Appears in the Proceedings of the IEEE European Solid-State Circuits Conf. (ESSCIRC-47), Sep. 2021

Layout-Based Evaluation of Read/Write Performance of SOT-MRAM and SOTFET-RAM

Olalekan Afuye, Shady Agwa, Christopher Batten, Alyssa Apsel
School of Electrical and Computer Engineering, Cornell University, Ithaca, NY

***Abstract*—This paper presents a comparison of array-level performance of non-volatile SOT-MRAM and SOTFET-RAM to conventional 6T CMOS SRAM using a specially developed simulation suite that merges physics-based compact models and layout-based parasitic extraction. Unlike prior work, our characterization framework generates a full layout of the memory array including all peripheral logic and routing. The framework uses an industry-standard parasitic extraction tool to generate the full netlist including parasitics which is then simulated using compact models for the appropriate emerging non-volatile device. Using this framework, we show about 1.8× energy savings for total read operations and write operations, and 2× area savings for the SOT-based memories relative to a comparable CMOS SRAM for a 256×128 array size. Our unique full-layout approach also enables important insights that challenge conventional wisdom based on higher-level modeling.**

I. Introduction

A number of non-volatile post-CMOS devices have been proposed over the the last two decades to address slowing scaling trends in CMOS SRAM-based on-chip memory hierarchies. These devices offer smaller bitcells and greater on-chip memory capacity as well as improved latency, throughput, and energy. To characterize these improvements, studies of new devices compare the array-level performance metrics to their corresponding SRAM-based CMOS alternative [1], [2]. Unfortunately, an accurate comparison is challenging due to poor integration of novel devices into process development toolkits (PDK) and conventional electronic design automation (EDA). As a result, prior comparisons usually do not capture layout extracted device and interconnect parasitics.

NVSim [3] and its derivatives are popular tools to analytically estimate the parasitics and extrapolate array-level performance based on single bitcell and sense amp characterization. Other studies use schematic-based simulations with simplified compact models and rough parasitics estimates. Characterization frameworks like these, based on single-bitcells, sometimes ignore layout considerations resulting in overly optimistic models and unrealistic performance estimates since the modeled design points are not actually functionally verified. Single bitcell layouts routinely ignore the impact of access device contacts or use minimum-width bitlines which would, for instance, result in significant IR drop or very slow RC response from poly routing. These frameworks often use first-order CMOS models for all device sizes and configurations leading to inaccuracies in scaled CMOS processes.

In this study, we construct a full array-level simulation leveraging both compact models and layout extracted parasitics which addresses some of these inadequacies and provides a more realistic estimate of the performance metrics of the proposed devices. To that end, we evaluate the performance of two emerging post-CMOS devices by generating full layouts of the RAM memories and simulating the post-layout extracted netlists using their compact models. We then compare their performance to a similarly generated CMOS-based SRAM.

Our specific contributions include: (1) the first quantitative performance evaluation of a RAM memory based on the recently proposed SOTFET device; (2) a layout-based performance characterization of SOT-MRAM and SOTFET-RAM using *in-situ* compact models and post-layout netlists extracted using an industry-standard parasitic extraction tool; and (3) a detailed comparison of SOT-MRAM, SOTFET-RAM, and standard CMOS-based SRAM showing potential area and energy savings in the SOT-based memories for both read and write operations.

A. SOT-MRAM and SOTFET-RAM

In this paper, we explore spin-orbit-torque magnetoresistive random access memory (SOT-MRAM) [4] and the recently proposed spin-orbit-torque field-effect transistor RAM (SOTFET-RAM) [5], [6].

1) SOT-MRAM: The SOT-MRAM is a three-terminal device consisting of two access CMOS transistors and an SOT magnetic tunnel junction (MTJ) device as shown in Fig. 1. The MTJ consists of a tunnelling oxide barrier between a free magnetic layer and a pinned magnetic layer. The direction of magnetization of the free layer determines the resistance of the tunnel junction through the tunnel magnetoresistance effect (TMR). If the pinned and free layers are in parallel orientations, the MTJ has a relatively lower resistance than when both layers are in anti-parallel orientations.

Fig. 1(a) and (b) show the SOT-MTJ and SOT-MRAM devices respectively. CMOS transistor *M1* gates write operations through the *WWL* wordline signal and another CMOS transistor *M0* gates read operations through the *RWL* signal. To write to the SOT-MRAM, charge current is passed through the SO layer by enabling *WWL* and connecting *BL* and *SL* to *VDD/0* or *0/VDD* depending on the desired magnetization direction. Reading is done by passing current through the tunnel junction and comparing the voltage to a reference voltage generated by adjacent reference cells. Since the resistance of the MTJ is determined by the magnetization state of the free layer, the detected voltage across the junction is dependent on the previously written magnetization.

2) SOTFET-RAM: As proposed in [5] and shown in Fig. 1, the SOTFET is a four-terminal device with a similar write mechanism to the the SOT-MTJ. Charge current through the SO layer generates a spin current in the SOT layer which exerts a spin orbit torque on ferromagnet (FM) layer. The FM's magnetization couples to the multiferroic (MF) layer through exchange coupling whose polarization is in turn assumed to be strongly coupled to its magnetization due to the Dzyaloshinskii–Moriya-Interaction (DMI). The multiferroic's polarization gates the semiconductor channel with current flowing when the magnetization state is $+1$ and no current flowing when the magnetization state is -1. Reading though a semiconductor channel results in transistor-like on-off ratios in contrast to the SOT-MTJ whose TMR is in the single digits.

In the SOTFET-RAM, *M0* and *M1* gate read and write operations respectively. The write procedure is the same as the SOT-MRAM while read operations can use a traditional precharge-based SRAM-like sense amplifier owing to the SOTFET's high on-off ratio. The simulation methodology described in this paper enables evaluating both precharge- and zero-precharge-based sensing schemes to show the corresponding trade-off between read energy and read latency.

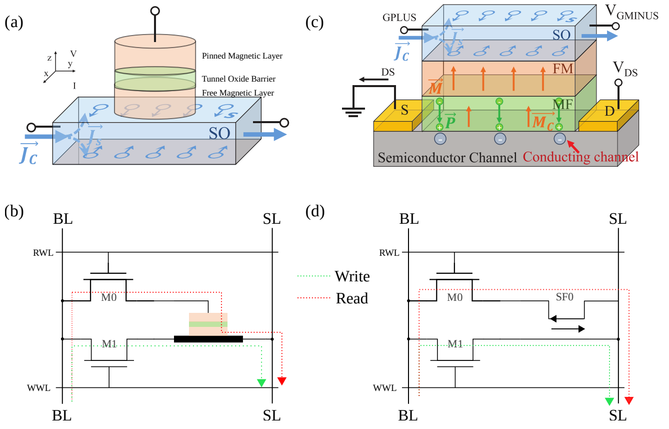

Fig. 1. Devices and MRAM bitcell schematics for SOT-MRAM and SOTFET-RAM. (a) shows the SOT-MTJ device which is of a MTJ on a spin orbit material (b) shows the SOT-MRAM consisting of the SOT-MTJ and access transistors *M0* and *M1*, (c) shows SOTFET device consisting of a SO material-ferromagnetic material-multiferroic material- gate stack on a semiconductor channel (d) shows the bitcell for the SOTFET-RAM including access transistors *M0* and *M1*

II. Benchmarking Methodology

A. Memory Generation in OpenRAM

The simulation tools used in this evaluation are adapted from OpenRAM [7], an open-source SRAM memory compiler. For each technology, we create a prototypical unit bitcell layout. Based on the bitcell layout and associated control logic needed, we generate the layout and the corresponding spice netlist of additional peripherals including the precharge cells, write driver array, sense amp array, wordline drivers, row and column decoders, and their associated control buffers.

To characterize the generated memory, we run design rule checks (DRC) and layout versus schematic (LVS) checks to ensure correctness. We then run a parasitic extraction (PEX) of the layout to produce a fully extracted spice netlist for simulation. The extracted netlist is post-processed to inject the compact model for the non-volatile device under consideration. We then simulate the post-processed netlist to measure the delay and energy consumption of the generated memory.

B. Compact Models for Simulation

We inject the compact model for each technology into individual bitcells by modifying the extracted netlist to instantiate the device model instance. Through this mechanism, we are able to retain the CMOS devices and interconnect parasitics while using the compact models for the simulation to ensure that we are verifying functionality using realistic models. For instance, if the post-layout magnitude or pulse width of the current generated by the write driver for the SOT-MRAM is below the critical threshold switching current (e.g., due to column muxes, interconnect resistance and capacitance) a write error is detected during post-simulation analysis.

1) SOT-MRAM Model: For the SOT-MRAM, we model the write operation using the LLGS equation as described and implemented in [8], [9]. The precessional motion of the magnetization $\vec{\mathbf{m}}$ of the free layer is described by

$$\frac{d\vec{\mathbf{m}}}{dt} = -\gamma\mu_0\vec{\mathbf{m}} \times \mathbf{H}_{\mathit{eff}} + \alpha\left(m \times \frac{d\vec{\mathbf{m}}}{dt}\right) + \frac{\gamma}{M_s}\vec{\tau}_{sot}$$

where $\mathbf{H}_{\mathit{eff}}$ is the effective magnetic field acting on $\vec{\mathbf{m}}$ and $\vec{\tau}_{sot}$ is the spin orbit torque term due to the spin current. $\vec{\tau}_{sot} = \frac{\hbar}{2e}\frac{J_c}{t}\left(\theta_{AD}\vec{\mathbf{m}} \times (\vec{\mathbf{m}} \times \vec{\mathbf{m}}_p) + \theta_{FL}\vec{\mathbf{m}} \times \vec{\mathbf{m}}_p\right)$ where J_c is the charge current density. θ_{AD} and θ_{FL} represent the charge current to spin current conversion efficiency for the anti-damping and field-like torques respectively. Joule heating and thermal noise effects are not included in the model to reduce model complexity. During read operations, the resistance of the MTJ is estimated as a function of the magnetization state and the bias voltage as described in [10].

2) SOTFET-RAM Model: Since the write mechanisms for the SOT-MRAM and the SOTFET-RAM are similar, we use the same LLGS-based model and parameters for both devices. The read model is an adaptation of the framework developed in [11] for the FeFET in which the charge in the multiferroic, MF, is equated to the charge in the MOSFET channel. The charge in the MF is modeled as

$$Q_{MF}(V_{MF}) = P_s \cdot m_z + V_{MF} \cdot C_{ferro} = Q_{MOS}(V_{MOS})$$

where m_z is the magnetization of the ferromagnet as in the SOT-MRAM model, P_s is the saturation polarization of the MF and C_{ferro} is the equivalent linear capacitance of the MF. Our model essentially replaces the Preisach model used in [11] for tracking the ferroelectric charge in the FeFET model with the LLGS equation for tracking the magnetization state m_z. Additional details of the SOTFET model are presented in [6].

C. LLGS Model Parameters

With the exception of α and the magnitude of the electric field used, the parameters are the same as those in [12] and [13]. While [13] reports that the parameters were verified with experimental data, several experimental studies [14] have reported that the single-domain based LLGS models tend to over-estimate the critical switching currents. Furthermore, the default value of α in [12] results in a severely under-damped transient response which can lead to write errors and much slower write times. We therefore select α and H_{ext} corresponding to a switching current of about 110 µA and switching time of about 500 ps consistent with the range of switching currents and times published in the spin orbit torque material exploration study in [2].

D. Sensing Schemes

Fig. 4 shows the sense amplifiers used for all four sensing schemes. The CMOS SRAM sense amp in Fig. 4(a) is a standard precharge based sensing amplifier in which the bitlines are precharged during the first half of the read cycle and the bitlines are conditionally discharged depending on the data in the bitcells during the second half of the cycle. For the SOT-MRAM, we equalize and discharge both bitlines to zero during the first half of the read cycle. In the second half of the read cycle, the bitline BL is charged up to a data dependent final voltage through the sense amp. We use the same sense amp design as in [15] as shown in Fig. 4(b) in which the final bitline voltage is compared to a reference voltage generated by two reference columns storing complementary data. The SOTFET-RAM zero-precharge sensing scheme as shown in Fig. 4(c) is similar to the SOT-MRAM except the reference voltage is generated externally. Finally, the precharged SOTFET-RAM sensing scheme shown in Fig. 4(c) is similar to the SRAM sensing scheme. In the first cycle, BL is precharged to VDD while SL is discharged. BL is then conditionally discharged depending on the stored data and compared to an externally generated reference voltage.

III. Results

Our framework enforces that layout considerations are taken into account at design time. All measurements were performed by reading

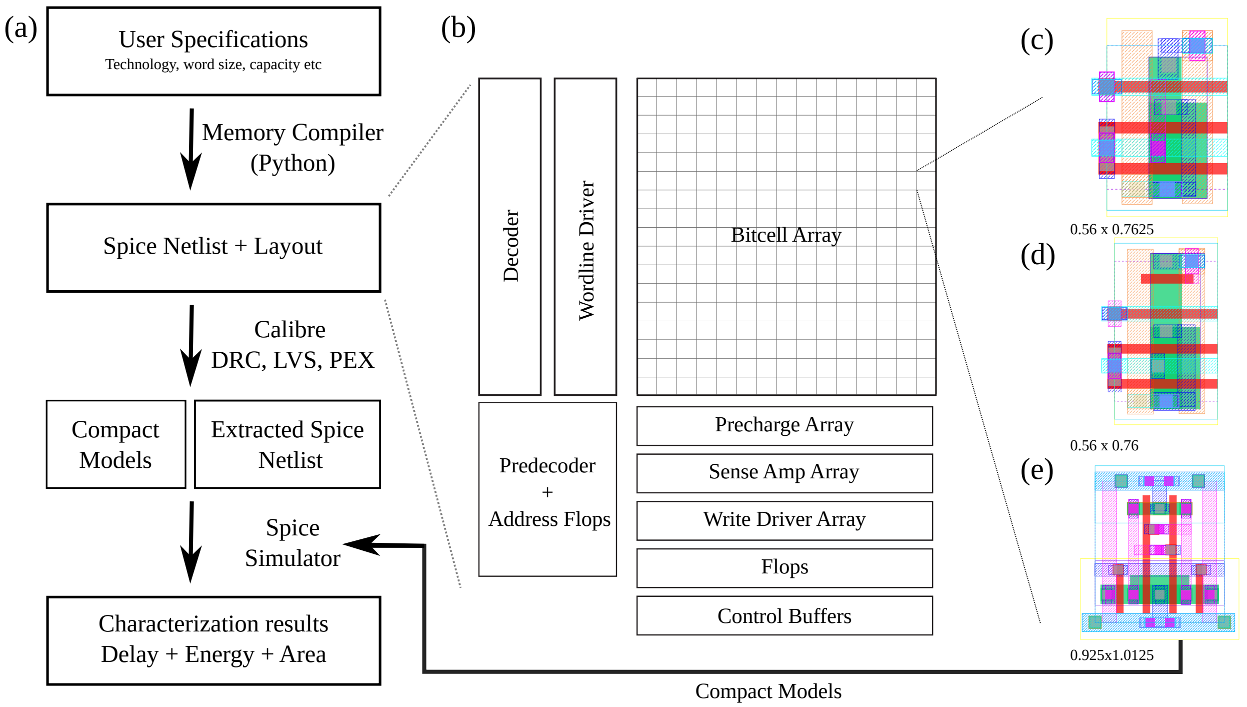

Fig. 2. OpenRAM Block Diagram: (a) shows the characterization flow, (b) shows a sample layout floor plan for the generated memory, (c), (d) and (e) show bitcells for SOT-MRAM, SOTFET-MRAM and 6T-SRAM, respectively.

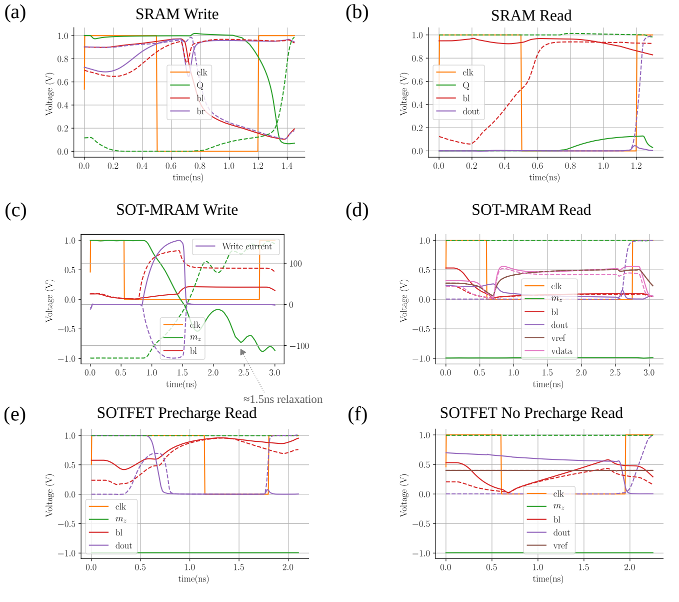

Fig. 3. Simulation waveforms for read and write operations. Dashed lines represent $Q = 1$ or $m_z = +1$; solid lines represent $Q = 0$ or $m_z = -1$.

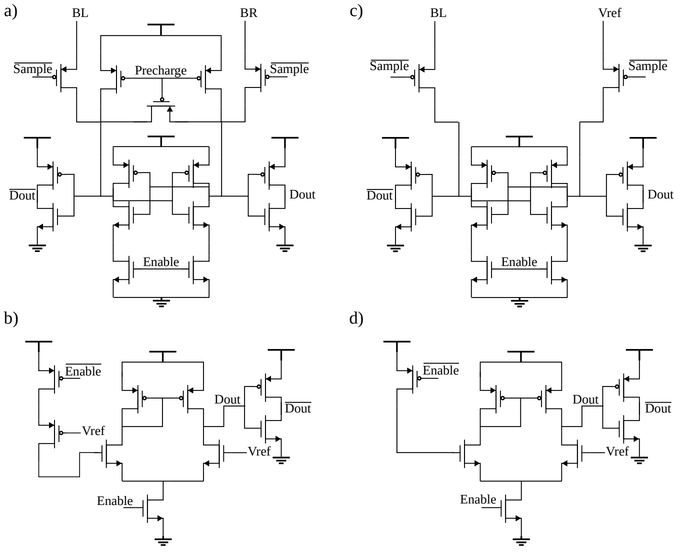

Fig. 4. Sense amplifiers: (a)-(d) show sense amplifiers for CMOS SRAM, SOT-MRAM, Precharge SOTFET, and Zero-Precharge SOTFET sense amps respectively

and writing the same data patterns to the memories and then reducing the cycle times until read/write errors occur; ensuring that the final design is functional. For example, both SOT-MRAM and SOTFET-RAM bitcell widths were determined by the wider metal4 layer pitch required for supplying the large write current. This is in contrast to previously published SOT-MRAM bitcell widths which usually use a minimum metal1/metal2 pitch or access transistor width.

Table I shows the performance metrics for the three technologies. The delay numbers in parenthesis represent the total delay including the magnetization relaxation after the initial current driven switch as depicted in Fig. 3(c). This additional delay (≈1.5ns) is only relevant for successive read, write operations to the same address. CMOS SRAM shows the lowest total delay at the cost of increased read energy relative to the zero-precharge based reading schemes.

Read and write energies were estimated by averaging the energy from running multiple read and write operations to random addresses in a random operation order. The memory is initialized with random

TABLE I
SRAM, SOT-MRAM, and SOTFET-RAM Performance Metrics

	CMOS	SOT-MRAM	SOTFET-RAM No Precharge/ Precharge
Read Delay (ns)	1.3	2.75	1.95/1.8
Write Delay (ns)	1.3	1.25 (2.75)	1.25 (2.75)
Read Energy (pJ)	21	7	9/45
Write Energy (pJ)	15	13	12/15
Total Energy (pJ)	36	20	21/60
Leakage Power (µW)	919	43	38
Area (µm^2)	42348	21280	23211

Array size = 128 rows × 256 columns; word size = 32 bits

data and subsequent write operations also write random data patterns. We have included the total energy to augment comparative analysis since overlap between consecutive read and write operations make it impossible to completely isolate the read and write energies. Our results spotlight the potential energy savings of the SOT based memories (and some other post-CMOS devices) which do not require precharge operations for correct operation. The standard 6T SRAM requires a precharge for both read and write operations to prevent accidental data override in deselected columns. For the SOT-MRAM and SOTFET-RAM, energy is only expended on the columns selected for read/write. This result holds even for small arrays and is in contrast to the results in [1] which show energy savings for the SOT-MRAM for large arrays only. The energy savings take advantage of zero-precharge-based sensing schemes which are not directly supported in traditional characterization tools (e.g., NVSim).

In comparison to the SOT-MRAM, the SOTFET-RAM can operate in a slightly higher energy consuming but faster zero-precharge read mode or operate in an even faster but also much more energy hungry precharge-based sensing mode. The faster zero-precharge read is due to the higher on-off ratio of the SOTFET versus about 150% TMR ratio for the SOT-MTJ. The read energy of the precharge based SOTFET-RAM is high due to the relatively large access devices (and corresponding parasitic bitline capacitance) required for driving the large write current.

Area savings are about 2× for the chosen array size but should be higher for larger array sizes up to a limit of the unit bitcell ratio. The SRAM bitcell is 2.19x bigger than the SOT-MRAM while the SOTFET-RAM bitcell is the same size as the SOT-MRAM despite the additional device since the drain of the additional device is shared across bitcell array rows while the SOT-MRAM requires space between adjacent rows. Finally, the leakage energy for the SOT-based memories are 21× times lower than those for the SRAM since leakage energy is dissipated only in the peripherals. The leakage energy savings should be even larger for bigger arrays and will be zero if power gating is employed for the non-volatile memories.

We note that the results presented only represent a design point in the energy-delay design landscape. Our choice of the older 45nm FreePDK process development kit (PDK) reflects our intention to publish the modified OpenRAM framework, device models and layouts in a future publication without violating non-disclosure agreements. For example, our internal OpenRAM generated push-rule-based SRAM generated in a 28nm process shows much lower delay and energy metrics which will not necessarily be reflected in the SOT-based memories since thermal stability concerns might limit MRAM technology scaling.

IV. Conclusion

In this paper, we evaluated and compared the read and write latencies and energies of the SOT-MRAM and SOTFET-RAM to a conventional 6T CMOS SRAM. Our results show potential energy savings for both read and write operations of the SOT-based memories even for low array sizes.

Acknowledgment

This work was supported in part by NSF E2CDA Award #1740136, the Semiconductor Research Corporation (SRC) as nCORE task 2758.002 and 2758.004. We acknowledge and thank Xiang Li, Huili Xing and Daniel Ralph for their helpful discussions on SOT and SOTFET modeling.

References

[1] F. Oboril, R. Bishnoi, M. Ebrahimi, and M. B. Tahoori, “Evaluation of hybrid memory technologies using SOT-MRAM for on-chip cache hierarchy,” *IEEE Trans. on Computer-Aided Design of Integr. Circuits and Syst.*, vol. 34, no. 3, pp. 367–380, 2015.

[2] Y.-C. Liao, P. Kumar, D. Mahendra, X. Li, D. Zhang, J.-P. Wang *et al.*, “Spin-orbit-torque material exploration for maximum array-level read/write performance,” in *IEEE Int’l Electron Devices Meeting*, 2020, pp. 13–6.

[3] 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. on Computer-Aided Design of Integr. Circuits and Syst.*, vol. 31, no. 7, pp. 994–1007, 2012.

[4] P. Gambardella and I. M. Miron, “Current-induced spin–orbit torques,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 369, no. 1948, pp. 3175–3197, 2011.

[5] X. Li, J. Casamento, P. Dang, Z. Zhang, O. Afuye, A. B. Mei *et al.*, “Spin–orbit torque field-effect transistor (SOTFET): Proposal for a magnetoelectric memory,” *Applied Physics Letters*, vol. 116, no. 24, p. 242405, 2020.

[6] O. Afuye, X. Li, F. Guo, D. Jena, D. C. Ralph, A. Molnar *et al.*, “Modeling and circuit design of associative memories with spin–orbit torque fets,” *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 5, no. 2, pp. 197–205, 2019.

[7] M. R. Guthaus, J. E. Stine, S. Ataei, B. Chen, B. Wu, and M. Sarwar, “OpenRAM: an open-source memory compiler,” in *IEEE/ACM Int’l Conf. on Computer-Aided Design*, 2016, pp. 1–6.

[8] J. Z. Sun, “Spin-current interaction with a monodomain magnetic body: A model study,” *Physical Review B*, vol. 62, no. 1, p. 570, 2000.

[9] P. Bonhomme, S. Manipatruni, R. M. Iraei, S. Rakheja, S.-C. Chang, D. E. Nikonov *et al.*, “Circuit simulation of magnetization dynamics and spin transport,” *IEEE Trans. on Electron Devices*, vol. 61, no. 5, pp. 1553–1560, 2014.

[10] 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. on Nanotechnology*, vol. 16, no. 3, pp. 387–395, 2017.

[11] K. Ni, M. Jerry, J. A. Smith, and S. Datta, “A circuit compatible accurate compact model for ferroelectric-FETs,” in *IEEE Symp. on VLSI Technology*, 2018, pp. 131–132.

[12] I. Ahmed, Z. Zhao, M. G. Mankalale, S. S. Sapatnekar, J.-P. Wang, and C. H. Kim, “A comparative study between spin-transfer-torque and spin-hall-effect switching mechanisms in PMTJ using SPICE,” *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 3, pp. 74–82, 2017.

[13] J. Kim, A. Chen, B. Behin-Aein, S. Kumar, J.-P. Wang, and C. H. Kim, “A technology-agnostic MTJ SPICE model with user-defined dimensions for STT-MRAM scalability studies,” in *IEEE Custom Integrated Circuits Conf.*, 2015, pp. 1–4.

[14] K. Garello, C. O. Avci, I. M. Miron, M. Baumgartner, A. Ghosh, S. Auffret *et al.*, “Ultrafast magnetization switching by spin-orbit torques,” *Applied Physics Letters*, vol. 105, no. 21, p. 212402, 2014.

[15] J. Kim, K. Ryu, S. H. Kang, and S.-O. Jung, “A novel sensing circuit for deep submicron spin transfer torque MRAM (STT-MRAM),” *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 1, pp. 181–186, 2010.