In-Memory Processing Paradigm for Bitwise Logic Operations in STT-MRAM

Wang Kang[1,2], *Member, IEEE,* Haotian Wang[2], Zhaohao Wang[2], *Member, IEEE,* Youguang Zhang[2], *Member, IEEE,* and Weisheng Zhao[1,2], *Senior Member, IEEE*

[1]Beijing Advanced Innovation Center for Big Data and Brain Computing (BDBC), Beihang Univeristy, Beijing, 100191, China
[2]Fert Beijing Research Institute, School of Electrical and Information Engineering, Beihang Univeristy, Beijing, 100191, China

In current big data era, the memory wall issue between the processor and the memory becomes one of the most critical bottlenecks for conventional Von-Newman computer architecture. In-memory processing (IMP) or near-memory processing (NMP) paradigms have been proposed to address this problem by adding a small amount of processing units inside/near the memory. Unfortunately, although have been intensively studied, prior IMP/NMP platforms are practically unsuccessful because of the fabrication complexity and cost efficiency by integrating the processing units and memory on the same chip. Recently, emerging nonvolatile memories provide new possibility for efficiently implementing the IMP/NMP paradigm. In this work, we propose a cost-efficient IMP/NMP solution in STT-MRAM without adding any processing units on the memory chip. The key idea behind the proposed IMP/NMP solution is to exploit the peripheral circuitry already existing inside memory (or with minimal changes) to perform bitwise logic operations. Such an IMP/NMP platform enables rather fast logic operations as the logic results can be obtained immediately through just a memory-like readout operation. Memory READ and logics NOT, AND/NAND, OR/NOR operations can be achieved and dynamically configured within the same STT-MRAM chip. Functionality and performance are evaluated with hybrid simulations under the 40 nm technology node.

***Index Terms*—STT-MRAM, in-memory processing (IMP), near-memory processing (NMP), nonvolatile memory.**

I. Introduction

In current big data era, the limited data bandwidth between the processor and the memory has become one of the most critical bottlenecks for conventional Von-Newman computer architectures. At the same time, the data transfer between the processor and the memory consumes high latency and energy, significantly degrading the system performance and efficiency. These issues are known as "memory wall" and "power wall", posing unprecedented challenges for computing capability to handle the ever-growing big data [1, 2]. In-memory processing or near-memory processing (IMP/NMP) paradigm, a decades-old concept, has been proposed to address this problem by adding a small amount of processing units inside/near the memory [3, 4]. The basic idea behind is that instead of moving all the raw data to the processor, it pre-processes the raw data and provides the processor only the intermediate results. Such a paradigm can reduce the overhead of data transfer bandwidth and power as well as improve performance by performing simple yet bandwidth-intensive logic operations in/near the memory. Unfortunately, although have been intensively studied, prior IMP/NMP platforms are practically unsuccessful, because manufacturing the performance-optimized processing units and the density-optimized memory on the same chip/die is rather complex and is not cost-effective.

Recently, emerging nonvolatile memories [5-7], such as resistive random access memory (ReRAM), phase change memory (PCM) and spin-transfer torque magnetic random access memory (STT-MRAM) etc., provide new possibility for efficiently implementing the IMP/NMP paradigm [3, 8-11]. On one hand, the 3D-stacking functionality of the nonvolatile memory devices enables to decouple the logic and memory circuits in different manufacturing processes with the back-end-of-line (BEOL) process technology, significantly alleviating the fabrication complexity and cost [12]. On the other hand, the resistance-based storage mechanism of the nonvolatile memory devices provides inherent processing capability, thus enabling energy-efficient logic computing within the memory. In this case, logic operations can be performed and the results are stored in the memory chip nonvolatilely. Therefore, IMP/NMP has recently reignited interest among industry and academic communities driven by the recent advances in these nonvolatile memory technologies [3, 13-18]. For example, "Pinatubo" in [3] has been designed for bulk bitwise logic operations in emerging nonvolatile memories, which, however, may not be employed directly for STT-MRAM owing to the low TMR ratio. Thereby, different strategies should be exploited for different nonvolatile memories, considering their intrinsic properties.

In this work, we propose a cost-efficient IMP/NMP solution for STT-MRAM having complementary array structure without adding any processing units on the memory chip. The key idea behind the proposed IMP/NMP solution is to exploit the peripheral circuitry (e.g., the sensing amplifier) already existing inside memory (or with minimal changes) to perform bitwise logic operations. Such an IMP/NMP platform enables fast and reliable logic operations as the logic results can be obtained immediately through just a memory-like readout operation. With the proposed IMP/NMP platform, memory READ and logic NOT, AND/NAND, OR/NOR operations can be achieved and dynamically configured within the same STT-MRAM chip.

The remainder of this paper is organized as follows. Section II briefly introduces the fundamentals of STT-MRAM. In Section III, we present the IMP/NMP paradigm in STT-MRAM. Section IV illustrates our simulation and discussion. Finally Section V concludes this paper.

Manuscript received March 10, 2017; Revised April 7, 2017.
Corresponding author: W. Zhao (e-mail: weisheng.zhao@buaa.edu.cn).
W. Kang and H. Wang contributed equally.
Digital Object Identifier (inserted by IEEE).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMAG.2017.2703863, IEEE Transactions on Magnetics

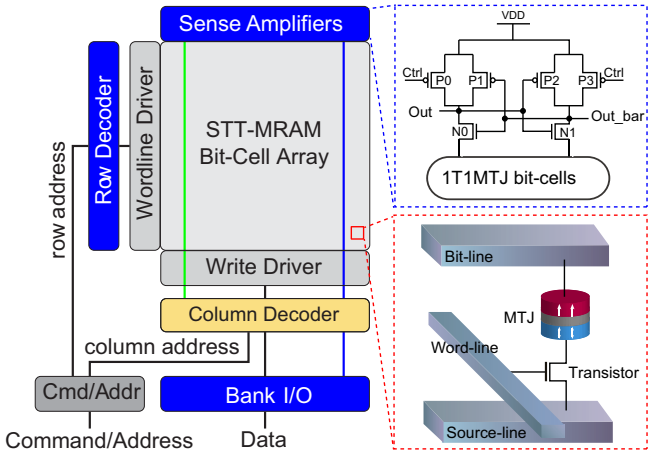


Fig. 1. Schematic of a STT-MRAM bank and the associated 1T1MTJ bit-cell structure and sensing amplifier.

II. STT-MRAM FUNDAMENTALS

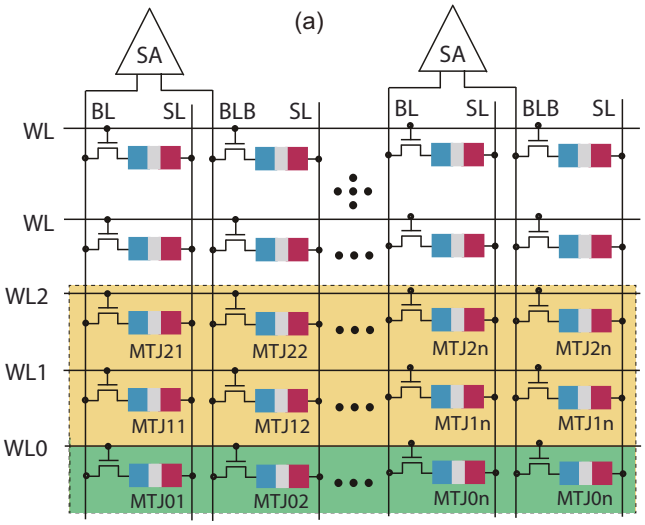
Fig. 1 shows the schematic of a STT-MRAM bank and the associated 1T1MTJ (i.e., one Transistor connected in series with one Magnetic Tunnel Junction) bit-cell structure and the peripheral sensing amplifier. The MTJ device is used as storage element while the transistor acts as an access device. An MTJ is mainly composed of three ultra-thin layers: one oxide barrier layer (e.g., MgO) sandwiched between two ferromagnetic (FM) layers (e.g., CoFeB). An MTJ can present two resistance states (i.e., low resistance, R_P and high resistance, R_{AP}) depending on the relative magnetization orientation (parallel (P) or anti-parallel (AP)) of the two FM layers. The resistance difference between the two stable resistance states of the MTJ device is characterized by the tunnel magneto-resistance (TMR) ratio (TMR $= (R_{AP} - R_P)/R_P$). The STT-MRAM memory bank is organized with an array of 1T1MTJ bit-cells through multiple bit-lines (BLs), word-lines (WLs), source-lines (SLs) and peripheral circuits, such as row/column decoders, write/read drivers as well as input/output (I/O) interfaces. To write data bit into the bit-cell, only a bi-directional current (through the STT mechanism [19]) is required through the write driver to change the magnetization orientation of the free layer of the MTJ. Due to the TMR effect, the digital information stored in the MTJ device of the bit-cell can be readout by distinguishing the resistance of the MTJ device in the memory bit-cell from that of a reference cell, in which the resistance of the MTJ is pre-known and is generally set to $(R_{AP} + R_P)/2$ [20, 21]. Here we consider a complementary memory cell array in STT-MRAM considering the limited TMR ratio (which is generally below 200% at room temperature) of the MTJ devices [8]. In such a configuration, we require no reference cell and the digital information stored in the memory is readout by comparing the resistance difference between the two complementary bit-cells (through BLs) via sensing amplifiers (SAs). Here we choose the pre-charge sensing amplifier (PCSA) in our design [22], as it provides fast sensing speed, high reliability and low power consumption. The read operation of the PCSA can be divided into two stages depending on the control signal and is described as follows. During the pre-charge (reset) phase (WL = 0 V and Ctrl = 0 V), the four PMOS transistors P0-P3 turn on and they charge the nodes "Out" and "Out_bar" to V_{dd}. During the discharging and evaluation stages (WL = V_{dd} and Ctrl = V_{dd}), nodes "Out" and "Out_bar" begins to discharge but with a different rate owing to the different resistances between the two complementary bit-cells (Bls). Then a differential voltage ΔV will build up between the two BLs (also between nodes "Out" and "Out_bar"). Once either "Out+" or "Out-" becomes less than the switching threshold voltage of the transistor in the cross-coupled inverters, composed of P1-P2 and N0-N1, they begin to amplify the voltage difference between "Out+" and "Out-" nodes, then either "Out-" or "Out+" node will be charged to V_{dd} or '1', while the other one will continue discharging to Gnd or '0'. Therefore the data stored in the bit-cells will be readout. Note that there are never any stationary currents during the sensing operation, only dynamic charging or discharging currents, we can expect very high speed and low power consumption. The details on the implementation and operation of the other peripheral circuits are not the focus of this paper and thus are omitted here.

III. PROPOSED IMP/NMP PARADIGM IN STT-MRAM

As discussed in the previous section, the data read operation in STT-MRAM is to compare the resistances of the MTJs in the two complementary bit-cells along the BLs through a SA. Since multiple WLs share the same set of SAs or BLs, the final output state of the SA depends solely on the resistances of the BLs connected to the two branches of the SA. In this configuration, we can perform bitwise logic operations by activating two (or more) WLs simultaneously, as shown in Fig. 2(a), which is an example of a 2-input AND or OR operation (see the truth table in Fig. 2(b)). Here the data stored in WL1 and WL2 are the two operands, while the data stored in WL0 are the control operands determining the logic types (either AND or OR). Specifically, if the data in WL0 are 1s, then the outputs of the SAs are bitwise OR of the data in WL1 and WL2. Otherwise, if the data in WL0 are 0s, then the outputs of the SAs are bitwise AND of the data in WL1 and WL2. Fig. 3 shows the detailed schematic combing the PCSA circuit. The data stored in the bit-cells along WL0 can be variable in order to perform different logic operations (either AND or OR) for different BLs. Furthermore, by adding an inverter (INV) and a multiplexer after the each SA, bitwise memory READ/NOT, NAND/NOR operations can then be achieved. Because of the regular bulk operation of bit-cells along the WLs in STT-MRAM, this approach naturally enables bulk bitwise READ/NOT, AND/OR and NAND/NOR logic operations through the SAs. Moreover, such an IMP/NMP platform enables rather fast logic operations as the results can be obtained immediately through just a memory-like readout operation. In addition, the processing is nondestructive and the logic operands and results can be statefully stored locally in the memory owing to the nonvolatility of the STT-MRAM bit-cells.

As can be seen, by modifying the row decoder and the SA, bitwise READ/NOT, AND/NAND, OR/NOR operations can be achieved with very little hardware overhead via our proposed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMAG.2017.2703863, IEEE Transactions on Magnetics



(b) Turth table for the AND and OR operations

Logic gate	Data in WL0 (MTJ 01, 02)	Data in WL1 (MTJ 11, 12)	Data in WL2 (MTJ 21, 22)	Output of the SA
AND	0 (P, AP)	0 (P, AP)	0 (P, AP)	0
	0 (P, AP)	0 (P, AP)	1 (AP, P)	0
	0 (P, AP)	1 (AP, P)	0 (P, AP)	0
	0 (P, AP)	1 (AP, P)	1 (AP, P)	1
OR	1 (AP, P)	0 (P, AP)	0 (P, AP)	0
	1 (AP, P)	0 (P, AP)	1 (AP, P)	1
	1 (AP, P)	1 (AP, P)	0 (P, AP)	1
	1 (AP, P)	1 (AP, P)	1 (AP, P)	1

Fig. 2. Illustration of (a) the proposed IMP/NMP paradigm for bitwise AND/OR operations in STT-MRAM, and (b) the corresponding truth table.

IMP/NMP paradigm within a STT-MRAM. The data stored in the STT-MRAM can be straightforwardly readout or processed by dynamically configuring the peripheral circuits depending on practical requirements. It should be noted that for data stored in the same STT-MRAM bank, the bitwise logic operations are rather simple. For data stored in different memory banks or sub-arrays, data movements within the memory chip are required to implement the logic operations, which, however, do not occupy the bandwidth between the processor and memory. In addition, system/software supports are generally required to implement the IMP/NMP paradigm, which are beyond the scope of this paper. More details about the system/software supports can refer to [3]. The proposed cost-effective IMP/NMP platform is very useful for some specific applications (e.g., bitmap) that rely on bulk bitwise logic operations on large bit-vectors.

IV. Evaluation and Discussion

By using a commercial CMOS 40 nm design kit and an in-house developed perpendicular anisotropy CoFeB/MgO STT-MTJ compact model [23], hybrid STT-MTJ/CMOS simulations are performed to demonstrate the functionality and performance of the proposed IMP/NMP platform in STT-MRAM. Here the

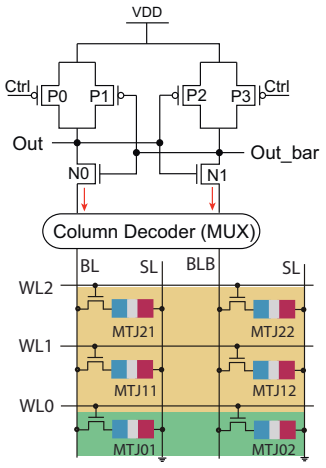


Fig. 3. Proposed IMP/NMP paradigm for bitwise 2-input AND/OR operations with the PSCA circuit.

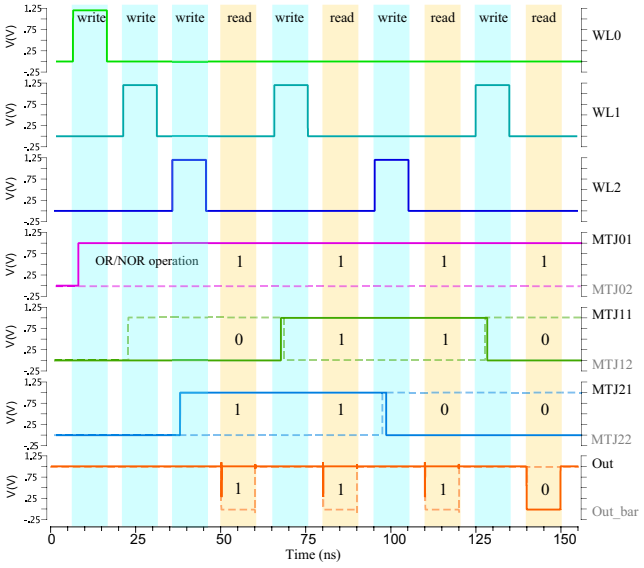


Fig. 4. Transient waveforms of the OR/NOR operations with the proposed IMP/NMP platform in STT-MRAM.

supply voltage is $V_{dd} = 1.2\,\mathrm{V}$, and the TMR = 100%. We use the minimum feature size for all the CMOS transistors and the default values for the STT-MTJ model can refer to [23].

Fig. 4 and Fig. 5 show the transient simulation waveforms of the OR/NOR and AND/NAND operations with the proposed IMP/NMP platform respectively. Initially, we activate each WL to write the initial data into the bit-cells. Then we change the data along the WLs to verify the functionality of the logic operations. As can be seen, the latency (~200 ps) for performing logic operations is similar to that of reading a data bit from the memory bit-cells. Further, we can also configure the data stored in the bit-cells along WL0 to dynamically configure the logic types for different BLs. The transient simulations validate the correctness of the proposed IMP/NMP design in STT-MRAM. It should be noted that as the bit-cells along the WLs are parallel connected during the logic operation mode, the real resistance difference between the two BLs will be degraded compared to that of the memory mode. In this case, the sensing reliability of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMAG.2017.2703863, IEEE Transactions on Magnetics

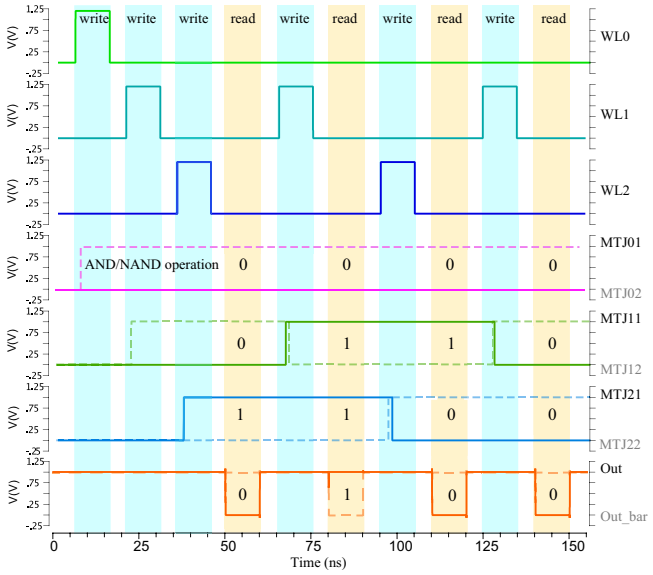


Fig. 5. Transient waveforms of the AND/NAND operations with the proposed IMP/NMP platform in STT-MRAM.

the SA will be affected because of reduced real TMR ratio and increased parasitic parameters. Therefore, higher TMR ratio of the MTJ devices and sensitive SAs (e.g., separated PCSA [24] and tunneling negative differential resistance-assisted SA [25]) are preferred in the proposed IMP/NMP platform.

V. Conclusion

A cost-efficient IMP/NMP platform is proposed here in STT-MRAM for bitwise logic operations to address the memory wall and power wall issues. Without adding any processing units, the proposed IMP/NMP platform performs logic operations within the memory through exploiting the peripheral circuitry already existing inside memory. Memory READ/NOT, AND/NAND, OR/NOR operations can be achieved within the STT-MRAM chip. Our simulations demonstrate that the proposed IMP/NMP platform enables rather fast logic operations as the logic results can be obtained via just a memory-like readout operation. More efforts on system or software hierarchy are expected to support the proposed IMP/NMP paradigm.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant 61501013 and 61571023).

References

[1] N. S. Kim, T. Austin, D. Baauw, T. Mudge, K. Flautner, J. S. Hu, M. J. Irwin, M. Kandemir, and V. Narayanan, "Leakage Current: Moore's Law Meets Static Power," *Computer*, vol. 36, no.12, pp. 68-75, Dec. 2003.

[2] W. A. Wulf and S. A. McKee, "Hitting the Memory Wall: Implications of the Obvious," *ACM SIGARCH Computer Architect. News*, vol. 23, no. 1, pp. 20-24, Mar. 1995.

[3] S. Li, C. Xu, Q. Zou, J. Zhao, Y. Lu, and Y. Xie, "Pinatubo: A Processing-in-memory Architecture for Bulk Bitwise Operations in Emerging Non-volatile Memorties," in *Proc. DAC*, Jun. 2016, pp. 1-6.

[4] J. Ahn, S. Hong, S. Yoo, O. Mutlu, and K. Choi, "A scalable processing-in-memory accelerator for parallel graph processing," in *ACM/IEEE ISCA*, 2015, pp. 105-117.

[5] W. Kang, L. Zhang, J. Klein, Y. Zhang, D. Ravelosona, and W. S. Zhao, "Reconfigurable Codesign of STT-MRAM under Process Variations in Deeply Scaled Technology," *IEEE Trans. Electron Devices*, vol. 62, no. 6, pp. 1769-1777, Jun. 2015.

[6] S. Wang, H. Lee, F. Ebrahimi, P. K. Amiri, K. L. Wang, and P. Gupta, "Comparative Evaluation of Spin-Transfer-Torque and Magnetoelectric Random Access Memory," *IEEE J. Emerg. Select. Top. Circ. Syst.*, vol. 6, no. 2, pp. 134-145, Apr. 2016.

[7] W. Kang, Y. Zhang, Z. Wang, J. Klein, C. Chappert, D. Ravelosona, G. Wang, Y. Zhang, and W. S. Zhao, "Spintronics: Emerging Ultra-low Power Circuits and Systems Beyond MOS Technology," *ACM JETC*, vol. 12, no. 16, pp. 1-42, Aug. 2015.

[8] S. Jain, A. Ranjan, K. Roy, and A. Raghunathan, "Computing in Memory with Spin-Transfer Torque Magnetic RAM," arXiv:1703.02118.

[9] D. Fan, M. Sharad, and K. Roy, "Design and Synthesis of Ultra Low Energy Spin-Memristor Threshold Logic," *IEEE Trans. Nanotechnol.*, vol. 13, no. 3, pp. 574-583, Mar. 2014.

[10] T. Hanyu, T. Endih, D. Suzuki, H. Koike, Y. Ma, N. Onizawa, M. Natsui, S. Ikeda, and H. Ohno, "Standby-Power-Free Integrated Circuits Using MTJ-Based VLSI Computing," *Proc. IEEE*, vol. 104, no. 10, pp. 1844-1863, Sept. 2016.

[11] S. Senni, L. Torres, G. Sassatelli, and A. Gamatie, "Non-volatile Processor Based on MRAM for Ultra-low-power IoT Devices," *ACM JETC*, vol. 13, no. 17, pp. 1-23, Dec. 2016.

[12] W. Kang, W. S. Zhao, Z. Wang, Y. Zhang, J. O. Klein, C. Chappert, Y. Zhang, and D. R. Ramasitera, "DFSTT-MRAM: Dual Functional STT-MRAM Cell Structure for Reliability Enhancement and 3-D MLC Functionality," *IEEE Trans. Magn.*, vol. 50, no. 6, pp. 1-7, Jun. 2014.

[13] H. Cai, Y. Wang, L. Naviner, and W. S. Zhao, "Robust Ultra-Low Power Non-Volatile Logic-in-Memory Circuits in FD-SOI Technology," *IEEE Trans. Circ. Syst. I: Reg. Papers*, vol. PP, no. 99, pp. 1-11, Nov. 2016.

[14] W. Kang, Z. Wang, Y. Zhang, J. O. Klein, W. Lv, and W. S. Zhao, "Spintronic Logic Design Methodology Based on Spin Hall Effect–Driven Magnetic Tunnel Junctions," *J. Phys. D: Appl. Phys.*, vol. 49, no. 6, pp. 1-11, Feb. 2016.

[15] D. Fan, "Low Power In-memory Computing Platform with Four Terminal Magnetic Domain Wall Motion Devices." in *IEEE/ACM NANOARCH*, Jul. 2016, pp. 153-158.

[16] T. Hanyu, D. Suzuki, A. Mochizuki, M. Natsui, N. Onizawa, T. Sugibayashi, S.Ikeda, T. Endoh, and H. Ohno, "Challenge of MOS/MTJ-Hybrid Nonvolatile Logic-in-Memory Architecture in Dark-Silicon Era", in *IEEE IEDM.*, 2014, pp. 28.2.1-28.2.3.

[17] B. Jovanovic, R. M. Brum, and L. Torres, "Comparative Analysis of MTJ/CMOS Hybrid Cells Based on TAS and In-plane STT Magnetic Tunnel Junctions," *IEEE Trans. Magn.*, vol. 51, no. 2, pp. 3400111, 2014.

[18] T. Hanyu, D. Suzuki, N. Onizawa, S. Matsunaga, M. Natsui and A. Mochizuki, "Spintronics-based nonvolatile logic-in-memory architecture towards an ultra-low-power and highly reliable VLSI computing paradigm," *in IEEE DATE.*, 2015, pp. 1006-1011.

[19] J. C. Slonczewski, "Current-driven Excitation of Magnetic Multilayers," *J. Magn. Magn. Mater.*, vol. 159, no. 1-2, pp. L1-L7, Jun. 1996.

[20] W. Kang, Z. Li, J. Klein, Y. Cheng, Y. Zhang, D. Ravelosona, C. Chappert, and W. S. Zhao, "Variation-tolerant and Disturbance-free Sensing Circuit for Deep Nanometer STT-MRAM," *IEEE Trans. Nanotechnol.*, vol. 13, no. 6, pp. 1088-1092, Nov. 2014.

[21] W. Kang, W. S. Zhao, J. Klein, Y. Zhang, C. Chappert, and D. Ravelosona, "High Reliability Sensing Circuit for Deep Submicron Spin Transfer Torque Magnetic Random Access Memory," *Electron. Lett.*, vol. 49, no. 20, pp. 1283-1285, Sep. 2013.

[22] W. S. Zhao, C. Chappert, V. Javerliac, and J. P. Noziere, "High Stability and Low Power Sensing Amplifier for MTJ/CMOS Hybrid Logic Circuits," *IEEE Trans. Magn.*, vol. 45, no. 10, pp. 3784-3787, Oct. 2009.

[23] Y. Zhang, B. Yan, W. Kang, Y. Cheng, J. Klein, Y. Zhang, Y. Chen, and W. S. Zhao, "Compact Model of Subvolume MTJ and Its Design Application at Nanoscale Technology Nodes," *IEEE Trans. Electron Devices*, vol. 62, no. 6, pp. 2048-2055, 2015.

[24] W. Kang, E. Deng, J.-O. Klein, Y. Zhang, Y. G. Zhang, C. Chappert, D. Ravelosona, and W. S. Zhao, "Separated Precharge Sensing Amplifier for Deep Submicrometer MTJ/CMOS Hybrid Logic Circuits," *IEEE Trans. Magn.* vol. 50, no. 6, pp. 3400305-5, Jun. 2014.

[25] S. Wang, A. Pan, C. O. Chui, and P. Gupta, "Tunneling Negative Differential Resistance-Assisted STT-RAM for Efficient Read and Write Operations," *IEEE Trans. Electron Devices*, vol. 64, no. 1, pp. 121-129, Dec. 2016.