İSTANBUL UNIVERSITY
CERRAHPAŞA

ORIGINAL ARTICLE
Electrica 2024

High-Performance Multi-level Cell Design Using Reduced Retention Time Spintronics Device

Alisha P. B., Tripti S. Warrier

Department of Electronics, Cochin University of Science and Technology, Kerala, India

Cite this article as: A. P. B. and T. S. Warrier, "High-performance multi-level cell design using reduced retention time spintronics device," *Electrica*, Published online September 9, 2024. doi: 10.5152/electrica.2024.23037.

ABSTRACT

Present-day computationally intensive on-chip applications consume much area and energy. Multi-level cells (MLCs) capable of storing two-bit information can reduce this area and energy consumption. Spintronics-based MLCs are suitable for on-chip memory with salient features like nonvolatility, high endurance, density, zero leakage, etc. Even though spintronic-based MLCs perform well regarding leakage power, speed, and thickness, they require a large switching current that increases energy dissipation during the write operation. Also, external sensing and feedback circuits lead to power and area overheads. This paper proposes a new high-performance Voltage Controlled Spin-Orbit Torque (VGSOT) based parallel MLC that uses Magnetic Tunnel Junctions (MTJs) with peculiar characteristics. The novel approach is to configure the MTJs with different retention times. The proposed model has reduced critical current for switching and write energy compared to conventional MLC due to the decreased energy barrier for switching in the MTJ. The energy barrier is reduced by tuning the retention time of the MTJ, which can be done by modulating the geometrical and physical parameters. Further, the flaws like leakage power, scaling, and the threshold voltage of conventional CMOS transistors in the read and write channel of the presented MLC are compensated by a Carbon nanotube Field Effect Transistor (CNFET). The comprehensive simulations are used to confirm the functionality of the proposed design using a 45 nm Cadence virtuoso. The VGSOT-based MLC's read-and-write circuit eliminates a sense amplifier and bistable feedback, making it radiation resistant. Compared to pioneering MLC architectures, the design reduces critical switching current by 50% and write energy by 42% for a 2-bit memory cell.

***Index Terms*—**Spintronics, Voltage controlled spin-orbit torque, Carbon nanotube field effect transistor, Magnetic tunnel junction.

I. INTRODUCTION

Multilevel non-volatile memories offer opportunities for future energy and area-efficient edge computing paradigms. Multilevel logic allows more data to be stored and sent per unit area of the chip, allowing for excellent integration efficiency [1]. Spintronics-based multi-level cells (MLCs) are the next-generation memory technology with energy-efficient performance. Spin torque transfer magnetic random access memory (STTMRAM) and Spin-orbit torque magnetic random access memory (SOTMRAM) have emerged as the primary contenders for multi-level cell (MLC) applications due to their advantageous features such as scalability and non-volatility [2-4]. These technologies offer the potential to achieve four distinct levels of resistance within the memory cell. This multilevel switching is achieved by combining the different states of two magnetic tunnel junctions (MTJs), configuring them in either a parallel or series arrangement The individual MTJs are configured with different critical switching current (I_c) and resistance variations (ΔR) [5].

Comparing this to traditional single-level SOTRAM results in space savings at increased writing energy costs. A SOTRAM-MLC with a shared diode (oxide-based) stacked over two in parallel MTJs for a read operation was proposed by [6]. This design increases the read voltage/energy to address the voltage drop across the diode. A double magnetic tunnel junction (DMTJ) with two perpendicular magnetic tunnel junctions (MTJ) of varying diameters that are coupled serially to each other to provide the four different resistance states suggested in [7]. In MLC, to ensure independent switching among the MTJs, it should follow non-identical configurations. To conduct read and write operations at separate nodes, both time and current dependent programming are used [8, 9]. In current dependent programming, the switching current I_c is modulated. Similarly, the voltage pulse width is modulated differently in time-dependent programming.

Corresponding author:
Alisha P. B.

E-mail:
pbalisha@cusat.ac.in

Received: March 18, 2023

Revision requested: July 3, 2023

Last revision received: April 26, 2024

Accepted: June 11, 2024

Publication Date: September 9, 2024

DOI: 10.5152/electrica.2024.23037

Critical memory properties like write energy, integration density, performance, and endurance are directly influenced by switching current and voltage pulse width. To configure energy and area-efficient spintronics MLC, most of the prior work focused on optimizing I_c [9, 10]. The aforementioned configurations are achieved by modulating the dimensions of MTJs, like the heavy metal layer (HM). Different heavy metal layer widths beneath each MTJ help to obtain different critical currents. This modification of HM has two immediate effects: (1) a rise in write and read error; and (2) a rise in the bit-cell area. In addition to being less energy-efficient and having a significant write current, STTMRAM/SOTMRAM-MLC has a vast area footprint, which negates the benefit of increased integration density.

The process of spin-orbit torque (SOT) switching has the capability not only to trigger the SOT effect but also to create exchange bias (HEX), which has practical applications. However, due to the relatively weak HEX and the fact that field-free SOT switching is involved, this leads to significant issues with the reliability of switching. Additionally, a substantial SOT switching current is necessary, resulting in high energy consumption during the switching process. To tackle these challenges, the study introduces the use of voltage-controlled magnetic anisotropy (VCMA) to assist in SOT switching. This innovative switching approach is termed voltage-gated SOT (VGSOT) switching [11].

This work modeled VGSOT-MTJ with different retention times to achieve energy-efficient MLC without any additional overhead. Retention time, a characteristic of non-volatile memory, is the anticipated period of time until a random bit-flip takes place in a memory cell [12]. The write latency and write current of spintronics memory are also decreased by reducing the retention time. Even though the presence of Complementary Metal Oxide Semiconductor (CMOS) transistors in the read and write path of spintronics devices (Fig. 1) will result in high leakage current due to the scale down of the channel length. Low channel mobility and a low threshold voltage of the MOS transistor make the scenario worse. This leads to degrading the performance of VGSOT-based MLC. This design replaces the CMOS transistors with the carbon nanotube field effect transistor (CNFET).

This manuscript proposes a feasible method to design the VGSOT-MLC to improve area and energy efficiency. The multiple MTJs with different retention times on a common antiferromagnetic strip are stacked parallel and share the CNFET as write and read access. Since the suggested design's read-and-write circuit lacks a sense amplifier and doesn't have bistable feedback, it is radiation-resistant. The objective of the work are:

- To design a VGSOT with reduced retention time.
- To propose a novel memory cell with RRT-VGSOT and CNFET logic with non-volatility and high integration capability.
- To design 2-bit parallel multilevel switching using VGSOT with modulating the retention time of MTJ.
- Evaluate the performance of the proposed design with various parameters like switching current, energy consumption, and process variation
- To compare the proposed design with current multilevel switching techniques using STTRAM and SOTRAM.

The proposed work is the first attempt to design an MLC structure using VGSOT with different retention times MTJ and CNTFET in the read and write path.

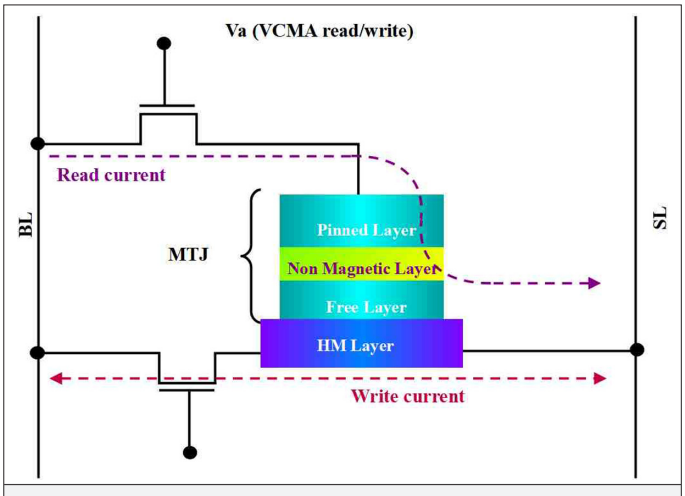

Fig. 1. Bit cell schematic of single level VGSOT.

II. PRELIMINARIES

This section discusses the basic structure of VGSOT and some state-of-the-art spintronics-based multilevel switching.

A. Device Structure

VGSOT is a crucial component in this MLC architecture. It is a three-terminal structure comprising two magnetic layers and one insulating layer. The compact model that is being employed here is made up of two ferromagnetic layers (CoFeB), one oxide barrier layer (MgO), and an Antiferromagnetic (AFM) layer (made of IrMn) placed on top of one another [11]. One of the layers, the free layer (FL), has a magnetization that may alternate between two stable orientations. In contrast, the other layer's magnetization is the fixed pinned layer (PL). Fig. 1 depicts the circuit diagram for a single-level cell. Depending on the direction of the FL magnetization, the MTJ might be parallel (P) or anti-parallel (AP), which correlates to resistances RP and RAP. The benefits of spin-orbit torque (SOT) effects and voltage control of magnetic anisotropy (VCMA) are combined in the VGSOT writing method. The applied positive VCMA voltage pulse at the PL of MTJ lowers the energy barrier (between P and AP states). It enables quick switching and decreased power usage. It enhances the lifespan of MTJ, decouples the read/write route, and offers reduced power consumption and speedier operation.

B. Literature Survey

Recent advancements in multilevel spintronics devices have primarily aimed to enhance storage density and cost-effectiveness. Multi-level Cell (MLC) techniques have been widely embraced for this purpose, and research has explored different avenues to achieve these goals.

One significant area of exploration involves leveraging high tunnel magnetoresistance (TMR) values to create distinct resistance states, enabling the growth of MLC techniques. STT-MRAM has received extensive attention as a candidate for MLC applications, particularly in embedded and on-chip scenarios [13]. However, increasing storage density in STTMRAM by merely stacking more MTJs and using STT switching schemes while preserving low power consumption, high speed, and reliability presents challenges [14]. In response, SOT-MRAM-based MLCs have emerged as a promising alternative.

These efforts focus on achieving low power consumption, high speed, and reliable MLCs [15-18]. Innovative approaches have been proposed to advance MLC technology further. Hybrid architectures, such as the Hybrid STT-SOT-based series triple-level cell (sTLC) [15], integrate multiple bits of data storage efficiently with a maximum of two writing steps. Additionally, designs incorporating high on/off ratio multi-level cells, achieved by integrating series MTJs with Schottky diodes, have been introduced to increase storage capacity [17]. SOT based MLC with any initialization proposed in [18]. A novel spintronic multilevel memory unit with a high on/off ratio designed for deep neural networks is introduced in [19]. Another avenue of research involves considering the combined effects of SOT and STT in the design of SOT-STT-based MLCs, making writing operations quicker and more energy-efficient than conventional MLCs. This approach includes a series triple-level cell (sTLC) architecture that maintains sufficient resistance margins between different resistance levels, necessary for representing various logic states. Furthermore, research efforts aim to achieve 4-bit data storage with improved latency and reduced energy consumption performance [20]. However, challenges persist, particularly in optimizing the read and write (R/W) current paths, given the varying current requirements in the mentioned MLC architectures. The application of high write voltages over the thin tunnel barrier of the MTJ during the writing process may introduce reliability concerns, especially in high-speed applications, resulting in additional energy penalties and area overhead. In summary, recent research in multilevel spintronics devices has explored diverse approaches, including SOT-MRAM, STTMRAM, and hybrid strategies, with the goal of enhancing energy and area efficiency in MLC operations. These efforts seek to strike a balance between high-density storage, low-power consumption, high speed, and reliability, paving the way for more efficient data storage solutions in the future.

III. EFFICIENT RRT-VGSOT- MULTI-LEVEL CELL

The main objective of this work is to design a 2-bit VGSOTMLC with reduced retention time that is optimal for area, energy, and read/write operation. Retention time modeling and an energy-efficient multilevel VGSOT architecture are discussed in this section.

A. Retention Time Modeling

The cross-sectional size of the HM and the distance between the MTJ are changed in the earlier research to achieve independent switching of two MTJs. This procedure, which is more crucial, delays the switching of the MTJ, reducing energy efficiency and jeopardizing reliability. An alternative strategy investigated in this study is to reduce the energy barrier between the pMTJ magnetization states by modulating their retention time.

Spintronics memory devices were initially designed to store data for up to 10 years without needing an external power source [12]. The retention time is the amount of time that data is kept in the device without power. Extended retention durations are unnecessary in on-chip caches because data is often only needed for less than one second [12]. Retention time, is exponentially correlated with the thermal stability factor through the following equation [21]:

$$t_{ret} = (1 / f_0) * e^{\Delta} \quad (1)$$

where f_0 is the frequency, which is roughly 1GHz. A thermal stability factor of more than 60 is required to satisfy the industry standard(memory devices) retention time of 10 years.

The Δ VGSOT device is given by [21]:

$$\Delta = \frac{\mu_0 M_s H_{Keff} t_F \pi r^2}{2 k_B T} \quad (2)$$

Where μ_0 is the permeability of the vacuum, k_B is the Boltzmann constant, T is the temperature, M_s is the saturation magnetization, H_{Keff} is the effective perpendicular anisotropy field (the external field in systems with uniaxial anisotropies of n-fold symmetry in a specific plane defined by the primary axis.), and t_F is the thickness of the free layer.

Reducing the thermal stability factor (Δ) facilitates fasttransition between P (parallel) and AP (antiparallel) states in MTJs. The Devices with shorter retention times have lower energy barriers compared to those with longer retention times. By adjusting the value of Δ, the VGSOT of a selected retention time can be modeled. The Δ value which is influenced by factors like junction diameter, oxide layer thickness, and free layer thickness. The choice of materials impacts key magnetic parameters such as saturation magnetization (M_s), effective anisotropy field (H_{Keff}), and damping constant (α). In VGSOT, external factors like doping and impurities affect the spin hall effect. Electron net polarization relies on these factors as well as geometric properties, free-layer thickness, planar area (A), and operating temperature (T) [22].

Furthermore, Δ in an MTJ is directly linked to the energy barrier ($Eb = \Delta\ k_B\ T$), where Eb is the energy barrier between stable states, k_B is the Boltzmann constant, and T is temperature. This implies that altering the free layer thickness can adjust Δ, potentially leading to a reduced energy barrier between P and AP states. The energy barrier primarily results from exchange interaction among magnetic layers and the shape anisotropy of the free layer, aligning the magnetic moment along its easy axis, often perpendicular to the layers' plane. Reducing Δ through adjustments in the free-layer thickness can lower the energy barrier, facilitating high speed switching [21]. This approach enables the modeling of MTJs with reduced retention times.

To summarize, the value Δ can be tuned to design a VGSOT with a specific retention time. Δ can be controlled by using physical parameters like the junction diameters and free-layer thickness of MTJ. Hence, a VGSOT with a reduced retention time can be designed by changing thermal stability values without any other structural modification.

B. Modeling Multilevel Reduced Retention Time VGSOT

This section discusses the design of a 2-bit storage memory cell using VGSOT with different retention times in a parallel configuration. Multilevel switching can be produced by including extra MTJs in the single-level bit cell. There are two design possibilities: 1) Series connection; 2) Parallel connection MTJ. The pMLC has opted for high-density and low-energy applications as compared to sMLC. Edge computing applications induce high latency. To reduce its latency by parallel architecture. So, in this design, the authors focused on parallel multilevel switching using VGSOT. The circuit schematic of the 2-bit VGSOT-pMLC is shown in Fig. 2. By virtue of permitting the use of PMTJ devices with smaller diameters and higher RA values, which reduce footprint area, the pMLC structure reduces the total resistance of a PMTJ stack.

The proposed design consists of MTJs with different retention times, 100 s (MTJ1-with Δ value in range of 35–40) & 10 s (MTJ2-with Δ

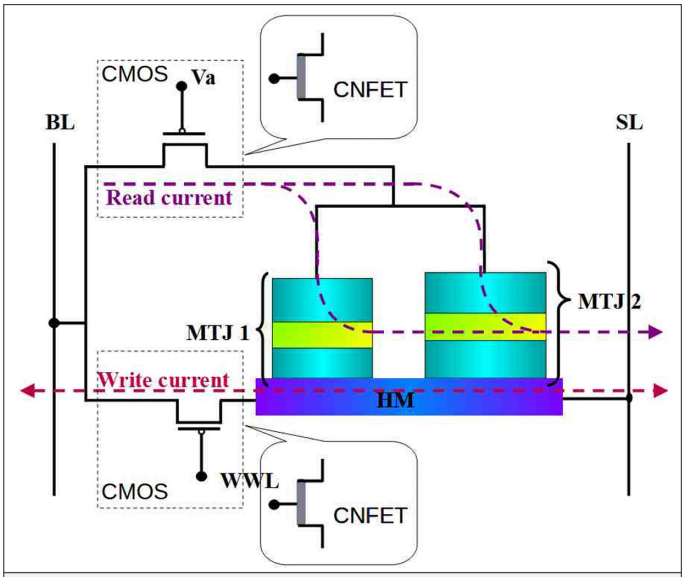

Fig. 2. Bit cell schematic of proposed reduced retention time multi-level of parallel VGSOT (VGSOT-MLC).

value in range of 28–32), respectively. MTJ1 represents the Least Significant Bit (LSB), whereas MTJ2 represents the Most Significant Bit (MSB). There will be two significant switching levels: 1) Soft transition, where either both bits have the same switching or only the LSB is switching (for example, "00" to "01," "00" to "11," or "10" to "11"), 2) Hard transition when just the MSB bit is shifted (for instance, "11" to "01" or "00" to "10" or "01" to "10"). A lower switching current is needed for a smooth transition, whereas a larger switching current is needed for a harsh transition. The MTJ1 has a long retention time (100s), which raises the switching current since the MSB bit demands a high current density. Conversely, the decreased retention time results in a lesser current for LSB (10 s). Present scenario, the primary design limitations for MTJs with variable retention times are; 1) optimization of the free layer thickness in both MTJs (100S, 10S), 2) optimization of MTJ dimensions, 3) optimization of enrgy barrier thickness to prevent read failure, and 4) tthe variance in the retention times of two different MTJs. The difference in switching current also grows if the retention time between two MTJs is significant. Write errors, process variance, and dependability concerns are therefore brought up.

The magnetic properties related to switching characteristics of MTJ in the VGSOT are evaluated using the object oriented micro magnetic framework simulations (OOMMF). This shows that the MLC below 100 nm is possible by carefully adjusting the material property. But the presence of transistors in the read-and-write path results in high leakage current due to the scaling down of the channel length. Low channel mobility and a low threshold voltage of the MOS transistor make the scenario worse. Even though the VGSOT avoids the bulk transistors, the MOS transistor used for read and write operations still results in short channel effects, high gate leakage, and low threshold voltage. These effects degrade the performance of VGSOT-MLC and negatively affect the area efficiency of the magnetic circuits [23].

To mitigate this problem, the authors propose to replace the read-and-write transistors with carbon nanotube field effect transistors (CNFET) in this design. Because of its exceptional electrical and thermal characteristics, the CNFET has been seen as a possible alternative to metal-oxide-semiconductor field-effect transistor (MOSFET)[24].

Furthermore, because of its programmable threshold voltage, CNTFET is an ideal technology for constructing MLC circuits that present data at several voltage levels. The MTJs are manufactured independently in a layer above the transistors, facilitating CNFET integration. A recently developed gate-all-around CNTFET model by Stanford University [25-26] is used for the proposed MLC. The paper details a 2-bit storage circuit that is both energy and space efficient. It consists of a VGSOT-MLC device with two MTJs that are not identical and uses CNFETs as access transistors (T) for read-and-write processes. In Fig. 2, the structure is depicted.

C. System-level Performance of RRT-VGSOT-Cache

The previously proposed compact model characteristics are used in this evaluation [27]. The impact of reduced retention time in VGSOT-based last level cache (L2) in a unicore system has been studied using the Gem5 simulator [28]. The parameters of L2 are modified based on RRT-VGSOT using the NVSim simulator [29]. When a cache block's retention duration exceeds the predicted value, it is invalidated and causes a miss on the next access. The effect of using an L2 cache with different retention times values is demonstrated in Fig. 3 on the miss rate of applications from the SPEC CPU2006 test suite [30]. Fig. 3 shows the normalized L2 miss rate of the SRAM-based L2. Computationally intensive Calculix and h264ref have the lowest miss rate when the retention time of L2 is 10 s.

Figs. 4 and Fig. 5 provide the normalized write energy and write latency of VGSOT, STTRAM, and SOTRAM w.r.t SRAM for different retention times. The figures show that even while VGSOT has substantially greater write energy and write latency than SRAM, the difference becomes less noticeable as retention time decreases. VGSOT outperforms STTRAM, SOTRAM, and shortened retention time STTRAM when comparing write energy and write latency.

D. The Effect of Temperature on Reduced Retention Time

The VGSOT cell is simulated using a 45-nm CMOS process within the Cadence Spectre environment. To accurately capture the MTJ cell's behavior, a previously proposed compact model is employed, which considers both field-like and damping-like torques [27]. This model accounts for the three-terminal MTJ device's operation. By utilizing this compact model, temperature and parameter variations can be analyzed for both magnetic and electric responses of the MTJ cell. This approach provides a holistic insight into the device's behavior across different conditions. The designed free layer-tuned VGSOT forms a pivotal element in MLC (Multi-Level Cell) circuits. An optimization process is carried out on the free layer's geometric parameters in order to enhance MLC functionality and energy efficiency. To investigate the impact of temperature on reduced retention time in VGSOT-MLC, temperature variations ranging from −50 °C to 100 °C are examined through DC analysis conducted using Cadence Virtuoso.

Fig. 6 depicts the cumulative distribution function of retention time for different temperatures. This demonstrates that at higher temperatures, the retention time is significantly decreased from its optimal value, while at lower temperatures, it is significantly increased. The temperature inversely affects thermal stability and accelerates the oxide breakdown. Dependence of write error rate (WER) on write voltage at different retention times VGSOT-MLC,

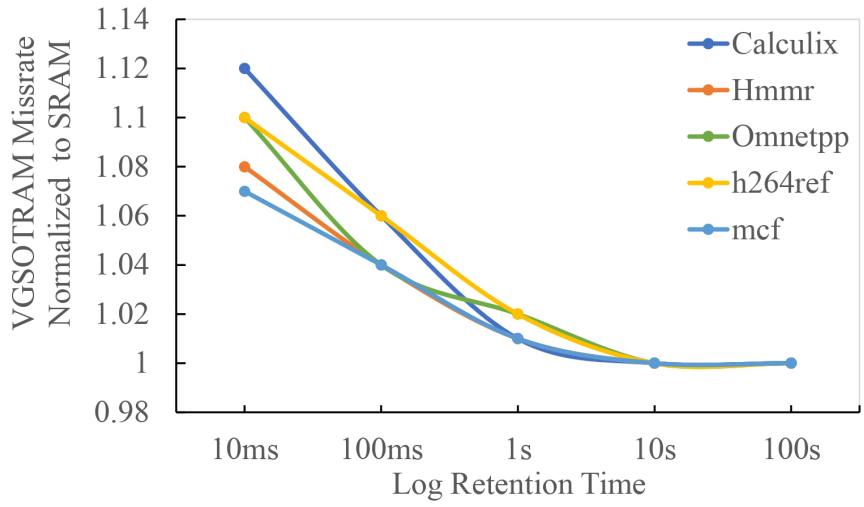

Fig. 3. L2 miss rate of VGSOT versus retention time normalized w.r.t. SRAM.

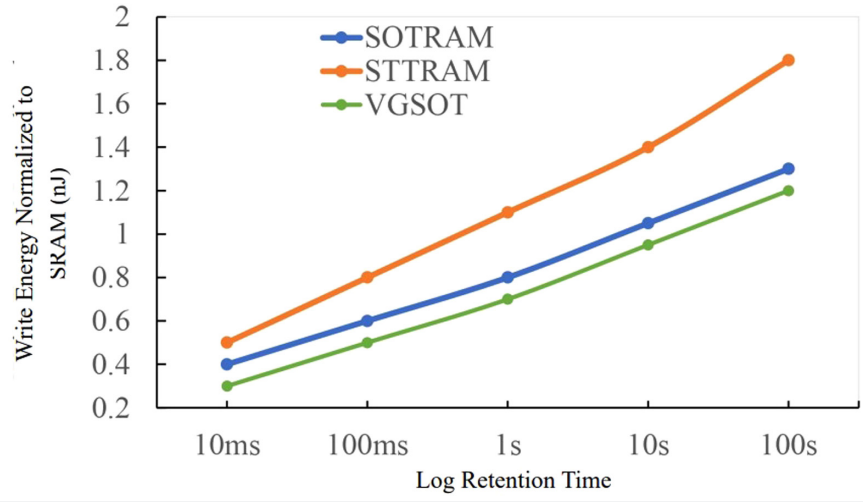

Fig. 4. Write energy of SOTRAM, STTRAM, and VGSOT for different values of retention time.

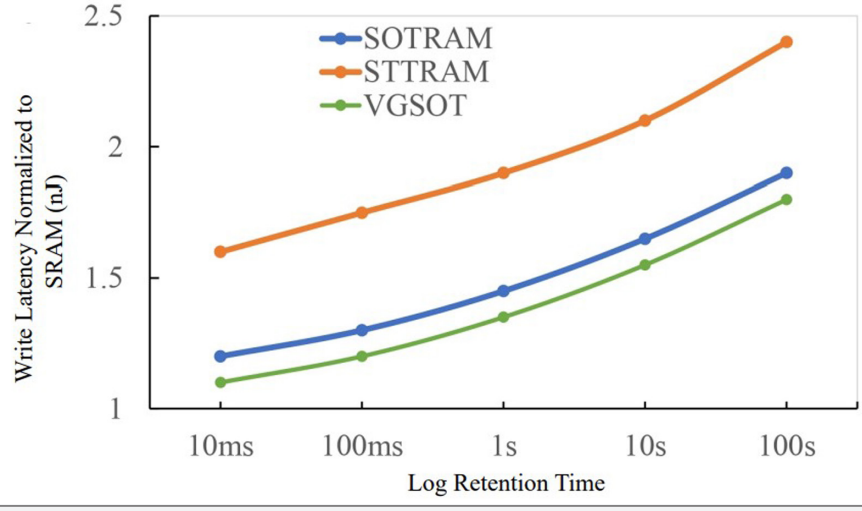

Fig. 5. Write latency of SOTRAM, STTRAM, and VGSOT for different values of retention time.

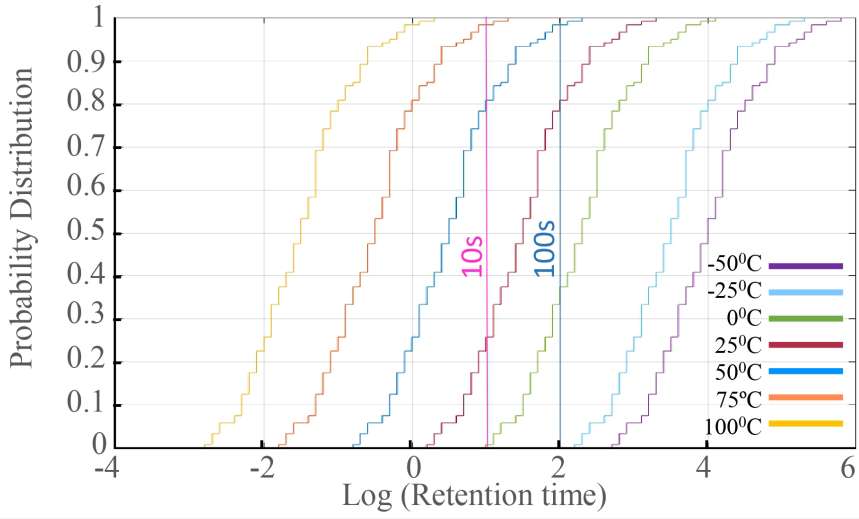

Fig. 6. CDF of retention time for VGSOT with retention time of 100 s and 10 s.

extracted from 5000 Monte-Carlo simulation runs as shown in Fig. 7. It depicts that the WER increases with decreasing retention time. Retention time reduction might increase the impact of thermal fluctuations during switching, which can result in writing errors. Retention time is decreased, which lowers the switching current and cuts down on switching energy and time. Yet the impacts of process variation, temperature fluctuations, and WER have a limit on this decrease in retention time. In exchange, VGSOT with varying retention durations between 10 and 100 seconds is appropriate for the multilevel design.

IV. EXPERIMENTAL RESULTS

This section presents the performance analysis of the optimized 2-bit VGSOT-pMLC. And also investigates how the performance of the pMLC designs is affected by process variation, switching current, and writing energy. Meanwhile, the state-of-the-art devices proposed using STTMRAM [13] SOTMRAM [16], and VGSOT [11] are employed to compare.

A. Simulation Framework

This section elaborates on the simulation framework used for device-to-architecture analysis. MTJ Device-level Modeling: In authors' investigation, the VGSOT cell was subjected to simulation within the Cadence Spectre environment, employing a 45-nm CMOS process. To accurately emulate the behavior of the MTJ cell, the authors utilized a previously proposed compact model that comprehensively encompasses both field-like and damping-like torques. Detailed information about this model can be found in reference [27]. The key parameters of the devices are tabulated in Table I. This model has been intricately designed to capture the functionality of a three-terminal MTJ device. Through the utilization of this compact model, the authors are able to explore the impacts of reduced retention time.

System-level Evaluation: For assessing the performance of the RRT-VGSOT in the last-level cache (L2-64kB), the authors employed the architectural simulator gem5 [28]. This allowed us to evaluate the energy benefits and the subsequent effects on application-level

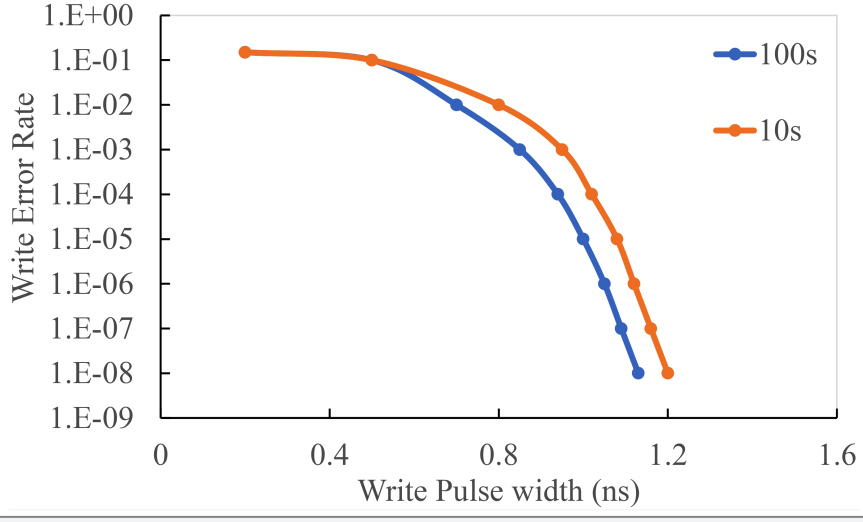

Fig. 7. Variation of WER with the retention time 100 s and 10 s respectively.

TABLE I. THE CRITICAL PARAMETERS OF THETHE VGSOT [27]

Gilbert Damping, a	0.0122
Sat. Magnetization, M_s	900×10^3 A/m
Dimension of free layer ($W_{FL} \times L_{FL} \times t_{FL}$)	120nm × 40 nm × 2 nm
Dimension of HM ($W_{HM} \times L_{HM} \times t_{HM}$)	120 nm × 80 nm × 2 nm
VCMA coefficient	60Fj/(V.m)
Spin polarization	0.58
Exchange bias	−50Oe
HM resistivity	200 mΩ.cm
Spin Hall angle, Θ_{SH}	0.3
Oxide thickness, t_{OX}	<1 nm
Critical current (7ns)	42 mA

quality. Within this evaluation, the authors considered three distinct retention times (1 s, 10 s, 100 s) for the L2 cache. The authors assessed the performance of on-chip cache by simulating computationally intensive applications from the SPEC 2006 benchmark suite. For estimating energy-related parameters of the L2 Cache, the authors utilized the modified NVSim simulator [29].

B. VGSOT-MLC Read and Write Operations

A two-step process is needed to read the value stored in a 2-bit MLC-VGSOT cell. The read operation involves measuring the resistance of an MLC Magnetic tunnel junction (MTJ) and comparing it to three resistance references, generated using two reference memory cells each. The sense amplifier initially compares the cell's resistance with Ref0 to ascertain the hard bit, and then compares it with either Ref1 or Ref2 to determine the soft bit. This two-step reading results in a read latency for an MLC cell that is 1.5 times longer than that of a single-level cell (SLC)[31].

Writing a 2-bit value into the cell is more complex due to the involvement of two magnetic tunnel junctions (MTJs) controlled by a single access transistor. This arrangement means that the write current always passes through both MTJs simultaneously [32]. Since the current needed to switch the more stable bit (MSB) is higher than that required for the less stable bit (LSB), changing the MSB invariably alters the LSB to have the same magnetic field, resulting in either "00" or "11" being written. In cases where the LSB differs from the MSB,

A smaller write current is applied in the second step to switch the LSB. The time taken to write each bit is comparable to that of single-level cell (SLC) writing, causing the latency of a 2-bit multi-level cell (MLC) write to roughly double that of an SLC write. However, it's worth noting that an MLC cell write can sometimes terminate early. This happens when either the LSB and MSB are the same, allowing the second step to be skipped, or when only the LSB needs to be changed while the MSB remains the same, allowing the first step to be skipped [33].

Concerning MLC based on VGSOT, the write-disturb failure issue is a critical concern. It occurs when writing to one memory cell unintentionally impacts another nearby cell, leading to incorrect data storage. This problem is effectively addressed in prioritized multi-level cell (P-MLC) design through a specific write sequence [34, 35]. In P-MLC design, the key strategy is to prioritize the writing of one memory cell over the other. For example, when dealing with two memory cells, MTJ1 and MTJ2, the write process starts by writing to MTJ2 first. Even if MTJ2 experiences disturbance during this initial write (due to shared write paths), the correct data for MTJ2 can be written in the subsequent cycle. This sequential approach ensures that each memory cell ultimately receives the correct data through a two-step write operation [35].

This strategy significantly reduces the likelihood of write-disturb failures, making P-MLC design a reliable solution for multi-level data storage in magnetic memory devices, even though it may have a slightly longer write latency compared to SLC [36].

C. Performance of VGSOT-pMLC

This section discusses various parameters associated with read and write to the proposed VGSOT-MLC. Fig. 8 and Fig. 9 demonstrate how the write current and write energy for switching are both decreased due to the decreased retention time. Inadequate write current causes errors and increases the likelihood of writing unwanted data. The write current of the current logic state is controlled by the critical switching current of the next logic level due to the multiple switching thresholds in the pMLC design. Higher write margins are mandatory to reduce unconditional flipping in the writing process.

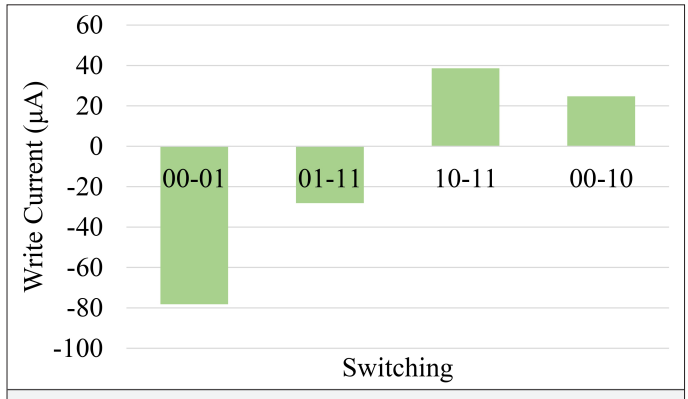

Fig. 8. Write and read performance of 2-bit VGSOT-pMLC with respect to the write current.

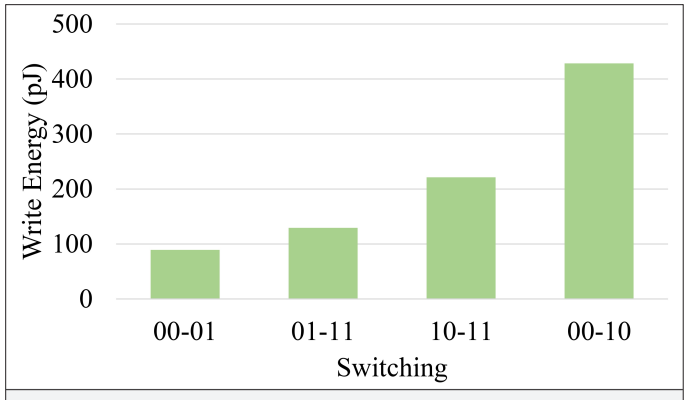

Fig. 9. Write and read performance of 2-bit VGSOT-pMLC with respect to the write energy.

Choosing the different retention times for MTJs have an adequate distinction between switching thresholds to obtain a higher write margin.

To identify all potential resistance states, a unidirectional read current that is less than the critical switching current is needed. To achieve minimal Read disturbance error, the read current in this architecture is selected to flow in the P-AP direction. The normalized difference between the read current and the critical switching current is known as the read margin (RM). The read current needs to be more than 20 uA in order to be reliable. If the read current is less than 20 uA, the sense amplifier might not read the proper logic value because of process variations. This design, the 2-bit pMLC design, is optimized with a read current more significant than 25 uA for a reliable and faster read operation. The Fig. 10 and Fig. 11 depict the Read current and read margin for the different switching levels.

Due to the unpredictable temperature fluctuation in the MTJ, the write operation in VGSOT is inherently stochastic. VGSOT offers a lower error rate than other spintronics counterparts due to its voltage controllability. Configuring the voltage levels with lower rise and fall times leads to better WER.

In addition, research on the impact of retention time on WER demonstrates that lower WER is seen with longer retention times. Additionally, despite the fact that the MTJ with a retention time of

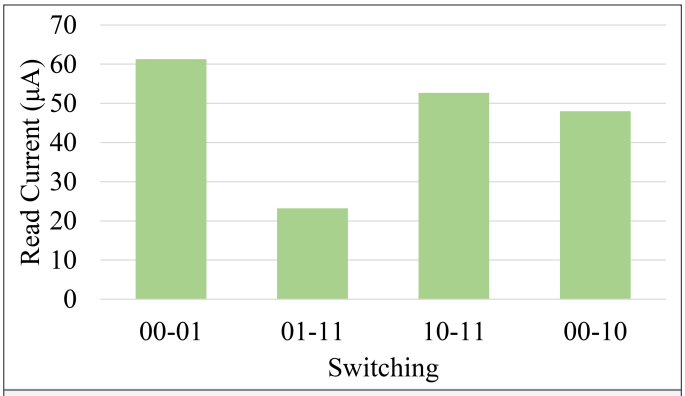

Fig. 10. Write and read performance of 2-bit VGSOT-pMLC with respect to the read current.

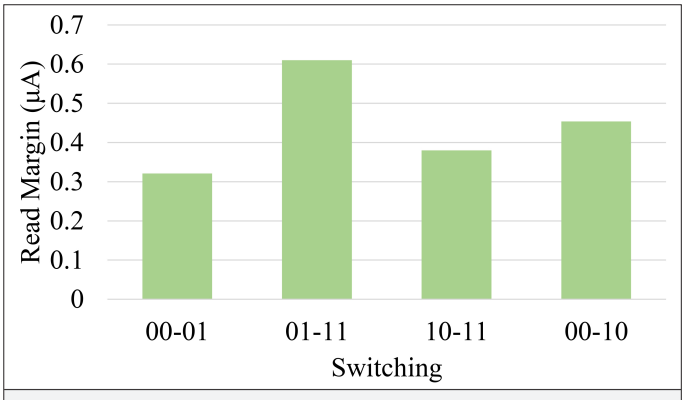

Fig. 11. Write and Read performance of 2-bit VGSOT-pMLC with respect to the read margin.

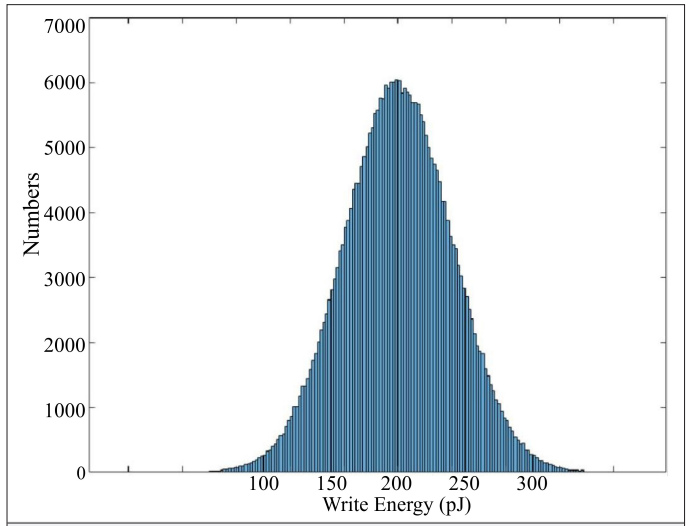

Fig. 12. Monte-Carlo simulations results for VGSOT-MLC write operation due to process variation.

seconds has precise writing, the error rates are acceptable level. The impact of temperature variations during the switching operation can be reduced by adjusting the damping factor values and changing the materials for MTJ. The process variations significantly impact the electrical and magnetic properties of devices. The vital components of process variations are tunnel free layer barrier, surface area, access transistor variations, MTJ shape variations, and fluctuations of magnetic anisotropy [30]. These variations in physical parameters also influence retention time. However, the write and read current variations and interconnect parasitic of spintronics devices lies within 10%. The Monte Carlo analysis also includes a bit line voltage variation of 10%. Monte Carlo simulations with 3 σ of 5% (MTJ dimensions free layer thickness) and average retention times in the 50 s and 100 s are used to examine the impact of process variation. As illustrated in Fig. 12, these variables are investigated for the resulting variance in the write energy. The logic bit "00" is assumed to have the fewest conflicts. These tests demonstrate that the MTJ retention time can vary according to process variation.

D. Comparison of VGSOT-pMLC with Pioneering MLC Architectures

The VGSOT cell may function in a high-frequency regime owing to the spin hall effect and VCMA. VGSOT with different retention times requires a small current for a write operation. The decoupled read-and-write path improves the lifetime of MTJ. The CNFET compensates for the flaws of the conventional transistors in scaling, leakage power, and threshold constraints. Review and compare the existing technology to verify the proposed design's performance. Comparing the performances of the proposed pMLC-VGSOT with the current technologies concerning the switching current and write energy are summarized in Fig. 13 and Fig. 14. Study shows that the suggested design's critical current for switching is 50% lower, and its write energy is 42% lower than that of STT-MLC. Compared to SOT-pMLC, the critical current for switching and write energy is reduced to 30% and 15%, respectively. The proposed pMLC also employs two parallel MTJs per cell without adding bit-cell area overhead, resulting in a 100% increase in integration density. VGSOT-pMLC with CNFET will consume 12% of the average write energy of VGSOT-pMLC with CMOS transistors. The MTJ with a small, different retention time can

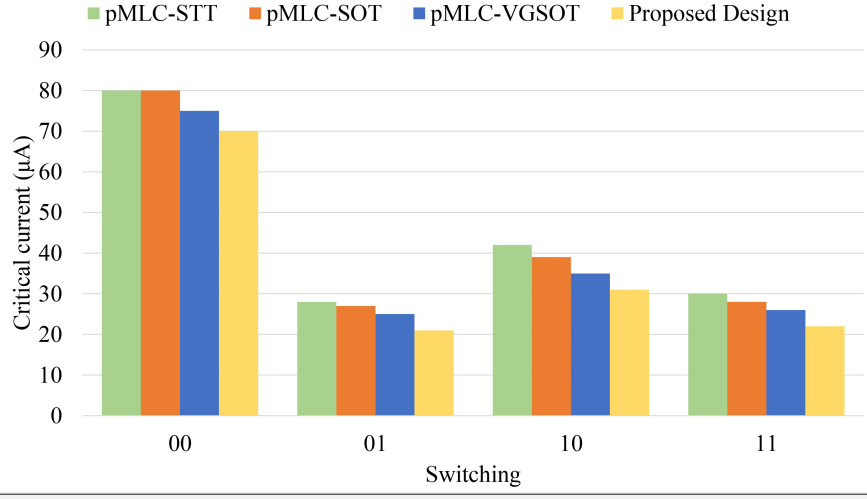

Fig. 13. Performance comparison of the proposed design with STT-pMLC, SOT-pMLC, VGSOT-pMLC with respect to write energy.

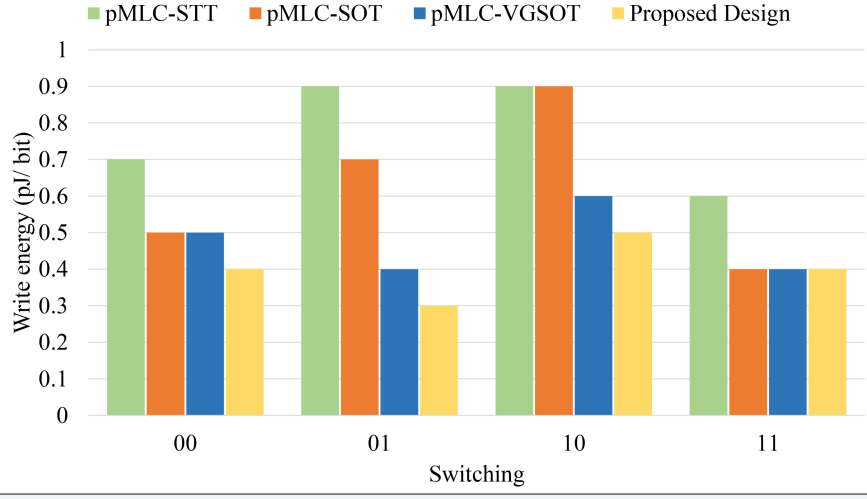

Fig. 14. Performance comparison of the proposed design with STT-pMLC, SOT-pMLC, VGSOT-pMLC with respect to critical switching current.

compensate for the area overhead due to the separate read and write path. From the analysis proposed energy-efficient design is compactable for on-chip cache designs.

V. CONCLUSION

The paper presents parallel MLC designs with different retention time MTJs to enhance the integration density of VGSOT. With the two parallel MTJs per cell employed by the suggested pMLC design, integration density is increased without experiencing bit-cell area overhead. Utilizing the low switching current feature of the reduced retention time MTJ and threshold voltage tunability of the GAA-CNTFETs, an energy-efficient VGSOT-pMLC is designed. CNFET-based spintronics devices outperform the CMOS-based designs in energy consumption and performance. By tuning the geometrical and physical parameters of MTJ, the authors can configure application-specific retention time for MTJ. The energy-efficient multilevel switching feature is added without any additional sense amplifier and bistable feedback. For a 2-bit MLC memory cell, the proposed design shows an average of 50% and 42% reduction in critical switching current and write energy, respectively. This enables using spintronics-based MLC in high-performance and low-power computing applications such as neuromorphic and IoT devices. In the future, the authors intend to work on a more practical adaptive retention-based MLC designs.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – P.B., S.W.; Design – P.B., S.W.; Supervision – S.W.; Resources – P.B.; Materials – P.B.; Data Collection and/or Processing – P.B., S.W.; Analysis and/or Interpretation – P.B., S.W.; Literature Search – P.B.; Writing – P.B., S.W.; Critical Review – P.B., S.W.

Declaration of Interests: The authors have no conflict of interest to declare.

Funding: The authors declared that this study has received no financial support.

REFERENCES

1. Z. Zheng *et al.*, "Enhanced spin-orbit torque and multilevel current-induced switching in W/Co– Tb/Pt heterostructure," *Phys. Rev. Appl.*, vol. 12, no. 4, p. 044032, 2019. [CrossRef]

2. S. Ning, H. Liu, J. Wu, and F. Luo, "Challenges and opportunities for spintronics based on spin orbit torque," *Fundam. Res.*, vol. 2, no. 4, pp. 535–538, 2022. [CrossRef]
3. Y. Chen, W. F. Wong, H. Li, C. K. Koh, Y. Zhang, and W. Wen, "On-chip caches built on multilevel spin-transfer torque RAM cells and its optimizations," *ACM J. Emerg. Technol. Comput. Syst. (JETC)*, vol. 9, no. 2, pp. 1–22, 2013. [CrossRef]
4. 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, 2020. [CrossRef]
5. A. Nisar, H. Nehete, G. Verma, and B. K. Kaushik, "Hybrid multilevel STT/DSHE memory for efficient CNN training," *IEEE Trans. Electron Devices*, vol. 70, no. 3, pp. 1006–1013, 2023. [CrossRef]
6. K. Ali, F. Li, S. Y. Lua, and C.-H. Heng, "Area efficient shared diode multilevel cell sot-mram," *IEEE Trans. Magn.*, vol. 54, pp. 1–5, 2018.
7. V. Nehra, S. Prajapati, T. N. Kumar, and B. K. Kaushik, "High performance computing-in-memory architecture using stt-/sotbased series triple-level cell mram," *IEEE Trans. Magn.*, vol. 57, pp. 1–12, 2021.
8. A. Sehgal, K. K. Das, S. Dhull, S. Roy, and B. K. Kaushik, "Variability analysis of multilevel spin-orbit torque MRAMs using machine learning," In 23rd International Conference on Nanotechnology (NANO), vol. 2023. IEEE Publications, 2023, pp. 793–797. [CrossRef]
9. 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, 2020. [CrossRef]
10. A. Nisar, S. Dhull, B. K. Kaushik, and S. Mittal, "High-performance voltage controlled multilevel MRAM cell," *Semicond. Sci. Technol.*, vol. 36, no. 12, p. 125013, 2021. [CrossRef]
11. S. Shreya, and B. K. Kaushik, "Modeling of voltage-controlled spin–orbit torque mram for multilevel switching application," *IEEE Trans. Electron Devices*, vol. 67, no. 1, pp. 90–98, 2019. [CrossRef]
12. D. Gajaria, K. Kuan, and T. Adegbija, "SCART: Predicting STT-RAM cache retention times using machine learning," In Tenth International Green and Sustainable Computing Conference (IGSC), vol. 2019. IEEE Publications, 2019, pp. 1–7. [CrossRef]
13. S. Prajapati, and B. K. Kaushik, "Area and energy efficient series multilevel cell stt-mrams for optimized read–write operations," *IEEE Trans. Magn.*, vol. 55, no. 1, pp. 1–10, 2018. [CrossRef]
14. S. Prajapati, and B. K. Kaushik, "Parallel multilevel cell stt-mrams for optimized area energy and read–write operations," *IEEE Trans. Magn.*, vol. 54, pp. 1–9, 2018.
15. V. Nehra, S. Prajapati, T. N. Kumar, and B. K. Kaushik, "Highperformance computing-in-memory architecture using stt-/sotbased series triple-level cell mram," *IEEE Trans. Magn.*, vol. 57, pp. 1–12, 2021.
16. 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, 2015. [CrossRef]
17. K. Ali, F. Li, S. Y. H. Lua, and C.-H. Heng, "Area efficient shared diode multilevel cell SOT-MRAM," *IEEE Trans. Magn.*, vol. 54, no. 11, 3401105, 2018. [CrossRef]
18. S. Zhang et al., "Spin-orbit-torque-driven multilevel switching in Ta/CoFeB/MgO structures without initialization," *Appl. Phys. Lett.*, vol. 114, no. 4, 042401, 2019. [CrossRef]
19. K. Zhang *et al.*, "High on/off ratio spintronic multilevel memory unit for deep neural network," *Adv. Sci. (Weinh)*, vol. 9, no. 13, p. e2103357, 2022. [CrossRef]
20. A. Nisar, S. Dhull, S. Mittal, and B. K. Kaushik, "Sot and stt-based 4bit mram cell for high-density memory applications," *IEEE Trans. Electron Devices*, vol. 68, no. 9, pp. 4384–4390, 2021. [CrossRef]
21. P. B. Alisha, and T. S. Warrier, "Optimizing free layer of Magnetic Tunnel Junction for true random number generator. Memories-materials, devices," *Circuits Syst.*, vol. 5, p. 100075, 2023.
22. B. Han, B. Zhang, S. Sun, B. Wang, Y. Guo, and J. Cao, "The thickness dependence of the field-like spin–orbit torque in heavy metal/cofeb/mgo heterostructures," *J. Appl. Phys.*, vol. 130, no. 21, p. 213902, 2021. [CrossRef]
23. S. Pathak, C. Youm, and J. Hong, "impact of spin-orbit torque on spin-transfer torque switching in magnetic tunnel junctions," Scientific reports 10, pp. 1–8, 2020.
24. F. Razi, M. H. Moaiyeri, R. Rajaei, and S. Mohammadi, "A variation-aware ternary spin-Hall assisted STT-RAM based on hybrid MTJ/GAA-CNTFET logic," *IEEE Trans. Nanotechnol.*, vol. 18, pp. 598–605, 2019. [CrossRef]
25. J. Deng, and H.-S. P. Wong, "Compact SPICE model for carbon-nanotube field-effect transistors including nonidealities and its application - Part II: Full device model and circuit performance benchmarking," *IEEE Trans. Electron Devices*, vol. 54, no. 12, pp. 3195–3205, 2007. [CrossRef]
26. F. Razi, M. H. Moaiyeri, R. Rajaei, and S. Mohammadi, "A variation-aware ternary spin-Hall assisted STT-RAM based on hybrid MTJ/GAA-CNTFET logic," *IEEE Trans. Nanotechnol.*, vol. 18, pp. 598–605, 2019. [CrossRef]
27. K. Zhang, D. Zhang, C. Wang, L. Zeng, Y. Wang, and W. Zhao, "Compact modeling and analysis of voltage-gated spin-orbit torque magnetic tunnel junction," in *IEEE Access*, vol. 8, pp. 50792–50800, 2020. [CrossRef]
28. N. Binkert *et al.*, "The gem5 simulator," *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, 2011. [CrossRef]
29. X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Trans. Comput. Aid. Des. Integr. Circuits Syst.*, vol. 31, no. 7, pp. 994–1007, 2012. [CrossRef]
30. J. L. Henning, "Spec cpu2006 benchmark descriptions,", *SIGARCH Comput. Archit. News*, vol. 34, no. 4, pp. 1–17, 2006. [CrossRef],"
31. W. Kang, Y. Ran, Y. Zhang, W. Lv, and W. Zhao, ''Modeling and exploration of the voltage-controlled magnetic anisotropy effect for the next-generation low- power and high-speed MRAM applications,'' *IEEE Trans. Nanotechnol.*, vol. 16, no. 3, pp. 387–395, 2017. [CrossRef]
32. H. Zhang *et al.*, "Spintronic processing unit within voltage-gated spin Hall effect MRAMs," *IEEE Trans. Nanotechnol.*, vol. 18, pp. 473–483, 2019. [CrossRef]
33. K. Zhang, D. Zhang, C. Wang, L. Zeng, Y. Wang, and W. Zhao, "Compact modeling and analysis of voltage-gated spin-orbit torque magnetic tunnel junction," *IEEE Access*, vol. 8, pp. 50792–50800, 2020. [CrossRef]
34. Y. C. Wu *et al.*, "Voltage-gate-assisted spin-orbit-torque magnetic random-access memory for high-density and low-power embedded applications," *Phys. Rev. Appl.*, vol. 15, no. 6, p. 064015, 2021. [CrossRef]
35. A. Lee *et al.*, "Low-energy shared-current write schemes for voltage-controlled spin-orbit-torque memory," *IEEE Trans. Electron Devices*, vol. 70, no. 2, pp. 478–484, 2023. [CrossRef]
36. A. Nisar, S. Dhull, B. K. Kaushik, and S. Mittal, "High-performance voltage controlled multilevel MRAM cell," *Semicond. Sci. Technol.*, vol. 36, no. 12, p. 125013, 2021. [CrossRef]

Alisha P B received a BTech degree in electronics and communication engineering from Cochin University of Science and Technology (CUSAT) in 2014 and an MTech degree in wireless technology from CUSAT in 2016. She is currently pursuing a PhD degree in electronics specializing in spintronics from CUSAT, Kerala, India. Her research interests include an experimental demonstration of novel spin-based devices for unconventional Computing.

Dr. Tripti S Warrier is an Assistant Professor at the Department of Electronics, Cochin University of Science and Technology. She received her PhD from the Department of Computer Science, Indian Institute of Technology Madras, India. Her industrial experience includes her tenure at the Naval Physical and Oceanographic Laboratory and Wipro Technologies both at Cochin, India. Her areas of interest are Domain specific RISC-V Architectures, Hardware acceleration for AI, Memory Systems- Non-Volatile memory designs, FPGA Based System Design, Digital System Design and Verification. She is the recipient of funding from Indian government agencies like the Department of Science and Technology, Ministry of Electronics and Information Technology and RUSA.