




Area Efficient Computing-in-Memory Architecture Using STT/SOT Hybrid Three Level Cell

SEEMA DHULL ¹ (Graduate Student Member, IEEE), ARSHID NISAR ¹ (Graduate Student Member, IEEE), RAKESH BHAT², AND BRAJESH KUMAR KAUSHIK ¹ (Senior Member, IEEE)

¹Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand 247667, India

²High-velocity Product Group (HPG), Intel Technology India Pvt. Ltd., Bengaluru 560103, India

CORRESPONDING AUTHOR: BRAJESH KUMAR KAUSHIK (e-mail: bkk23fec@iitr.ac.in)

ABSTRACT Spintronic-based computing-in-memory (CiM) architecture has emerged as one of the efficient solutions to counter the latency/bandwidth bottleneck of conventional von-Neumann architecture. However, computation within a small area while achieving low power consumption still remains a challenge. Multi-bit spintronic storage device is a suitable solution to improve the integration density of such architectures. This paper focuses on using spin-transfer torque (STT)/spin-orbit torque (SOT) based hybrid three-level cell (TLC) in CiM application for implementing logic circuits such as AND, XOR, and magnetic full adder (MFA). Moreover, the performance of the STT/SOT-TLC-based MFA is compared with other full adder designs. The results show that the proposed MFA is 75% more area-efficient in comparison to two-bit STT and SOT-based designs, and 50% more area-efficient in comparison to differential spin hall effect (DSHE) based designs

INDEX TERMS Compute-in-memory, magnetic memory, multi-level cell, magnetic full adder, triple-level cell.

I. INTRODUCTION

The main setback with traditional von Neumann architecture is that it consumes significant energy for communicating between memory and processing unit. This separation causes high data traffic and processing time which has become a performance limitation. New architectures such as near-memory computation and CiM are capable of overcoming this shortcoming [1]–[3]. At present, the CiM architecture needs to be explored more since it has higher scope than any other architecture because it is faster and requires the least amount of energy for data computation [4], [5].

Spintronics based non-volatile memories are potential candidates to realize CiM architecture due to zero static power, compact structure, and high compatibility with CMOS technology. The popular spin device that made an impact commercially is the STT-magnetic tunnel junction (MTJ). The MTJ is a fundamental storage element that consists of a non-magnetic material sandwiched between two ferromagnetic (FM) layers. The magnetization in one of the FM layers is

pinned to a particular direction and the other is configured in either parallel (P) or antiparallel (AP) to it. This results in two resistive states that help in storing logic states. The device has high endurance, high access speed, and low static power but faces reliability challenges such as oxide breakdown and erroneous read/write operation due to the same read-write path [6]. Furthermore, SOT-MTJ is explored that overcomes the issues associated with STT-MTJ device. It uses the spin hall effect (SHE) and provides high write speed along with high reliability because of the separated read-write path [7]. The value of switching current and write energy can be reduced further by the mechanism of voltage-controlled magnetic anisotropy (VCMA) [8]. However, perpendicular magnetic anisotropy SOT-MTJ devices require an external magnetic field for magnetization reversal that degrades the performance of the device. It is possible to switch the magnetization without applying any external field with the interplay of both STT and SOT [9]. All the aforementioned devices are capable of storing only a single bit per device. These are

not sufficient for high-density memory applications due to the large footprint of access transistors [10], [11]. It increases the scope for multi-level devices that can store more than one bit per device and reduce cost per bit. A typical multilevel cell (MLC) can be built by connecting two or more MTJs in series or parallel [12]. However, it comes with several hurdles such as write speed and sensing margin. A dual-bit cell (DLC) requires two steps for the write operation leading to higher latency [13]. Moreover, its capability of 2-bit storage per cell does not solve the storage density issue [14]. One solution is to build an MLC with more than 2 bits while maintaining considerable latency and energy consumption. Generally, a regular 3-bit MLC requires three steps to complete a single write operation making the latency worse. Besides, the division of resistance states into eight distinct levels reduces the sensing margin to distinguish two states while reading the device [15]. STT/SOT hybrid TLC is a device that stores 3-bits per cell and solves the aforementioned issues of 3-bit MLC [16]. It requires only 2 steps to complete a single write operation and provides an improved sensing margin for better readability. Moreover, these aforementioned spin devices are capable of performing both memory and logic operations altogether. Hence, spintronics-based magnetic memories are potential candidates for CiM applications owing to their remarkable aforementioned features [17], [18]. This work focuses on the implementation of logic circuits in a memory array of TLC. Making use of the TLC device in CiM comes with the aforementioned advantages of the device along with the reduction in the overall area utilization. The key contributions of the proposed work are as follow:

- The three-bit MLC is used for the implementation of CiM based logic gates and magnetic full adder (MFA). The TLC based MFA in CiM architecture achieves 75%, 75% and 50% improvement in area efficiency when compared to STT-DLC, SOT-DLC, and DSHE based circuits, respectively.
- The combination of SOT and STT switching scheme is used to achieve optimal latency and energy consumption in TLC device. The complete write operation is performed using only 2 steps and provides an improved sensing margin using self-referencing read operation.

The rest of the paper is organized as: Section II presents the structure of the TLC. Section III presents the design of logic gates and MFA using TLC. It also presents the performance analysis of these circuits. Finally, Section IV concludes the paper.

II. STT/SOT HYBRID TRIPLE LEVEL CELL

The STT/SOT TLC structure is comprised of three MTJs stacked one upon another and the free layer of base MTJ resting on heavy metal layer as shown in Fig. 1(a). Here, PL and FL stand for pinned layer and free layer of MTJ, respectively. HM is the heavy metal attached to the free layer of MTJ₃ which is responsible for the SOT switching mechanism. The MTJ₂ and MTJ₃ are equal in dimension and larger as compared to MTJ₁. This equality in dimension results in only

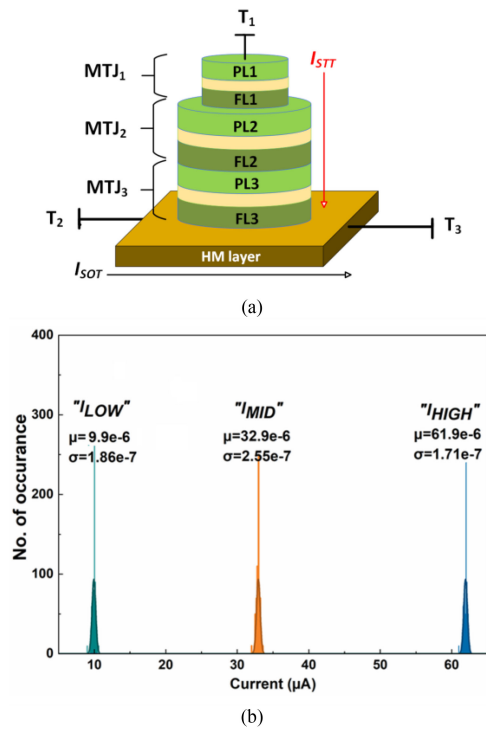


FIGURE 1. (a) Structure of three level cell (b) Statistical distribution of operating STT currents.

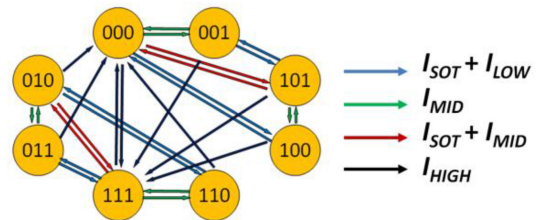


FIGURE 2. State transition diagram for the write operation of TLC.

six distinct resistance states instead of eight. The magnetization state of each MTJ can be switched by passing a current through the TLC and thus logical bits can be stored in each of them by configuring the magnetization state either in parallel or antiparallel. Magnetization switching of MTJ₁ and MTJ₂ is achieved with the help of the STT mechanism and MTJ₃ is switched with the assistance of the SOT mechanism. Hence, there are two currents i.e., I_{STT} and I_{SOT} controlling the write operation of the cell where I_{STT} flows from T_1 to T_3 and I_{SOT} flows from T_2 to T_3 . I_{STT} has three current values i.e., I_{LOW} , I_{MID} , and I_{HIGH} . Here, I_{MID} is sufficient to switch the magnetization of only MTJ₁, I_{HIGH} can switch the magnetization of all the MTJs, and I_{LOW} alone cannot switch any of the MTJs but it can be used to switch MTJ₃ along with the assistance of I_{SOT} . Fig. 1(b) shows the statistical distribution of all STT currents. Fig. 2 shows the state transition diagram for the write operation. The simulation parameters for the device are shown in Table I. The dynamics of magnetization of the free layer

TABLE I Simulation Parameters Of STT/SOT-TLC

Parameter	Value
Diameter	34nm (MTJ ₁), 40nm (MTJ ₂ /MTJ ₃)
Heavy Metal Size (l × w × d)	60nm × 40nm × 3nm
Saturation Magnetization	1.25 × 10 ⁶ A/m
Magnetic Anisotropy	163000 A/m (MTJ ₁), 143000 A/m (MTJ ₂ /MTJ ₃)
Free Layer Thickness	1nm
Damping Constant	0.03
Spin Hall Angle	0.3

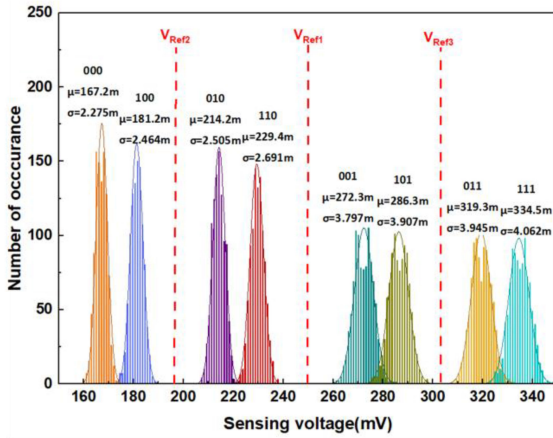


FIGURE 3. Statistical distribution of sensing voltages of all states of TLC.

is governed by Landau–Lifshitz–Gilbert (LLG)–Slonczewski equation and is represented as [19]:

$$\frac{\partial \vec{m}}{\partial t} = -\gamma \mu_0 \vec{m} \times \vec{H}_{eff} + \alpha \vec{m} \times \frac{\partial \vec{m}}{\partial t} - \xi P J_{STT} \vec{m} \times (\vec{m} \times \vec{m}_r) - \xi \eta J_{SHE} \vec{m} \times (\vec{m} \times \vec{\sigma}_{SHE}) \quad (1)$$

Where, terms on the right-hand side represents torques on free layer magnetization (\vec{m}) due precession, Gilbert damping, STT and spin hall effect. \vec{m}_r represents the magnetization of reference layer and \vec{H}_{eff} is the effective magnetic field. J_{STT} and J_{SHE} are STT and SHE write current densities, respectively. $\vec{\sigma}_{SHE}$ is the polarization of pure current injected into the free layer. γ represents the gyromagnetic ratio and α is the damping factor. The effective magnetic field (\vec{H}_{eff}) is mainly composed of perpendicular magnetic anisotropy field (\vec{H}_{PMA}), exchange field bias (\vec{H}_{EX}), and thermal noise field (\vec{H}_{TH}). A special self-referencing read scheme has been used to distinguish the overlapping resistance levels [16]. Fig. 3 shows the statistical distribution of sense voltages of all the states and three reference voltages that are used to differentiate the states. In addition, the TLC has a significant reduction in latency and energy consumption.

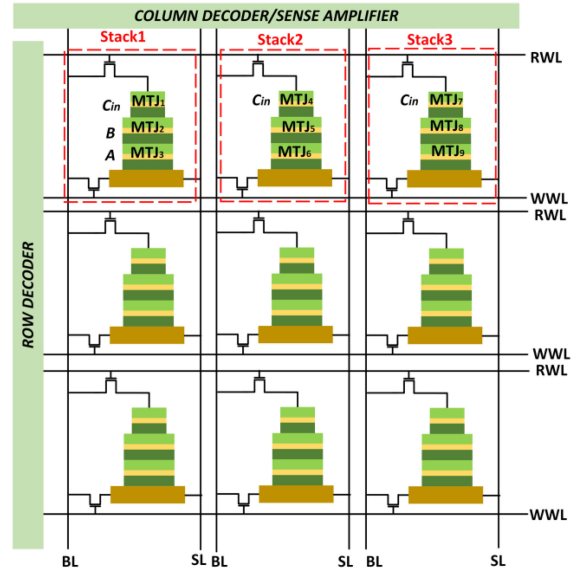


FIGURE 4. Implementation of logic circuits in TLC memory array.

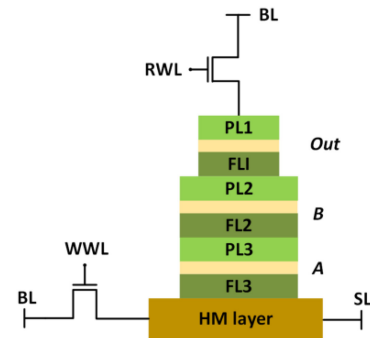


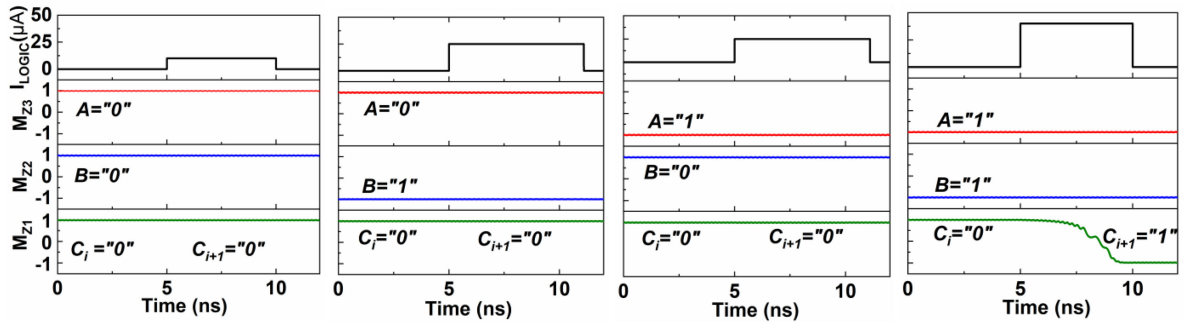
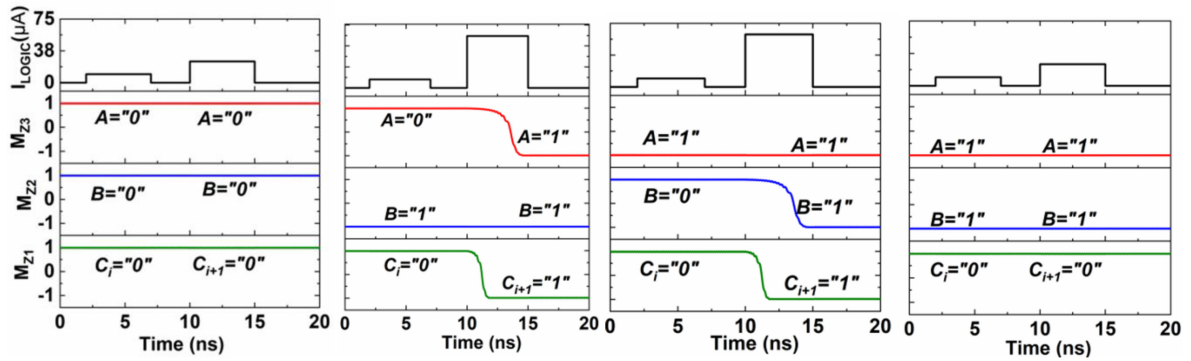
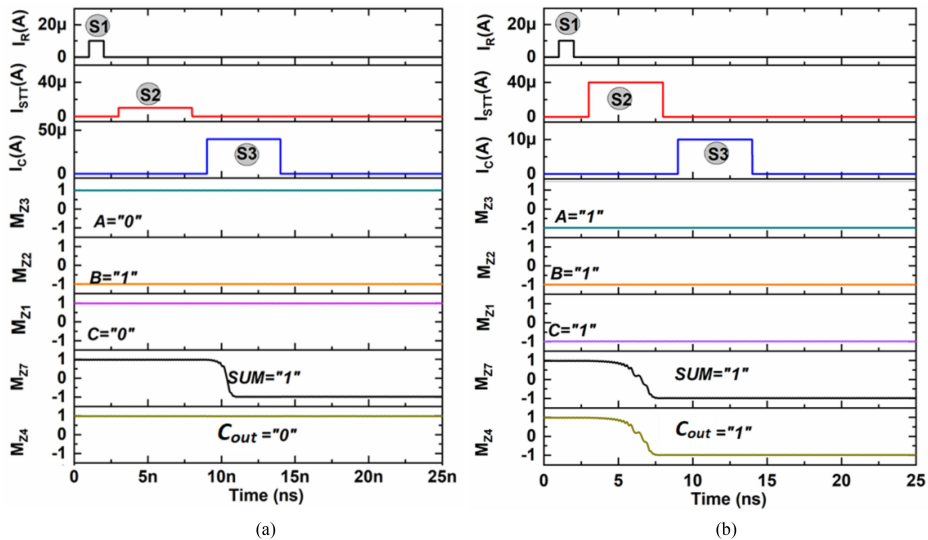
FIGURE 5. TLC used as a logic gate.

III. COMPUTE-IN-MEMORY USING TLC

In the proposed CiM architecture, as shown in Fig. 4, the STT/SOT TLC memory device performs the functions of both the data storage and the computation. The two operations can be achieved by selecting either memory mode or compute mode. In memory mode, the data is written to or read from memory. The data is written by hybrid STT and SOT switching techniques as discussed in Section II. The read operation is performed by passing a low read current through the TLC memory and the data is detected by using a sense amplifier. The computation mode is activated to perform any Boolean or non-Boolean operation.

A. LOGIC GATES IMPLEMENTATION

In Fig. 5, the TLC is configured as a logic gate. MTJ₁ is used to store the output whereas MTJ₂ and MTJ₃ are for inputs. Depending upon the initial state of the input data, the current encoding scheme is used to switch the magnetization of the output MTJ device. The logic operations NAND, AND, OR, NOR, XOR, and XNOR are performed within the same TLC


FIGURE 6. Transient analysis of the TLC based AND operation.

FIGURE 7. Transient analysis of the TLC based XOR operation.

FIGURE 8. Transient analysis of magnetic full adder for input combinations of (a) 010 (b) 111.

device. The implementation of logic operations is expressed by the following equations:

$$AND/OR = A.B + A.C + B.C \quad (2)$$

where C acts as a control signal and is represented by the initial state of the MTJ₁. When $C=0$, AND operation is executed whereas when $C=1$, OR operation is performed. Fig. 6 shows

the transient analysis of AND operation. Table II shows the encoding currents for AND logic implementation.

$$NAND/NOR = \bar{A}.\bar{B} + \bar{A}.C + \bar{B}.C \quad (3)$$

where C acts as a control signal and is represented by the initial state of the MTJ₁. When $C=0$, NAND operation is

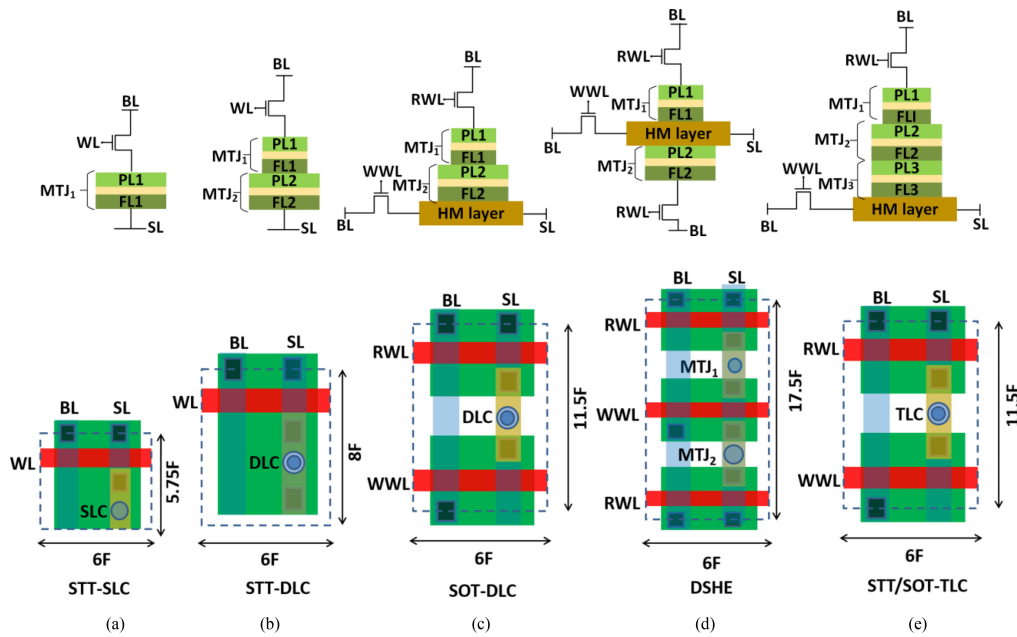


FIGURE 9. Schematic of device structure and layout of (a) STT-SLC (b) STT-DLC (c) SOT-DLC (d) DSHE (e) STT/SOT TLC.

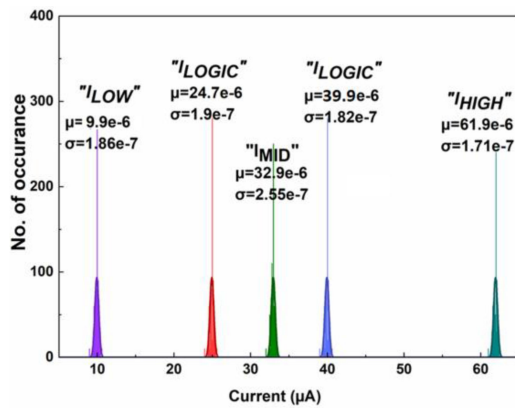


FIGURE 10. Statistical distribution of all logic currents.

TABLE II Current Encoding for 2-Bit AND Operation

A	B	I_{LOGIC} (μA)	AND
0	0	10	AP
0	1	25	AP
1	0	25	AP
1	1	40	P

executed whereas when $C=I$, NOR operation is performed.

$$XOR/XNOR = \bar{A}.\bar{B}.C + \bar{A}.B.\bar{C} + A.B.C + A.\bar{B}.\bar{C} \quad (4)$$

where C acts as a control signal and is represented by the initial state of the MTJ_1 . When $C=0$, XOR operation is executed whereas, when $C=I$, XNOR operation is performed.

TABLE III Current Encoding For 2-Bit Xor Operation

Input State (AB)	I_{STT1} (μA)	I_{STT2} (μA)	XOR
00	10	25	AP
01	-10	62	P
10	10	62	P
11	-10	25	AP

TABLE IV Stepwise Operation Of Proposed TLC Based MFA

Steps	Stack1	Stack2	Stack3
S1	Input (A, B, C_{in})		
S2		Copy C_{in} to MTJ_4	Copy C_{in} to MTJ_7
S3		Compute C_{out}	Compute Sum

Fig. 7 shows the transient analysis of XOR operation. Table III shows the encoding currents for XOR logic implementation.

B. MAGNETIC FULL ADDER

To implement the full adder operation within the CiM array, the input data A , B , and C_{in} are stored in one TLC stack while the computation and storage of Sum and Carry (C_{out}) results are performed in the second and third TLC stack as shown in Fig. 4. The input data C_{in} is first copied from the TLC stack1 and stored to the MTJ_4 of TLC stack2. Depending upon the initial state of MTJ_4 , C_{out} and Sum operations are computed and stored in MTJ_4 of the TLC stack2 and MTJ_7 of the TLC stack3, respectively. The Sum is computed using the XOR and XNOR operation whereas the C_{out} is computed AND and OR operations as described by the equations (2) and (4), respectively. The steps for full adder operation as shown in Table IV, are represented by

TABLE V Area Consumption Of Various Cells

Parameter	STT-SLC	STT-DLC	SOT-DLC	DSHE	STT/SOT TLC
Area (F ²)	34.5	48	69	105	69
Density (F ² /bit)	34.5	24	34.5	52.5	23
Write energy (fJ/bit)	480	520	236	240	27.8
Write Latency (ns/bit)	5	5	3.5	0.93	7.8

TABLE VI Area Comparison Of Various Full Adders

Parameters	CMOS based FA [23]	STT-SLC MFA [21]	SOT-SLC MFA [22]	DSHE-MFA [20]	STT-DLC MFA	SOT-DLC MFA	Proposed STT/SOT-TLC MFA
No. of transistors	46	18	18	12	24	24	6
No. of MTJs	0	9	9	8	12	12	9
Total area(F ²)	1587	621	621	414	828	828	207

S1: Read C_{in} , S2: Write C_{in} to MTJ₄ of Stack2 and MTJ₇ Stack3, S3: Sum and C_{out} computation. Fig. 8 shows the transient analysis of MFA for input combinations “010 and “111”.

The performance of STT/SOT TLC has compared with other STT and SOT based MLCs at device and circuit level. Fig. 9 shows the schematic of various STT and SOT based MLCs and their layout. The device level area assessment shows that the total area consumed by the SOT/STT TLC cell is $69F^2$. Therefore, the area consumed per bit is $23F^2$. It is 34%, 4.1%, 33.3%, and 54% more area efficient as compare to STT-SLC, STT-DLC, SOT-DLC and DSHE, respectively as shown in Table V. Owing to the area efficiency of STT/SOT-TLC, the area of the proposed MFA design is significantly reduced by 75% in comparison to STT-DLC and SOT-DLC based designs, and 50% in comparison to DSHE based design [20]. Moreover, the proposed MFA is 66.6% more area efficient when compared to STT-SLC [21] and SOT-SLC [22]. The performance comparison of STT/SOT-TLC based MFA with other MLC based designs is presented in Table VI. The usage of a multilevel cell plays huge role in area reduction and achieving high integration density. Along with the performance analysis, this work also focuses on exploring the effect of parametric variation on the MFA operation. 1000 Monte Carlo simulations with 3% variation in TLC device parameters namely oxide layer thickness, free layer thickness and TMR and 10% variation in CMOS transistor length and width have been performed for all the write currents, logic currents as well as sensing resistances. The statistical distribution of write currents, sensing voltages, and logic currents are shown in Fig. 1(b), Fig. 3, and Fig. 10, respectively. It is evident that there are enough margins between the different currents and sensing voltages that enable reliable write and read operations, respectively.

IV. CONCLUSION

The proposed work presents an implementation of area efficient STT/SOT based TLC MRAM based CiM architecture for logic circuits. The STT/SOT TLC based MFA shows a significant reduction in area consumption by 66.6%, 75%, 75%, and 50%, as compared to STT-SLC, STT-DLC, SOT-DLC, and DSHE based MFA, respectively. The future scope of

this work could include image processing and neural network applications using STT/SOT-TLC arrays.

REFERENCES

- [1] S. K. Nam *et al.*, “Leakage current: Moore’s law meets static power,” *Computer*, vol. 36, no. 12, pp. 68–75, Dec. 2003, doi: [10.1109/MC.2003.1250885](https://doi.org/10.1109/MC.2003.1250885).
- [2] W. A. Wulf and S. A. McKee, “Hitting the memory wall: Implications of the obvious,” *SIGARCH Comput. Architecture News*, vol. 23, no. 1, pp. 20–24, Mar. 1995, doi: [10.1145/216585.216588](https://doi.org/10.1145/216585.216588).
- [3] X. Huang, C. Liu, Y.-G. Jiang, and P. Zhou, “In-memory computing to break the memory wall,” *Chin. Phys. B*, vol. 29, no. 7, Jul. 2020, Art. no. 078504, doi: [10.1088/1674-1056/ab90e7](https://doi.org/10.1088/1674-1056/ab90e7).
- [4] J. Ahn, S. Yoo, O. Mutlu, and K. Choi, “PIM-enabled instructions: A low-overhead, locality-aware processing-in-memory architecture,” in *Proc. 42nd Annu. Int. Symp. Comput. Architecture-ISCA’15*, Portland, OR, 2015, pp. 336–348, doi: [10.1145/2749469.2750385](https://doi.org/10.1145/2749469.2750385).
- [5] M. Gao, G. Ayers, and C. Kozyrakis, “Practical near-data processing for in-memory analytics frameworks,” in *Proc. Int. Conf. Parallel Architecture Compilation*, San Francisco, CA, USA, Oct. 2015, pp. 113–124, doi: [10.1109/PACT.2015.22](https://doi.org/10.1109/PACT.2015.22).
- [6] S. Bhatti, R. Sbiaa, A. Hirohata, H. Ohno, S. Fukami, and S. N. Piramanayagam, “Spintronics based random access memory: A review,” *Mater. Today*, vol. 20, no. 9, pp. 530–548, Nov. 2017, doi: [10.1016/j.mattod.2017.07.007](https://doi.org/10.1016/j.mattod.2017.07.007).
- [7] C. Zhang, S. Fukami, H. Sato, F. Matsukura, and H. Ohno, “Spin-orbit torque induced magnetization switching in nano-scale Ta/cofeb/mgO,” *Appl. Phys. Lett.*, vol. 107, no. 1, Jul. 2015, Art. no. 012401, doi: [10.1063/1.4926371](https://doi.org/10.1063/1.4926371).
- [8] H. Lee *et al.*, “Analysis and compact modeling of magnetic tunnel junctions utilizing voltage-controlled magnetic anisotropy,” *IEEE Trans. Magn.*, vol. 54, no. 4, pp. 1–9, Apr. 2018, doi: [10.1109/TMAG.2017.2788010](https://doi.org/10.1109/TMAG.2017.2788010).
- [9] M. Wang *et al.*, “Field-free switching of a perpendicular magnetic tunnel junction through the interplay of spin-orbit and spin-transfer torques,” *Nature Electron*, vol. 1, no. 11, pp. 582–588, Nov. 2018, doi: [10.1038/s41928-018-0160-7](https://doi.org/10.1038/s41928-018-0160-7).
- [10] Emerging research devices, “International technology road map for semiconductors,” Aug. 2011. Accessed: Jun. 20, 2021. [Online]. Available: <https://www.semiconductors.org/wp-content/uploads/2018/08/2011ERD.pdf>
- [11] B. Tallis, “IBM and everspin announce 19TB NVMe SSD with MRAM 277 write cache,” *AnandTech*, Aug. 2018. Accessed: Jun. 18, 2021. [Online]. Available: <https://www.anandtech.com/show/13174/ibm-and-everspin-announce>
- [12] Y. Kim, X. Fong, K.-W. Kwon, M.-C. Chen, and K. Roy, “Multilevel spin-orbit torque MRAMs,” *IEEE Trans. Electron Devices*, vol. 62, no. 2, pp. 561–568, Feb. 2015, doi: [10.1109/TED.2014.2377721](https://doi.org/10.1109/TED.2014.2377721).
- [13] L. Jiang, B. Zhao, Y. Zhang, and J. Yang, “Constructing large and fast multi-level cell STT-MRAM based cache for embedded processors,” in *Proc. 49th Annu. Des. Automat. Conf.*, New York, NY, USA, 2012, pp. 907–912, doi: [10.1145/2228360.2228521](https://doi.org/10.1145/2228360.2228521).
- [14] Y. Deguchi, S. Suzuki, and K. Takeuchi, “Write and read frequency-based word-line batch V_{TH} Modulation for 2-D and 3-D-TLC NAND flash memories,” *IEEE J. Solid-State Circuits*, vol. 53, no. 10, pp. 2917–2926, Oct. 2018, doi: [10.1109/JSSC.2018.2852748](https://doi.org/10.1109/JSSC.2018.2852748).

- [15] Z. Li, B. Yan, L. Yang, W. Zhao, Y. Chen, and H. Li, "A new self-reference sensing scheme for TLC MRAM," in *Proc. IEEE Int. Symp. Circuits Syst.*, Lisbon, Portugal, May 2015, pp. 593–596, doi: [10.1109/ISCAS.2015.7168703](https://doi.org/10.1109/ISCAS.2015.7168703).
- [16] Y. Xu, B. Wu, Z. Wang, Y. Wang, Y. Zhang, and W. Zhao, "Write-Efficient STT/SOT hybrid triple-level cell for high-density MRAM," *IEEE Trans. Electron Devices*, vol. 67, no. 4, pp. 1460–1465, Apr. 2020, doi: [10.1109/TED.2019.2963421](https://doi.org/10.1109/TED.2019.2963421).
- [17] W. Kang, L. Chang, Z. Wang, and W. Zhao, "In-memory processing paradigm for bitwise logic operations in STT-MRAM," in *Proc. IEEE Int. Magn. Conf.*, Dublin, May 2017, pp. 1–1, doi: [10.1109/INT-MAG.2017.8008048](https://doi.org/10.1109/INT-MAG.2017.8008048).
- [18] A. Nisar, S. Dhull, S. Shreya, and B. K. Kaushik, "Energy-Efficient advanced data encryption system using spin-based Computing-in-Memory architecture," *IEEE Trans. Electron Devices*, vol. 69, no. 4, pp. 1736–1742, Apr. 2022, doi: [10.1109/TED.2022.3150623](https://doi.org/10.1109/TED.2022.3150623).
- [19] Z. Wang, W. Zhao, E. Deng, J.-O. Klein, and C. Chappert, "Perpendicular-anisotropy magnetic tunnel junction switched by spin-Hall-assisted spin-transfer torque," *J. Phys. D: Appl. Phys.*, vol. 48, no. 6, 2015, Art. no. 065001.
- [20] S. Shreya, A. Jain, and B. K. Kaushik, "Computing-in-memory architecture using energy-efficient multilevel voltage-controlled spin-orbit torque device," *IEEE Trans. Electron Devices*, vol. 67, no. 5, pp. 1972–1979, May 2020, doi: [10.1109/TED.2020.2978085](https://doi.org/10.1109/TED.2020.2978085).
- [21] M. Zabihi, Z. I. Chowdhury, Z. Zhao, U. R. Karpuzcu, J. Wang, and S. S. Sapatnekar, "In-memory processing on the spintronic CRAM: From hardware design to application mapping," *IEEE Trans. Comput.*, vol. 68, no. 8, pp. 1159–1173, Aug. 2019, doi: [10.1109/TC.2018.2858251](https://doi.org/10.1109/TC.2018.2858251).
- [22] M. Zabihi *et al.*, "Using spin-hall MTJs to build an energy-efficient in-memory computation platform," in *20th Int. Symp. Qual. Electron. Des.*, Santa Clara, CA, USA, Mar. 2019, pp. 52–57, doi: [10.1109/ISQED.2019.8697377](https://doi.org/10.1109/ISQED.2019.8697377).
- [23] E. Deng, Y. Zhang, J.-O. Klein, D. Ravelsona, C. Chappert, and W. Zhao, "Low power magnetic full-adder based on spin transfer torque MRAM," *IEEE Trans. Magn.*, vol. 49, no. 9, pp. 4982–4987, Sep. 2013, doi: [10.1109/TMAG.2013.2245911](https://doi.org/10.1109/TMAG.2013.2245911).