Relaxing Non-Volatility for Energy-Efficient DMTJ Based Cryogenic STT-MRAM

Esteban Garzón, Raffaele De Rose, Felice Crupi, Lionel Trojman, Adam Teman, Marco Lanuzza

▶ **To cite this version:**

Esteban Garzón, Raffaele De Rose, Felice Crupi, Lionel Trojman, Adam Teman, et al.. Relaxing Non-Volatility for Energy-Efficient DMTJ Based Cryogenic STT-MRAM. Solid-State Electronics, 2021, pp.108090. 10.1016/j.sse.2021.108090 . hal-03254839

HAL Id: hal-03254839
https://hal.science/hal-03254839v1

Submitted on 9 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Relaxing Non-Volatility for Energy-Efficient DMTJ Based Cryogenic STT-MRAM

Esteban Garzón[a,b,*], Raffaele De Rose[a], Felice Crupi[a], Lionel Trojman[c], Adam Teman[b], Marco Lanuzza[a]

[a] *Department of Computer Engineering, Modeling, Electronics and Systems (DIMES), University of Calabria, Rende 87036, Italy*
[b] *Emerging Nanoscaled Integrated Circuits & Systems (EnICS) Labs, Faculty of Engineering, Bar-Ilan University, Ramat-Gan 5290002, Israel*
[c] *Laboratoire d'Informatique, Signal, Image, Télécommunication et Électronique (LISITE), Institut Supérieur d'Électronique de Paris (ISEP), Paris 75006, France*

Abstract

Spin-transfer torque magnetic random-access memory (STT-MRAM) is considered a potential universal memory that could replace conventional six-transistors static random-access memory (6T-SRAM) in processor caches. This paper explores STT-MRAMs based on double-barrier magnetic tunnel junction (DMTJ), while operating at cryogenic temperatures (77 K). To deal with large dynamic energy and long write latency of magnetic memories we suggest to significantly relax the non-volatility requirement (i.e., by reducing cross-section area) of the DMTJ devices at room temperature, while maintaining the typical ten year storage retention time at the target operating temperature. This leads the DMTJ-based STT-MRAM to be more energy-efficient under both read/write operations than 6T-SRAM, while maintaining a smaller area footprint.

Keywords: Double-barrier magnetic tunnel junction (DMTJ), STT-MRAM, cryogenic computing, thermal stability relaxation, 77 K

1. Introduction

Cryogenic electronics is nowadays gaining great interest in view of its potential to go over the power/memory wall of classical room-temperature computing [1]. To efficiently support cryogenic computing, low-energy and high-density memories are required. Recent studies have focused on conventional memories operating down to the liquid-nitrogen temperature (77 K) [2, 3]. Moreover, cryogenic operation of spin-transfer torque magnetic random-access memory (STT-MRAM) based on standard single-barrier MTJ (SMTJ) was recently evaluated in [4], proving to outperform its SRAM counterpart for architectures with medium to large storing capability (i.e., typical L2/L3-caches), mainly due to the better scalability of the memory interconnection structure [5]. On the other hand, for smaller memory sizes, the SMTJ-based STT-MRAM is penalized by the high bitcell latency and energy consumption under write access, both increased as temperature reduces [2, 3]. A feasible solution to deal with the above issue consists of replacing conventional SMTJs with double-barrier MTJ (DMTJ) devices with two reference layers. Due to the reduced switching current of DMTJ, faster and less energy demanding write operation is potentially allowed [6–10].

In the above context, this work investigates DMTJ-based STT-MRAM operating at cryogenic conditions by using a cross-layer simulation framework [5, 6], starting from the device- up to the memory architecture-level. In our study, we considered cache memory sizes from 64 kB up to 2 MB for a target operating temperature of 77 K. To improve DMTJ energy and performance characteristics at cryogenic temperatures, we suggest to significantly relax the non-volatility requirement (i.e., by reducing cross-section area of the device) at room temperature, while maintaining the typical ten year storage retention time at the target operating temperature. In this way, under 77 K operating temperature condition, the DMTJ-based STT-MRAM proves to be ≈ 60% and ≈ 20% less energy-hungry than conventional six-transistor static random access memory (6T-SRAM), under read and write operations, respectively. The rest of the paper is organized as follows. Section 2 discusses the DMTJ device compact model used in this work. Section 3 presents the circuit- and architecture-level results of our simulation study. Finally, Section 4 summarizes our work by remarking on its main conclusions.

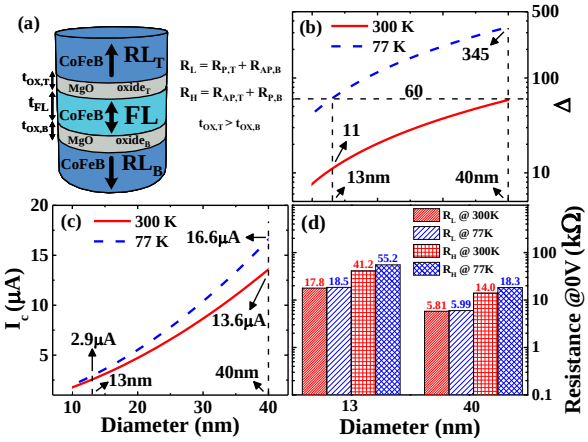

Figure 1: (a) Sketch of the DMTJ device and its temperature-dependent characteristics as a function of the cell diameter for 300 K and 77 K: (b) thermal stability factor (Δ), (c) critical switching current (I_c), (d) resistance in low (R_L) and high (R_H) states.

*Corresponding author
Email address: esteban.garzon@unical.it (Esteban Garzón)

Table 1: DMTJ parameters

Parameter	Description	DMTJ - 40 nm (300 K, 77 K)	DMTJ - 13 nm (300 K, 77 K)
t_{FL}	FL thickness	1.2 nm	
$t_{OX,T}$	Top barrier thickness	1 nm	0.85 nm
$t_{OX,B}$	Bottom barrier thickness	0.5 nm	0.4 nm
RA	Resistance-area product	~7 Ω·μm^2	~2 Ω·μm^2
P	Spin polarization factor	(0.66, 0.73)	
M_S	Saturation magnetization	(1.58, 1.80) T	
α	Gilbert damping factor	0.03	
K_i	Interfacial perpendicular anisotropy constant	(1.3×10^{-3}, 1.73×10^{-3}) J/m^2	
TMR(0)	Tunnel magnetoresistance @0V	(141, 205) %	(131, 198) %

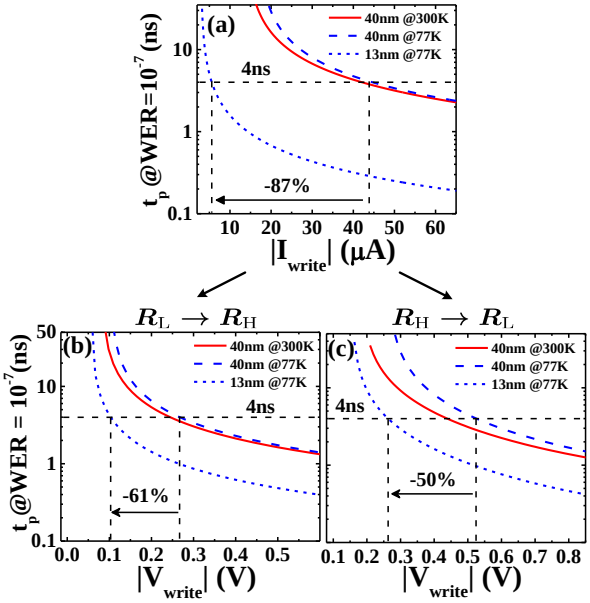

Figure 2: Switching behavior of the DMTJ device with 40 nm diameter (at 300 K and 77 K) and 13 nm diameter (at 77 K) in terms of the write pulse width (t_p) to ensure a write error rate WER = 10^{-7} as a function of (a) write current (I_{write}), and write voltage (V_{write}) for (b) low-to-high ($R_L \rightarrow R_H$) and (c) high-to-low ($R_H \rightarrow R_L$) transitions.

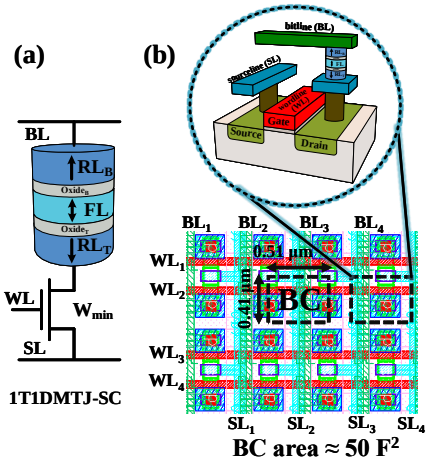

Figure 3: (a) Schematic of the one transistor-one MTJ STT-MRAM bitcell (BC) in standard configuration (1T1MTJ-SC) and (b) its minimum-area layout (corresponding to a BC area of 50 F^2). Single BC is highlighted with dashed lines within the memory array.

2. Device-level modeling

Fig. 1(a) shows the sketch of a perpendicular DMTJ, where the CoFeB free layer (FL) with variable magnetization orientation is sandwiched between two MgO barriers with different thickness, each one interfaced with a CoFeB polarizing reference layer (RL). To allow reduced switching current, the two RLs feature fixed magnetization orientations in antiparallel configuration [10]. Table 1 reports the DMTJ physical parameters considered in this work for two different devices featuring circular perpendicular magnetic anisotropy (PMA) geometry and diameter of 40 nm and 13 nm, respectively. Temperature-dependent parameter values have also been made explicit for two operating temperatures (300 K and 77 K). Note that, in order to assure fully compatibility with the CMOS process, the resistance of the shrinked device is scaled down according to the trends reported in literature [11].

At the device-level, the DMTJ is described by our macrospin-based Verilog-A compact model [10], which also mimics the temperature dependence of the device physical parameters [12]. Figs. 1(b)-(d) show the temperature-dependent characteristics for the considered DMTJs. The 40 nm DMTJ exhibits a thermal stability factor $\Delta \approx 60$ at 300 K, which is enough to ensure 10-years data retention [6]. However, as shown in Fig. 1(b), the Δ overly grows at the lowest temperature ($\Delta \approx 345$ at 77 K), while increasing by $\approx$ 20% the switching current, with detrimental effect on energy consumed during the write operation (see Fig. 1(c)). By shrinking the FL cross-section area (i.e., by scaling the DMTJ diameter) to achieve a data retention time of about 10-years at 77 K ($\Delta \approx 60$ for 13 nm diameter), the switching current can be considerably reduced (see Fig. 1(c)), with the added benefit of increased storage density. As shown in Fig. 2(a), the suggested approach leads to reduce the switching current by more than 80% at the parity of pulse duration for a target write error rate (WER) of 10^{-7} [6]. From Figs. 2(b-c), it can be observed that the required write voltage pulse amplitudes (for both the $R_L \rightarrow R_H$ and $R_H \rightarrow R_L$ transitions) are slightly increased for the 40 nm DMTJ as the temperature goes down to 77 K. This energy-adverse effect is counteracted by the relaxed non-volatility of the 13 nm DMTJ device, which allows the required write voltage amplitudes to be considerably reduced (56%, on average) at cryogenic temperature.

It is worth pointing out that, the increase of the required write voltage of the 40 nm DMTJ (see Fig. 2), owing to the decrease in temperature from 300 K to 77 K, implies a greater device degeneration in terms of cycling endurance [13]. To evaluate the DMTJ stack cycling endurance, we considered the semi-empirical time-dependent dielectric breakdown (TDDB) model proposed in [13]. According to this model and referring to data provided in Figs. 2(b)-(c) in terms of required voltages and pulse width = 4 ns, the 13 nm DMTJ with relaxed non-volatility at 77 K improves the endurance lifetime by orders-of-magnitude when compared to its 40 nm counterpart operating at the same temperature.

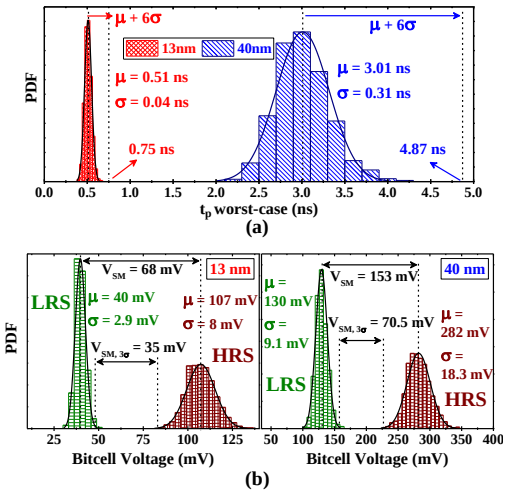

Figure 4: DMTJ-based STT-MRAM bitcell-level Monte Carlo (MC) simulations for (a) write (@WER = 10^{-7}) and (b) read (@RDR = 10^{-9}) operations at 77 K.

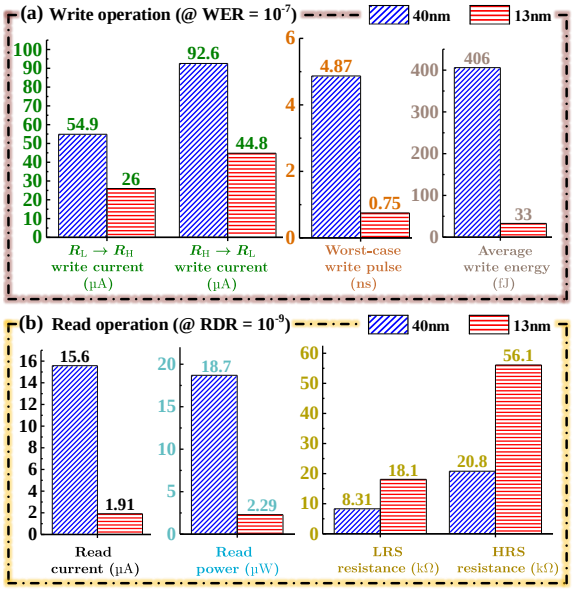

Figure 5: DMTJ-based STT-MRAM bitcell-level results at 77 K for devices with 40 nm and 13 nm diameters: (a) write operation ensuring write error rate WER = 10^{-7} and (b) read operation ensuring read disturbance rate RDR = 10^{-9}.

3. Circuit- and architecture-level analysis at 77 K

3.1. Circuit-level results

The DMTJ-based bitcell shown in Fig. 3 has been simulated in the Cadence Virtuoso environment by exploiting our DMTJ compact model [10] along with a commercial 65 nm CMOS technology experimentally characterized down to the liquid nitrogen temperature [4]. Extensive Monte Carlo (MC) simulations were performed for an operating temperature of 77 K, while considering both DMTJ and CMOS variability. While the CMOS process variations have been taken into account by using the statistical models provided by the process design kit (PDK), the DMTJ manufacturing uncertainty is considered by our Verilog A compact model which allows defining the variability (σ/μ) of some process parameters, whose variations are assumed to follow a Gaussian distribution [10]. In particular a $\sigma/\mu = 5\%$ was set for the cross-section area, while a $\sigma/\mu = 1\%$ was imposed for $t_{OX,B}$, $t_{OX,T}$, and t_{FL} [6, 14, 15].

Fig. 4(a) compares MC simulation results for the 40 nm and the 13 nm DMTJs in terms of switching time (t_p) referred to the worst-case state transition which assures the target WER of 10^{-7}. The 6σ switching time ($\mu + 6\sigma$) considered for architecture-level simulations (discussed below) is also highlighted. According the reported results, the 13 nm DMTJ reduces the 6σ switching time by the 85% as compared to its 40 nm counterpart. Fig. 4(b) is related to MC simulation results for the read operation. Here, we have referred to the conventional voltage sensing scheme [16], which involves applying a fixed read current (I_{read}) to the the bitline (BL) and measuring its voltage (V_{BL}) for both low-resistance state (LRS) and high-resistance state (HRS). The (I_{read}) has to be properly set to ensure a reasonably low read disturbance rate (RDR), which represents the probability of an unwanted switching of the stored bit during the reading operation [17]. In our simulations, the (I_{read}) has been set considering a target RDR of 10^{-9} [16]. From the reading MC results is possible to estimate the voltage sensing margin ($V_{SM} = V_{BL(HRS)} - V_{BL(LRS)}$), along with its 3σ value.

A summary of the bitcell-level MC results is provided in Figs. 5(a)-(b), where speed and energy benefits for the write operation carried out by the 13 nm DMTJ are clearly shown (lower switching time and energy by 6.5 × and 12.3 ×, as compared to the 40 nm device). These advantages are obtained at the expense of about halved sensing margin during the read operation.

3.2. Architecture-level results

DMTJ-based bitcell (BC) characteristics obtained from MC circuit-level simulations at 77 K were used as input to the memory architecture-level DESTINY estimation tool [18]. For the sake of accuracy and to keep consistency with the above circuit-level study, we properly calibrated the DESTINY tool [4], by embedding in its code temperature-dependent transistor (i.e., threshold voltage, on/off current, gate capacitance, mobility, etc.) and interconnection (i.e., resistivity) characteristics extracted from the adopted 65 nm cryogenic PDK in Cadence Virtuoso. This allows an accurate evaluation of the behavior of peripheral memory circuits. DMTJ-based STT-MRAMs have been benchmarked against 6T-SRAM for different cache sizes, ranging from 64 kB up to 2 MB.

Figs. 6(a)-(d) show the STT-MRAM read/write latency and energy results (normalized to conventional 6T-SRAM). As compared to the 40 nm DMTJ-based magnetic memory, the STT-MRAM based on 13 nm DMTJ devices allows write energy savings of about 50% (on average), while ensuring comparable read energy. As a consequence, the use of DMTJ devices with relaxed non-volatility allows the STT-MRAM to outperform its 6T-SRAM counterpart in terms of energy consumption under cryogenic temperatures. Moreover, the STT-MRAM based on 13 nm DMTJ devices considerably reduces

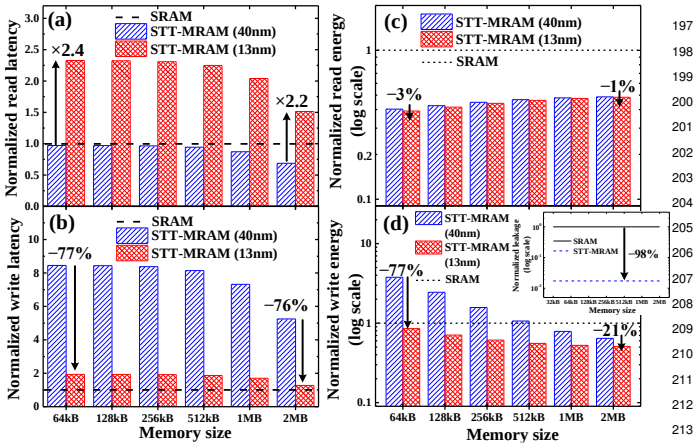

Figure 6: DMTJ-based STT-MRAM (devices with 40 nm and 13 nm diameters) versus 6T-SRAM results at the architecture-level for 77 K: (a) read latency, (b) write latency, (c) dynamic read energy, (d) dynamic write energy (leakage power in the inset). Data is normalized to conventional 6T-SRAM.

write latency of more than 70% in comparison to its 40 nm DMTJ-based counterpart. This allows to significantly reduce the STT-MRAM delay penalty with respect to the conventional 6T-SRAM, under write access. The above benefits come at the only cost of increased read latency (about 2.3 ×) in comparison to the 40 nm DMTJ-based STT-MRAM. Leakage power was also evaluated, as shown in the inset of Fig. 6(d). Note that the STT-MRAMs based on 40 nm and 13 nm DMTJ devices present almost overlapped static power consumption, which is about the 98% less than that of the conventional 6T-SRAM.

4. Conclusions

In this work, we propose to exploit DMTJ devices with relaxed non-volatility to define energy/performance efficient cryogenic caches. Our simulation study performed at circuit and memory architecture-levels shows that by shrinking the DMTJ cross-section area (i.e. by relaxing the non-volatility requirement at room temperature) allows to improve energy and performance under memory write operation, while maintaining adequate retention time at 77 K. In this way, the write delay penalty with respect to conventional 6T-SRAM is greatly reduced, while higher energy-efficiency under both write and read accesses is achieved.

References

[1] IRDS-IEEE, International Roadmap for Devices and Systems - Cryogenic Electronics and Quantum Information Processing (2020). URL http://irds.ieee.org

[2] F. Ware, L. Gopalakrishnan, E. Linstadt, S. A. McKee, T. Vogelsang, K. L. Wright, C. Hampel, G. Bronner, Do Superconducting Processors Really Need Cryogenic Memories? The Case for Cold DRAM, in: Int. Symp. Mem. Syst., 2017, p. 183–188. doi:10.1145/3132402.3132424.

[3] D. Min, I. Byun, G.-H. Lee, S. Na, J. Kim, CryoCache: A Fast, Large, and Cost-Effective Cache Architecture for Cryogenic Computing, in: 25th Int. Conf. Arch. Supp. Program. Lang. Oper. Syst. (ASPLOS), 2020, p. 449–464. doi:10.1145/3373376.3378513.

[4] E. Garzón, R. De Rose, F. Crupi, A. Teman, M. Lanuzza, Exploiting sttmrams for cryogenic non-volatile cache applications, IEEE Transactions on Nanotechnology 20 (2021) 123–128. doi:10.1109/TNANO.2021.3049694.

[5] E. Garzón, R. De Rose, F. Crupi, L. Trojman, M. Lanuzza, Assessment of STT-MRAM performance at nanoscaled technology nodes using a device-to-memory simulation framework, Microelectronic Engineering 215 (2019) 111009. doi:10.1016/j.mee.2019.111009.

[6] E. Garzón, R. De Rose, F. Crupi, L. Trojman, G. Finocchio, M. Carpentieri, M. Lanuzza, Assessment of STT-MRAMs based on double-barrier MTJs for cache applications by means of a device-to-system level simulation framework, Integration - the VLSI Journal 71 (2020) 56–69. doi:10.1016/j.vlsi.2020.01.002.

[7] G. Hu, J. Lee, J. Nowak, J. Sun, J. Harms, A. Annunziata, S. Brown, W. Chen, Y. Kim, G. Lauer, et al., STT-MRAM with double magnetic tunnel junctions, in: 2015 IEEE International Electron Devices Meeting (IEDM), IEEE, 2015, pp. 26–3. doi:10.1109/IEDM.2015.7409772.

[8] G. Hu, J. Nowak, G. Lauer, J. Lee, J. Sun, J. Harms, A. Annunziata, S. Brown, W. Chen, Y. Kim, et al., Low-current spin transfer torque MRAM, in: 2017 International Symposium on VLSI Design, Automation and Test (VLSI-DAT), IEEE, 2017, pp. 1–2. doi:10.1109/VLSI-DAT.2017.7939701.

[9] G. Wang, Y. Zhang, Z. Zhang, J. Nan, Z. Zheng, Y. Wang, L. Zeng, Y. Zhang, W. Zhao, Compact modeling of high spin transfer torque efficiency double-barrier magnetic tunnel junction, in: 2017 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH), IEEE, 2017, pp. 49–54. doi:10.1109/NANOARCH.2017.8053705.

[10] R. De Rose, M. d'Aquino, G. Finocchio, F. Crupi, M. Carpentieri, M. Lanuzza, Compact Modeling of Perpendicular STT-MTJs With Double Reference Layers, IEEE Transactions on Nanotechnology 18 (2019) 1063–1070. doi:10.1109/TNANO.2019.2945408.

[11] D. Apalkov, A. Khvalkovskiy, S. Watts, V. Nikitin, X. Tang, D. Lottis, K. Moon, X. Luo, E. Chen, A. Ong, A. Driskill-Smith, M. Krounbi, Spin-Transfer Torque Magnetic Random Access Memory (STT-MRAM) 9 (2) (May 2013). doi:10.1145/2463585.2463589.

[12] R. De Rose, M. Lanuzza, M. d'Aquino, G. Carangelo, G. Finocchio, F. Crupi, M. Carpentieri, A compact model with spin-polarization asymmetry for nanoscaled perpendicular MTJs, IEEE Transactions on Electron Devices 64 (10) (2017) 4346–4353. doi:10.1109/TED.2017.2734967.

[13] R. Carboni, S. Ambrogio, W. Chen, M. Siddik, J. Harms, A. Lyle, W. Kula, G. Sandhu, D. Ielmini, Modeling of breakdown-limited endurance in spin-transfer torque magnetic memory under pulsed cycling regime, IEEE Transactions on Electron Devices 65 (6) (2018) 2470–2478. doi:10.1109/TED.2018.2822343.

[14] R. De Rose, M. Lanuzza, F. Crupi, G. Siracusano, R. Tomasello, G. Finocchio, M. Carpentieri, Variability-aware analysis of hybrid MTJ/CMOS circuits by a micromagnetic-based simulation framework, IEEE Transactions on Nanotechnology 16 (2) (2016) 160–168. doi:10.1109/TNANO.2016.2641681.

[15] R. De Rose, M. Lanuzza, F. Crupi, G. Siracusano, R. Tomasello, G. Finocchio, M. Carpentieri, M. Alioto, A variation-aware timing modeling approach for write operation in hybrid CMOS/STT-MTJ circuits, IEEE Transactions on Circuits and Systems I: Regular Papers 65 (3) (2017) 1086–1095. doi:10.1109/TCSI.2017.2762431.

[16] K. T. Quang, S. Ruocco, M. Alioto, Boosted sensing for enhanced read stability in STT-MRAMs, in: 2016 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2016, pp. 1238–1241. doi:10.1109/ISCAS.2016.7527471.

[17] Y. Ran, W. Kang, Y. Zhang, J.-O. Klein, W. Zhao, Read disturbance issue and design techniques for nanoscale STT-MRAM, Journal of Systems Architecture 71 (2016) 2–11. doi:10.1016/j.sysarc.2016.05.005.

[18] M. Poremba, S. Mittal, D. Li, J. S. Vetter, Y. Xie, DESTINY: A Tool for Modeling Emerging 3D NVM and eDRAM caches, in: Design, Autom. and Test in Europe Conf. Exh. (DATE), 2015, pp. 1543–1546. doi:10.7873/DATE.2015.0733.