Energy Efficient Reduced Area Overhead Spin-Orbit Torque Non-Volatile SRAMs

Karim Ali [(1)], Fei Li [(2)], Sunny Y.H. Lua, and Chun-Huat Heng [(1)]
[(1)] National University of Singapore, ECE
[(2)] Institute of Microelectronics, Agency for Science, Technology and Research (A*STAR), Singapore

***Abstract*— This paper proposes two spin orbit torque nonvolatile static random-access memories (SOT-NVSRAMs) with reduced area overhead. One of the proposed cells comprises nine transistors (9T) and a pair of complementary SOT-MTJs and the other comprises seven transistors (7T), two MOM diodes stacked over a pair of complementary SOT-MTJs. The proposed 7T cell has only one overhead transistor compared to the standard 6T SRAM cell, achieving at least 50% reduction compared to other SOT-NVSRAM designs in the literature. Moreover, the always complementary MTJs aid to increase the read margin and speed, leading to at least 20% lower restore energy of our proposed 9T cell compared to others. Furthermore, the two SOT-MTJs are written serially resulting in at least 32% lower store energy of both designs than the counterparts in the literature. In addition, the two SOT-MTJs are totally disconnected from the 6T cell in the nominal SRAM operation, which aids in maintaining the conventional SRAM high performance. Our proposed cell achieves at least 2.2x and 1.9x improvement in a figure-of-merit defined as energy and area product compared to its counterparts in the literature.**

***Index Terms*— Spintronic memory and logic, magnetic random-access-memory, NVSRAM, spin-orbit torque.**

I. Introduction

Artificial Intelligence of Things (AIoT) brings intelligence to millions of nodes connected to the Internet of Things (IoT) infrastructure. These smart nodes could include sensors, wearables and hearables and mobile devices deployed at the edge. These small-scale systems implement low depth convolutional neural networks (CNN) due to limited hardware resources such as power and area budget available. Most implementations of CNN accelerators [15-17] require large amount of static random-access memory (SRAM) as on-chip weight storage into which network weights are loaded from off-chip storage such as dynamic random-access memory (DRAM) or non-volatile memory (NVM) at the system startup. Such data movement has been proved to be more energy consuming than computation [18]. Besides, the volatile nature of SRAM requires constant power supply at nominal or lowered values to sustain stored information resulting in significant static power consumption in both active and standby operations [1]. The relentless scaling of CMOS technology nodes increases the transistor leakage current, which worsens the issue further [2].

Power gating (PG) [3] has been employed to circumvent the increasing leakage problem, where the inactive modules are disconnected from the power supply to eliminate its leakage contribution. This could be applied for the entire memory during standby durations or to specific idle memory modules during the active durations. However, due to the volatility nature of the purely CMOS based SRAM, the existing stored content of the power gated blocks need to be first backed up to a non-volatile storage. This is to ensure that the correct data can be restored back to these modules once they are reactivated. Unfortunately, the non-volatile storage is usually off-chip, which incurs significant data path delay and energy consumption. This reduces the possible power gating duration and minimizes the overall PG efficiency.

Another approach is to employ a non-volatile memory (NVM) such as magnetic random-access memory (MRAM) or resistive RAM (ReRAM) instead of the conventional SRAM [4]. However, currently, the write energy for these technologies is order of magnitudes higher than the conventional SRAM, which renders this approach energy inefficient. Thus, it has been explored to incorporate a non-volatile memory device (NVMD) within the SRAM cell to implement a non-volatile SRAM (NVSRAM), where the NVMD is 3D stacked over each SRAM cell on the same chip. Various non-volatile devices such as ReRAM, phase change memory (PCM) and magnetic tunnel junctions (MTJ) are compatible with CMOS technology, and have been employed to implement the corresponding NVSRAM technology [5, 6]. This approach allows fast store/restore operation for the required data from/to the volatile SRAM cells to/from the corresponding stacked NVMD in parallel manner. This reduces the energy overhead (OH) and permits applying PG with finer grain to further reduce the leakage contribution. Furthermore, this approach maintains the SRAM required high performance, especially if the employed NVSRAM uses the NVMD in such a way that it does not interfere with SRAM nominal operation.

Among the various NVSRAM proposals, MTJ based solutions are the stand-out option due to its multiple advantages compared to other technologies. These advantages include high endurance, lower write energy, smaller resistance distribution, faster switching speed, better scalability, long state retention time and higher compatibility with CMOS process. These advantages hold true for both MTJ writing technologies such as spin transfer torque (STT) and spin orbit torque (SOT). Between the two, SOT-MTJ offers more benefits than STT-MTJ [7]. Firstly, it has higher energy efficiency due to the faster switching speed and lower critical current requirement. Secondly, the write and read paths are separated, which aids in achieving a more optimum design. Finally, the write current

This work was supported by the SpOT-LITE programme (A*STAR Grant No. A18A6b0057) through RIE2020 funds from Singapore.

flows through the low resistance heavy metal (HM) electrode rather than the MTJ itself. This permit using lower write voltages and improves the device reliability. Thus, SOT-MTJ is the stand-out solution to achieve energy efficient, reliable and high performance NVSRAM.

Different SOT-NVSRAM designs have been reported in the literature [1, 8]. [1] proposed an 11T SOT-NVSRAM that is capable of fully disconnecting the SOT-MTJs from the conventional 6T SRAM under nominal operation. Whereas, [8] proposed an 8T SOT-NVSRAM that achieves reduced area overhead. However, both designs suffer from few drawbacks that results in higher energy consumption or larger area overhead as further elaborated in section II.

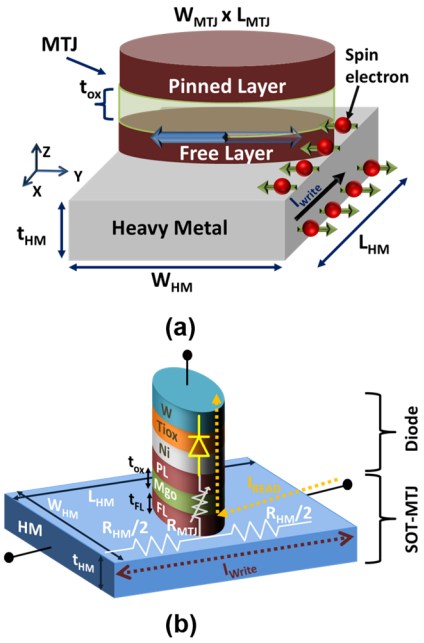

Fig. 1 (a) Spin-orbit torque magnetic tunnel junction (SOT-MTJ). (b) Diode based SOT-MTJ (D-SOT-MTJ), where a MOM diode is stacked over the SOT-MTJ in 3D manner.

In this paper, we propose an energy efficient 9T SOT-NVSRAM with reduced area overhead that avoids the above-mentioned issues. The proposed design comprises the conventional 6T SRAM cell with additional three transistors and two complementary SOT-MTJs. The proposed design has only one added transistor along both write and read paths of the SOT-MTJs, which reduces their path resistance and energy consumption. In addition, both SOT-MTJs are written with identical write current and voltages, which results in symmetrical switching and smaller store energy. Furthermore, the SOT-MTJs can be totally disconnected from the volatile SRAM cell during nominal operation to preserve SRAM performance. This is possible because the OH transistors are not embedded within the 6T SRAM cell, which minimizes their parasitic impact. Furthermore, a 7T SOT-NVSRAM that relies on a diode-based SOT-MTJ to reduce the area OH is also proposed.

The paper is organized as follows: Section II describes the SOT-MTJ fundamentals and operation. Section III presents our proposed SOT-NVSRAM and illustrates its operation. Thereafter, the simulation and evaluation of our proposed cell is provided in Section IV. Finally, the paper in concluded in Section V.

II. Review

A. Spin-Orbit Torque Magnetic Tunnel Junction (SOT-MTJ)

A SOT-MTJ, shown in Fig.1 (a), is composed of an MTJ that is placed in-contact with a HM electrode from the MTJ's free layer (FL) side. An MTJ [9] consists of a tunneling oxide barrier sandwiched between two ferromagnetic layers (FM), namely, the pinned layer (PL) and the free layer (FL). The PL has a pinned magnetization direction while the FL magnetization can be programmed to be either in parallel (P) or anti-parallel (AP) to the PL. The MTJ exhibits low and high resistance states in P and AP configurations which can be detected through peripheral circuits to represent logic '1' and '0'. The FL magnetization can be reversed using the SOT effect [10] as the FL is placed in contact with the HM electrode. When a charge current flows through the HM, spin-orbit interaction induces spin accumulation in the FL which assists in the change of the FL magnetization to either P or AP state depending on the current direction relative to the FL easy axis. A HM electrode with large spin-orbit interaction and low resistivity is used (e.g. W, Td, Pt) as it allows an injection of spin-polarized current that can be even larger than the supplied charge current (spin current / charge current >1).

The behavior of the FL switching in the SOT technology depends mainly on the employed MTJ type (either with perpendicular to film plane easy axis (PMTJ) or in plane easy axis (IMTJ)) and its orientation over HM electrode. A PMTJ or an IMTJ with easy axis in parallel direction to current flow mainly requires an external magnetic field to achieve deterministic switching. Whilst, an IMTJ with easy axis perpendicular to current flow direction in HM can achieve deterministic switching without applying external magnetic field [11]. Hence, the latter is considered in this work.

It worth highlighting that the current only flows through the MTJ in the read operation and this current is relatively small (10's of µA). Hence, it is possible to implement a diode-based SOT-MTJ (D-SOT-MTJ), as depicted in Fig. 1(b), where a uni-directional metal-oxide-metal (MOM) diode is 3D stacked over the MTJ [12, 13]. This diode permits controlling the read path of the SOT-MTJs without the need for an additional silicon-based transistor, and thus reduces the area OH.

B. Previous SOT-NVSRAM

In the literature there are mainly two SOT-NVSRAM proposals [1, 8]. [1] proposed an 11T SOT-NVSRAM, shown in Fig. 2(a). Their 11T design can fully disconnect the SOT-MTJs from the conventional 6T SRAM cell during nominal operation by pulling Ctrl, BEN and REN signals low with the goal to maintain a similar performance to the conventional 6T SRAM in the nominal mode. However, four of the overhead (OH) transistors are connected to internal nodes of the 6T SRAM cell, highlighted in red in Fig. 2(a), which increases the capacitive parasitic to these nodes, and affects the

SRAM performance. It also suffers from large area overhead, where 5 additional transistors beside the two complementary SOT-MTJs are required. Moreover, the high series resistance of the 2 (2) transistors along the write and read paths of the SOT-MTJs leads to higher write/read voltages, which increases the energy consumption. [8] proposed an 8T SOT-NVSRAM depicted in Fig. 2(b). This design achieves reduced area overhead with only two additional transistors to the conventional 6T. However, as the two complementary SOT-MTJs are directly connected to the conventional 6T SRAM cell (i.e. HM in the path between bit line (BL) and latch), it affects SRAM performance during nominal operation. The 6T SRAM performance also depends on the HM resistance of the employed SOT-MTJ technology leading to higher timing overhead. Furthermore, during the store operation, voltage boosting and unsymmetrical writing for both SOT-MTJs are required, which increases the store energy significantly. Voltage boosting is needed to prevent the cross coupled inverter from flipping. Whereas, the unsymmetrical writing is due to the 6× ratio between the writing voltages used to write both SOT-MTJ (i.e. (boosted VDD – $V_{intermediate}$) for one SOT-MTJ and $V_{intermediate}$ for the other). This results in much slower switching for the SOT-MTJ with smaller write voltage and increases the overall store duration.

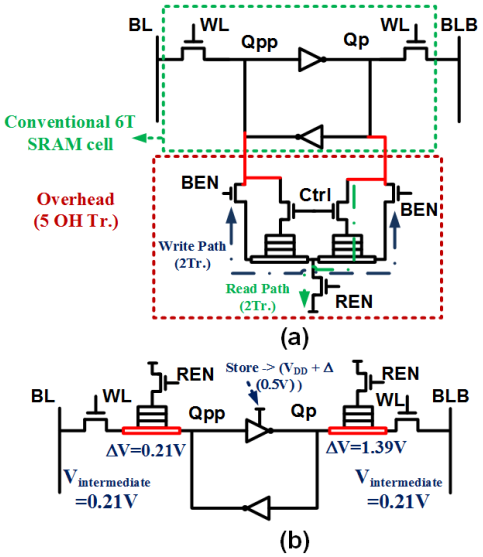

Fig. 2 (a) Schematic of the 11T SOT-NVSRAM proposal in [1]. (b) Schematic of the 8T SOT-NVSRAM proposal in [8].

III. Proposed Spin Orbit Torque Non-volatile Static Random-Access Memory (SOT-NVSRAM)

A. 9T SOT-NVSRAM

Our proposed 9T SOT-NVSRAM, shown in Fig. 3, comprises the conventional 6T SRAM cell with three OH transistors and two always complementary SOT-MTJs to add the non-volatility features. The two non-volatile SOT-MTJs for each cell are 3D stacked over the corresponding nine transistors in the BEOL process to store the data before power gating the cell during the sleep mode. Two complementary SOT-MTJs are employed to increase the read margin and speed in the restore operation. The two SOT-MTJs are connected in-series with a transistor so that the two MTJs can be written simultaneously with identical write current to reduce the store delay and energy [7]. The additional two transistors are connected to the MTJs to control the read current during the restore operation. All the OH transistors are not directly connected to the 6T cell, which minimize their impacts on the SRAM performance.

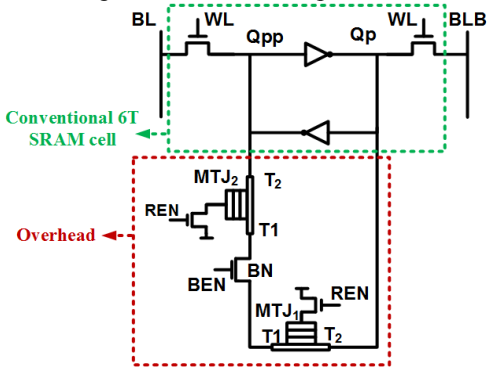

Fig. 3 Schematic of the proposed 9T SOT-NVSRAM.

Our proposed SOT-NVSRAM has four modes of operation, which are nominal mode, store mode, sleep mode and restore mode. During the nominal mode, the OH transistors are turned off by pulling both BEN and REN signals low, which fully disconnect the SOT-MTJs from the conventional 6T cell. Hence, the proposed design approximates a conventional 6T SRAM with similar performance and energy consumption. During the nominal mode, the data is written (read) to (from) the selected SRAM cell by activating the proper row and column through the row and column decoders, respectively. The data is then supplied (read) to (from) the BL and BLB. The store mode needs to be activated before power gating the SRAM in the sleep mode to preserve the SRAM data in the non-volatile device before disconnecting the power supply during the sleep mode. During the store mode, the data in the cross-coupled inverters of all the cells are stored in the corresponding MTJs in parallel by setting the BEN signal high,

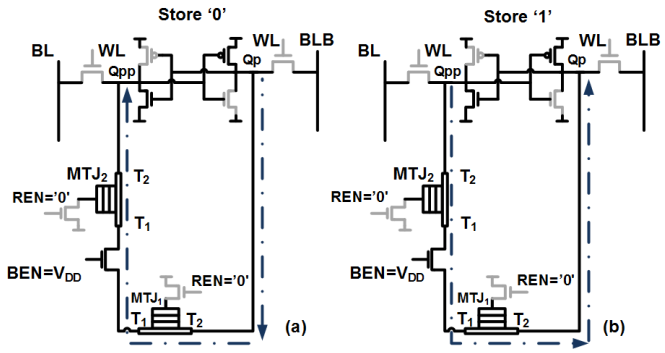

Fig.4 Proposed SOT-NVSRAM configuration and current flow direction during storing (a) data='0' and (b) data = '1', respectively.

while pulling the REN and WL (i.e. row select signal) signals low. This permits the cross coupled inverters to generate a write current to flow through the HM electrodes of the two SOT-MTJs to switch their FL magnetization to the corresponding states, as illustrated in Fig. 4. The MTJs final states are always complementary, which is determined by the write current direction based on the data in the cross coupled

inverter.

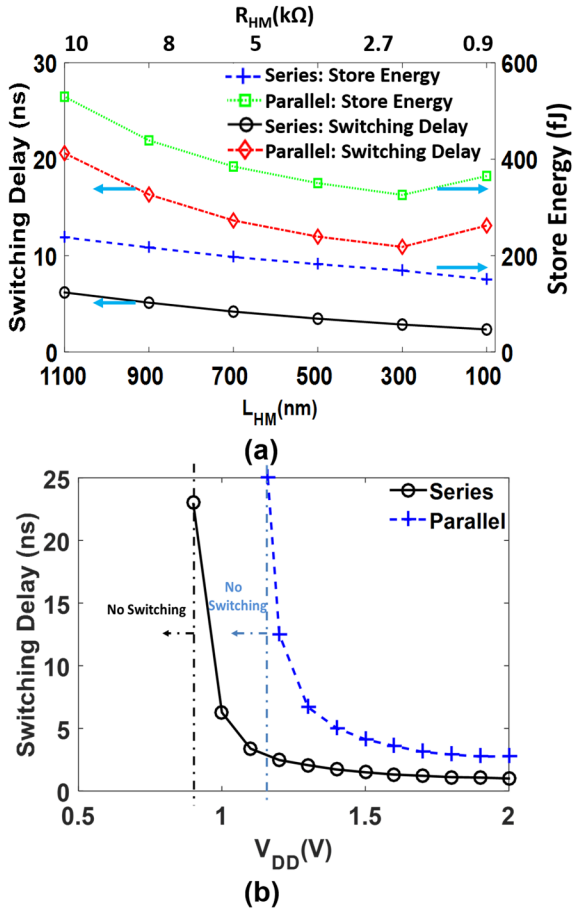

Fig.5 Comparison of SOT-MTJs series and parallel writing (a) switching delays and store energy with HM length scaling and (b) switching delay with supply voltage scaling, respectively.

It is noteworthy that only one OH transistor exists along the SOT-MTJ's write path, therefore a low resistance overhead and backup energy consumption can be achieved. It is noteworthy that in store mode, series writing is employed rather than parallel writing. This is because series writing offers lower energy consumption and switching delay compared to parallel writing across a wide range of HM length and resistance as depicted in Fig. 5(a). Furthermore, it offers better voltage scalability as it can achieve successful switching for MTJs at lower supply voltages as indicated in Fig. 5(b). The reason can be attributed to the need to supply the switching write current to only one branch in series writing, whilst, the current is supplied to two branches in parallel writing. Thereafter, to return to the nominal mode after the required sleep duration, a restore mode must be first initiated. During the restore mode, the data stored in the MTJs are copied back to the corresponding cross coupled inverters to ensure functional operation in the nominal mode. The restore operation occurs by ramping up the supply voltage from 0 to VDD [14], while raising up the REN signal high and pulling the BEN low, which results in the circuit configuration in Fig. 6. During the slow VDD ramping, the branch with the low resistance MTJ state (P-state) will conduct higher current than the other branch with high resistance MTJ state (AP-state). Thus, the low resistance branch discharges faster to ground. Due to the regenerative nature of the cross coupled inverters, the other branch is pulled up to VDD. As only one OH transistor in each branch, both the read margin and energy efficiency are improved.

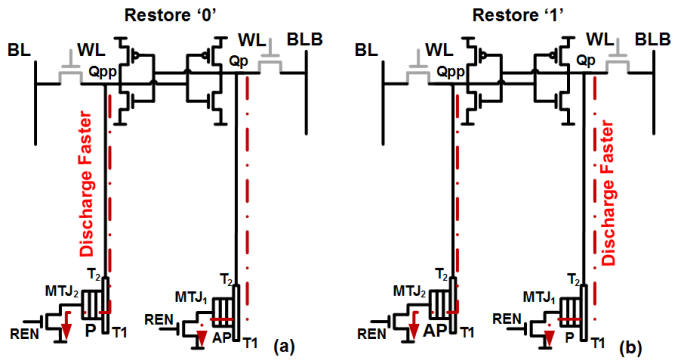

Fig.6 Proposed SOT-NVSRAM configuration and current flow direction during restoring (a) data='0' and (b) data = '1', respectively.

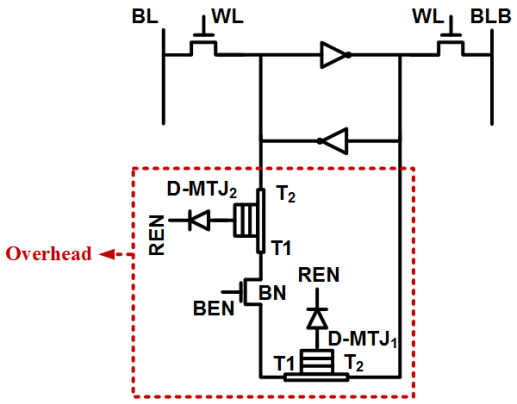

Fig.7 Schematic of the proposed 7T diode-based SOT-NVSRAM.

B. *7T D-SOT-NVSRAM*

The silicon area OH of our proposed 9T SOT-NVSRAM can be reduced by employing the D-SOT-MTJ shown in Fig. 1(b). This eliminates two transistors required for the restoration operation. Since the MOM diodes are stacked on top of the two SOT-MTJs, no extra area overhead is incurred. Thus, only seven transistors (7T) are needed as illustrated in Fig. 7. Our proposed 7T diode-based SOT-NVSRAM shares identical modes of operation as the 9T SOT-NVSRAM. The main difference is in the REN signal state in the various modes. In this design, the REN is high during the nominal and store modes to reverse bias the diodes and prevent any current to flow through the MTJs. Whereas, the REN signal is pulled low in the restore mode to forward bias the diodes and permit the read current to flow through the MTJs. Employing diode in the restore operation relatively increases the restore energy compared to the non-diode-based designs. This is due to the larger employed read voltage needed for the larger voltage headroom required for the diodes. It may also increase the restore duration depending on the transient behavior of the employed MOM diode [12, 13].

IV. Evaluation

To evaluate our proposed design, the Verilog-A model for the SOT-MTJ in [7] and the device parameters in Table I is employed. A 2×2 array of SOT-NVSRAM is simulated under various modes of operations with 32-nm technology node with 1.2 V supply voltage. Fig. 8 shows the NVSRAM under the nominal mode in which the SOT-MTJ are disconnected from the conventional 6T SRAM cell by setting both BEN and REN signals low. During this mode, various data are written to and read from the top two cells of the array. The targeted cell is selected by activating the proper world line (i.e. WL1a=VDD) and setting the column select signal (CS) to the required row and column comprising this cell. The conventional write operation for the selected cell occurs by asserting the write enable signal (RW) high, while the read operation is activated by setting the read enable signal (RE) high. Thereafter, it is required to power gate the SRAM and turn off the supply voltage during the sleep mode. Hence, a store mode is firstly activated by raising the BEN signal high (REN signal is low), which allows storing the data of the various cross-coupled inverters to the corresponding MTJS in parallel manner. Subsequently, the restore mode is activated before returning to the nominal operation. The restore operation takes place by raising the REN signal high (BEN signal is low), which results in moving the data back from each complementary MTJs pair to the corresponding latch. Once in nominal mode, the data is read back subsequently from the top two SRAM cells in the array. Similarly, the 7T design will have a similar simulation waveform except the REN signal as indicated in Section III. The 7T design is simulated using the same models elaborated in [12, 13].

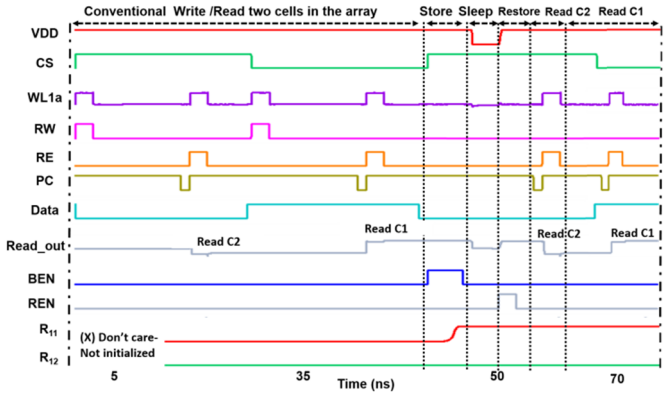

Fig.8 Simulated four modes of operation of our proposed SOT-NVSRAM. The waveform shows the successful write/read for the various data in a 2x2 memory array in nominal operation. It also shows successful store and restore for the different data. RW is the write enable signal to conventional SRAM cell, WL1a is the write word line of the first row, CS is the column select signal, RE is read enable signal, BEN is the backup enable signal to control store operation, REN is the restore enable signal to control restore operation, PC is pre-charge enable signal to restore the data to the SRAM cell, Data is the input data, R_{ij} represent the resistance value of MTJ connected to the Qpp node during restore in row i and column j, Read-out is the data read from the selected cell for any column, VDD is supply voltage. C2 is 2nd column and C1 is 1st column.

Table I. SOT-MTJ Device Parameters.

Symbol	MTJ Parameter	Value
W_{MTJ} x L_{MTJ}	MTJ Dimensions (WxL)(nm^2)	50x100
Ku V / KT	Thermal Stability	46
t_{MgO}	Tunnel Barrier Thickness (nm)	1.5
t_{HM}	Heavy Metal Thickness (nm)	3.3
W_{HM}	Heavy Metal Width (nm)	120
Ms	Magnetization Saturation (emu/cm^3)	1114
RAP	High Resistance of MTJ value (kΩ)	22
TMR	Tunnel Magneto-resistance Ratio (%)	83
θ_{SHE}	Spin Hall Angle	0.3

Table II summarizes the performance metrics of different SOT-NVSRAM designs. Under nominal operation, both our designs and the 11T SOT-NVSRAM design [1] disconnect the SOT-MTJs from the 6T SRAM cell, whereas for [8] the SOT-MTJs are still connected to the SRAM cell and interferes the nominal operation. Hence, our designs and [1] only incur 1.8% timing overhead over the conventional 6T SRAM cell while the 8T NVSRAM in [8] worsens switching delay by 170% in nominal operation. This improvement could be further enhanced with the increase of the employed SOT-MTJ HM resistance, which is expected due to SOT-MTJ future scaling [7]. This is because the switching delay in the nominal operation for [8] increases with the increase of the HM resistance, whilst for our designs it is nearly unaffected by the increase of the HM resistance. Furthermore, if the post layout parasitic of transistors are considered, our design can offer ~10% lower delay compared to [1] as the OH transistors are not connected to the internal node of the 6T SRAM cell.

During the store operation, our designs offer at least 32% lower store energy compared to the other two designs due to the employed series writing technique. Series writing supplies the write current to only one branch to write the two MTJs with identical current amplitude and achieves similar switching time. Moreover, only one overhead transistor is added along the SOT-MTJs write path, which reduces the path's resistance and the consumed store energy. On the contrary, [1] has two transistors added along the SOT-MTJs write path, which increases its resistance and consumed energy. Whereas in [8], voltage boosting is required in addition to unsymmetrical writing for both SOT-MTJs, which results in significantly higher store energy.

During restore operation, our proposed 9T design offers at least 20% less restore energy compared to others. This is because our design has only one OH transistor for each SOT-MTJ branch, whereas [1] has two OH transistors in each branch of the MTJs during the restore operation. Furthermore, our 9T design does not require voltage boosting during the restore operation, whilst [8] requires voltage boosting on certain transistors to reduce its effective resistance. On the other hand, our proposed 7T design consumes higher restore energy due to the use of diode, which requires larger read voltage for voltage headroom consideration.

TABLE II. COMPARISON OF DIFFERENT SOT-NVSRAM DESIGNS.

		8T NVSRAM [8]	11 T NVSRAM [1]	Prop. 9T NVSRAM	Prop. 7T NVSRAM
Store	delay (ns)	6	4.4	2	2
	Energy (E_S) (fJ)	1162	211	144	144
Restore	delay (ns)	1	2.5	1	2
	Energy (E_r) (fJ)	72	66	53	80
Switching delay (t_{sd}) OH to the 6T SRAM (%)		172	1.8	1.8	1.8
Transistor no. / cell (Tr. #)		8	11	9	7
FOM (Es*Er*Tr. #)		9.744	2.2301	1	1.17
FOM (Es*Er*t_{sd}*Tr. #)		26.1	2.2301	1	1.17

Though our 9T design has one more OH transistor than the 8T design [8], at least 55% improvement in a figure-of-merit (FOM), defined as energy and number of transistors product can be achieved as compared to the other designs. Whereas the 7T design offers 47% improvement in the FOM compared to other designs in the literature with the advantage of reduced area overhead.

V. CONCLUSION

This paper proposes two energy efficient, low area overhead 9T and 7T SOT-NVSRAM designs. The SOT-MTJs are totally disconnected from the 6T SRAM cell in the nominal operation to maintain the similar conventional SRAM high performance. Moreover, the OH transistors are not connected to the internal nodes of SRAM to eliminate their parasitic contribution. Two always complementary SOT-MTJs are employed to increase the read margin and speed. Furthermore, the two SOT-MTJs are connected in-series in the store operation to achieve symmetrical writing, which reduces the write energy consumption. Thus, the proposed design offers at least 32% lower backup energy compared to other proposals in the literature. In comparison to other proposals in the literature, our proposed 9T and 7T designs achieve at least 2.2x and 1.9x improved FOM defined as energy and area product.

REFERENCES

[1] W. Kang, W. Lv, Y. Zhang, and W. Zhao, "Low Store Power High-Speed High-Density Nonvolatile SRAM Design With Spin Hall Effect-Driven Magnetic Tunnel Junctions," *IEEE Transactions on Nanotechnology,* vol. 16, no. 1, pp. 148-154, 2017.

[2] S. Mukhopadhyay, H. Mahmoodi-Meimand, C. Neau, and K. Roy, "Leakage In Nanometer Scale CMOS Circuits," (in English), *2003 International Symposium on VLSI Technology, Systems, and Applications, Proceedings of Technical Papers,* pp. 307-312, Oct. 2003.

[3] T. Ohsawa *et al.*, "A 1 Mb Nonvolatile Embedded Memory Using 4T2MTJ Cell With 32 b Fine-Grained Power Gating Scheme," *IEEE Journal of Solid-State Circuits,* vol. 48, no. 6, pp. 1511-1520, 2013.

[4] Y. Seo, K. Kwon, X. Fong, and K. Roy, "High Performance and Energy-Efficient On-Chip Cache Using Dual Port (1R/1W) Spin-Orbit Torque MRAM," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems,* vol. 6, no. 3, pp. 293-304, 2016.

[5] A. M. S. Tosson, A. Neale, M. Anis, and L. Wei, "8T1R: A Novel Low-power High-speed RRAM-based Non-volatile SRAM Design," presented at the 2016 International Great Lakes Symposium on VLSI (GLSVLSI), 18-20 May 2016, 2016.

[6] K. Chen, J. Han, and F. Lombardi, "On the Restore Operation in MTJ-Based Nonvolatile SRAM Cells," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 23, no. 11, pp. 2695-2699, 2015.

[7] K. Ali, F. Li, S. Y. H. Lua, and C. Heng, "Energy- and Area-Efficient Spin–Orbit Torque Nonvolatile Flip-Flop for Power Gating Architecture," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 26, no. 4, pp. 630-638, 2018.

[8] A. Raha, A. Jaiswal, S. S. Sarwar, H. Jayakumar, V. Raghunathan, and K. Roy, "Designing Energy-Efficient Intermittently Powered Systems Using Spin-Hall-Effect-Based Nonvolatile SRAM," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 26, no. 2, pp. 294-307, 2018.

[9] M. Julliere, "Tunneling between Ferromagnetic-Films," (in English), *Physics Letters A,* vol. 54, no. 3, pp. 225-226, 1975.

[10] J. E. Hirsch, "Spin Hall effect," (in English), *Physical Review Letters,* vol. 83, no. 9, pp. 1834-1837, Aug 30 1999.

[11] S. Fukami, T. Anekawa, C. Zhang, and H. Ohno, "A Spin–orbit Torque Switching Scheme with Collinear Magnetic Easy Axis and Current Configuration," *Nat Nano,* Article vol. 11, no. 7, pp. 621-625, Jul. 2016.

[12] K. A. Ahmed, F. Li, S. Y. H. Lua, and C. Heng, "Area-Efficient Multibit-per-Cell Architecture for Spin-Orbit-Torque Magnetic Random-Access Memory With Dedicated Diodes," *IEEE Magnetics Letters,* vol. 9, pp. 1-5, 2018.

[13] K. Ali, F. Li, S. Y. H. Lua, and C. Heng, "Area Efficient Shared Diode Multi-Level Cell SOT-MRAM," *IEEE Transactions on Magnetics,* pp. 1-5, 2018.

[14] S. Yamamoto, Y. Shuto, and S. Sugahara, "Nonvolatile delay flip-flop using spin-transistor architecture with spin transfer torque MTJs for power-gating systems," (in English), *Electronics Letters,* vol. 47, no. 18, pp. 1027-U1562, Sep 1 2011.

[15] J. Sim, J. Park, M. Kim, D. Bae, Y. Choi, and L. Kim, "14.6 A 1.42TOPS/W deep convolutional neural network recognition processor for intelligent IoE systems," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, 2016, pp. 264-265.

[16] S. Park, K. Bong, D. Shin, J. Lee, S. Choi, and H. Yoo, "4.6 A1.93TOPS/W scalable deep learning/inference processor with tetra-parallel MIMD architecture for big-data applications," in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, 2015, pp. 1-3.

[17] L. Cavigelli and L. Benini, "Origami: A 803-GOp/s/W Convolutional Network Accelerator," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 27, no. 11, pp. 2461-2475, 2017.

[18] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2014, pp. 10-14.