from the RWL is 50% on to the OUT_{AP} reaches 100 mV or OUT_P reaches 900 mV) of the OUT signal achieves 0.245 ns in the worst scenario SS/80°C. It is 3.2 times faster compared to the RRAM-based digital CIM design [12].

III. Performance Evaluation in Macro-Operation

Fig. 7 shows the architecture of the proposed MRAM-based digital CIM macro chip. Each local array consists of 8x16 PE, each PE stores 4b weight, and 4to1 RWL MUX is

TABLE I COMPARISON OF STATE-OF-THE-ART

	JSSC 21 [10]	ISSCC 21 [11]	TCAS-I 22 [12]	VLSI tech 20 [15]	VLSI tech 19 [16]	IEDM 18 [17]	This work
Technology	65nm	22nm	65nm	22nm	22nm	22nm	28nm
Memory	SRAM	SRAM	RRAM	SOT MRAM	SOT MRAM	STT MRAM	SOT MRAM
Bitcell Area (F^2)	2492	783	200.5	140.5	140.5	100	318.9
Supply Voltage (V)	0.6-0.8	0.72	0.8-1	–	–	–	1
CIM Mode	Digital	Digital	Digital	Analog	Analog	Analog	Digital
Ron/Roff (Ω)	NA	NA	10k/500k	6M/15M	10k/21k	1.4k/3.92k	10k/19k
Weight bit	1-16	4/8/12/16	1-8	1.7	–	–	1-8
Input bit	1-16	1-8	1-16	–	–	–	1-16
Output bit	1-16	16(4b/4b) 24(8b/8b)	8-24(1b IN) 16-32(8b IN)	4	4	4	8-16(1b IN) 24-32(1b IN)
Cycle Time (ns)	120(1b/1b) 3590(16b/16b)	10(4b/4b) 18(8b/8b)	6(1b/8b) 48(8b/8b)	–	–	–	5(1b/8b) 42(8b/8b)
Energy efficiency (TOPS/W)	2.06 (16b) 117.3 (1b)	24.7 (8/8/24b) 89 (4/4/16)	27.28 (1/8/15b) 114.41 (1/1/8b)	19.6	9.2	2.5	25.43 (1/8/15b) 129.83 (1/1/8b)

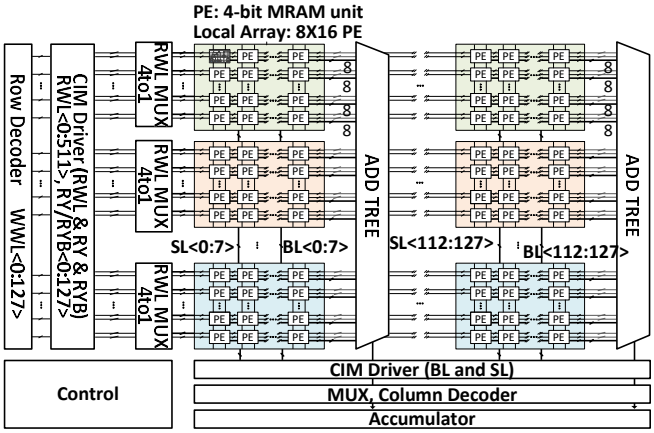

Fig. 7. Overall structure of the proposed MRAM-based 11T4-in-1M digital CIM macro unit.

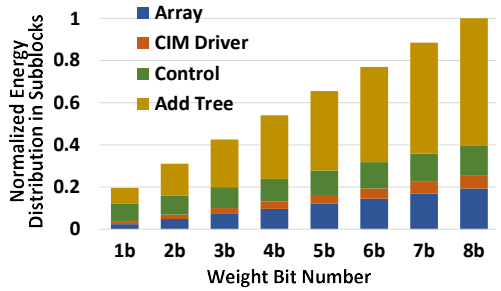

Fig. 8. The normalized energy distribution in subblocks of the proposed 11T4-in-1M CIM macro with 1b input / 1-8b weight / 8-15b output precision.

placed to choose a single weight each time, so one local array supports 1-8b weight computing operation. A single Add Tree is shared with 8 columns. The same mechanism as in previous digital CIM works [10-12], to implement a higher weight bit number, we will add on the local array in the same columns to obtain a maximum 15b weight. The maximum input is 16b, which means for each local array we can turn on up to 16 rows simultaneously. The capacity of the whole array is 128x128 unit cells which store 512 x 128 weights.

Fig. 8 exhibits the energy distribution in each subblock of the proposed macro chip under 1b input and 1-8b weight with 8–15b output in one cycle. The time of one cycle is 5ns for 1b input 1-8b weight. The power consumption in the memory array and CIM driver are simulated with a duration 0.4n window slot for RWL on. During the computing operation, a shorter window slot consumes less direct current flow in the memory array, while the CIM driver will eat up more power for a shorter rise and fall time. The power dissipation of the Add Tree with various bit numbers of output is referred from the previously published works [10-14]. From weight 1b to weight 8b, the percentage of energy consumption in the memory array is increased from 12.3% to 19.3%, and in the Add Tree, it is increased from 38.4% to 60.2%. CIM driver and control have lower energy increases when expanding the bit number of weights.

Table I shows the comparison of the proposed MRAM-based digital CIM design with state-of-the-art. The throughput for 1b/8b/8b precision is 819.2 GOP/S. It achieves the TOPS/W of 129.83 for precision 1b/1b/8b. The proposed design has a significantly smaller area and leakage compared with SRAM-based CIM. RRAM has a compact area due to the 1T1R structure. However, MRAM obtains faster computing speed, the latency is 5ns for 1b input 1-8b weight. [15-17] MRAM in analog CIM designs are taking much smaller sizes than digital CIM cells, but they either apply high resistance values (exceeding proper limits) or perform low energy efficiency.

IV. CONCLUSION

For advanced CIM-based edge computing chips, large capacity is preferred to avoid frequent data caching. The analog CIM designs with large capacities suffer from the process and environmental variations which degrade the accuracy. SRAM-based digital CIM has a large area and leakage due to its structure. The proposed 11T4-in-1M SOT MRAM-based digital CIM design achieves 40% of the SRAM+NOR design [11] and the leakage current at the typical condition is 1.16%. We have run the statistic simulations to verify the reliability of the proposed design in 28nm technology. It obtains 129.83 TOPS/W at 1b/1b/8b precision.

ACKNOWLEDGMENT

This research is supported by Programmatic grant no. A18A6b0057 (SpOT-Lite), Singapore RIE2020 Advanced Manufacturing and Engineering domain.

REFERENCES

[1] A. Trout and X. Cao, "Tradeoff and Solutions of Data Storage in Edge Computing," 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021, pp. 0432-043.

[2] J. -W. Su et al., "16.3 A 28nm 384kb 6T-SRAM Computation-in-Memory Macro with 8b Precision for AI Edge Chips," 2021 IEEE International Solid-State Circuits Conference (ISSCC), 2021, pp. 250-252.

[3] A. Nisar, S. Dhull, S. Mittal and B. K. Kaushik, "SOT and STT-Based 4-Bit MRAM Cell for High-Density Memory Applications," in IEEE Transactions on Electron Devices, vol. 68, no. 9, Sept. 2021, pp. 4384-4390.

[4] K. Ali, F. Li, S. Y. H. Lua and C. -H. Heng, "Area Efficient High Through-put Dual Heavy Metal Multi-Level Cell SOT-MRAM," in IEEE Transactions on Nanotechnology, vol. 19, pp. 613-619, 2020.

[5] Y. Seo and K. Roy, "High-Density SOT-MRAM Based on Shared Bitline Structure," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 26, no. 8, Aug. 2018, pp. 1600-1603.

[6] C. Wang, Z. Wang, B. Wu and W. Zhao, "Design and Optimization of an Area-efficient SOT-MRAM," 2019 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC), 2019, pp. 1-3.

[7] A. Nisar, S. Dhull, S. Mittal, and B. K. Kaushik, "SOT and STT-Based 4-Bit MRAM Cell for High-Density Memory Applications," IEEE Trans. on Electron Device, vol. 68, No. 9, September 2021, pp. 4384-4390.

[8] G. Yuan, X. Ma, S. Lin, Z. Li, J. Deng, and C. Ding, "A DNN Compression Framework for SOT-MRAM-based Processing-In-Memory Engine," IEEE 33rd International System-on-Chip Conference (SOCC), 2020, pp.37-42.

[9] C. Wang, Z. Wang, Y. Xu, J. Yang, Y. Zhang, and W. Zhao, "Computing-in-Memory Architecture based on FieldFree SOT-MRAM with Self-Reference Method," 2020 IEEE International Symposium on Circuits and Systems (ISCAS), 2020, pp. 1-4.

[10] H. Kim, T. Yoo, T. T. -H. Kim and B. Kim, "Colonnade: A Reconfigurable SRAM-Based Digital Bit-Serial Compute-In-Memory Macro for Processing Neural Networks," in IEEE Journal of Solid-State Circuits, vol. 56, no. 7, July 2021, pp. 2221-2233.

[11] Y.-D. Chih et al., "16.4 An 89TOPS/W and 16.3TOPS/mm2 all-digital SRAM-based full-precision compute-in memory macro in 22nm for machine-learning edge applications," in IEEE ISSCC Dig. Tech. Papers, Feb. 2021, pp. 252–254.

[12] V. Sharma, H. Kim and T. T. -H. Kim, "A 64 Kb Reconfigurable Full-Precision Digital ReRAM-Based Compute-In-Memory for Artificial Intelligence Applications," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 69, no. 8, Aug. 2022, pp. 3284-3296.

[13] B. Moons and M. Verhelst, "An Energy-Efficient Precision-Scalable ConvNet Processor in 40-nm CMOS," in IEEE Journal of Solid-State Circuits, vol. 52, no. 4, April 2017, pp. 903-914.

[14] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim and H. -J. Yoo, "UNPU: An Energy-Efficient Deep Neural Network Accelerator With Fully Variable Weight Bit Precision," in IEEE Journal of Solid-State Circuits, vol. 54, no. 1, Jan. 2019, pp. 173-185.

[15] J. Doevenspeck et al., "SOT-MRAM Based Analog in-Memory Computing for DNN Inference," 2020 IEEE Symposium on VLSI Technology, 2020, pp. 1-2.

[16] K. Garello et al., "Manufacturable 300mm platform solution for Field-Free Switching SOT-MRAM," 2019 Symposium on VLSI Technology, 2019, pp. T194-T195.

[17] O. Golonzka et al., "MRAM as Embedded Non-Volatile Memory Solution for 22FFL FinFET Technology," 2018 IEEE International Electron Devices Meeting (IEDM), 2018, pp. 18.1.1-18.1.4.