Ultra-fast and high-reliability SOT-MRAM: from cache replacement to normally-off computing

Guillaune Prenat, Kotb Jabeur, Pierre Vanhauwaert, Gregory Di Pendina, Fabian Oboril, Rajendra Bishnoi, Mojtaba Ebrahimi, Nathalie Lamard, Olivier Boulle, Kévin Garello, et al.

Ultra-Fast and High-Reliability SOT-MRAM: From Cache Replacement to Normally-Off Computing

Guillaume Prenat, Kotb Jabeur, Pierre Vanhauwaert, Gregory Di Pendina, Fabian Oboril, Rajendra Bishnoi, Mojtaba Ebrahimi, Nathalie Lamard, Olivier Boulle, Kevin Garello, Juergen Langer, Berthold Ocker, Marie-Claire Cyrille, Pietro Gambardella, Mehdi Tahoori, and Gilles Gaudin

Abstract—This paper deals with a new MRAM technology whose writing scheme relies on the Spin Orbit Torque (SOT). Compared to Spin Transfer Torque (STT) MRAM, it offers a very fast switching, a quasi-infinite endurance and improves the reliability by solving the issue of "read disturb", thanks to separate reading and writing paths. These properties allow introducing SOT at all-levels of the memory hierarchy of systems and adressing applications which could not be easily implemented by STT-MRAM. We present this emerging technology and a full design framework, allowing to design and simulate hybrid CMOS/SOT complex circuits at any level of abstraction, from device to system. The results obtained are very promising and show that this technology leads to a reduced power consumption of circuits without notable penalty in terms of performance.

1 Introduction

For several years, the scaling of microelectronics has been facing physical limits, mainly due to leakage currents, heating issues and process variations. For the performance and capabilities of circuits to keep on improving, several solutions are investigated, from device to system levels. For instance, the use of innovative devices besides or in replacement of standard CMOS (refered as "More than Moore") is a promising concept.

Among these new technologies, STT-MRAM (for Spin Transfer Torque Magnetic Random Access Memory) seems particularly promising and is studied by most of the major microelectronics companies. It is an emerging memory technology, which combines non-volatility with high writing and reading speed, low-power consumption, high density and a high endurance. This unique set of performance allows integrating this technology in the memory hierachy of systems to reduce the power consumption without degrading the performance, or even offering new functionalities. However, some limitations of STT-MRAM make it difficult to be used for applications requiring very high operating speed or in very low levels of the memory hierarchy.

In this paper, we present an emerging technology called SOT-MRAM (for Spin Orbit Torque Magnetic Random Access Memory), whose properties allow addressing specific applications that could not be addressed easily by STT-MRAM. In order to be able to design hybrid CMOS/SOT circuits, it is necessary to introduce the SOT technology in the standard design flows of microelectronics. The paper is organized as follows: the first part is dedicated to the description of the technology and elementary device as well as the underlying physics. The second part describes a compact model of the device for electrical simulations of circuits embedding the SOT technology. The third part presents a full design framework for the evaluations of systems embedding SOT-MRAM in the memory hierarchy and promising results of evaluations using it. Compared to previous system-level studies using MRAM for cache memories, this work is the first that compares perpendicular STT-MRAM and SOT-MRAM for single- and multi-core processors in terms of area, energy efficiency and performance. In contrast, previous work focused either on single-core processors with small (a few MByte) last-level caches and/or the much slower and less efficient in-plane STT-MRAM.The last part deals with the introduction of SOT in the logic itself to further reduce the total power consumption.

- *G. Prenat, K. Jabeur, P. Vanhauwaert, G. Di Pendina, N. Lamard, M.C. Cyrille, G. Gaudin, and O. Boulle are with the Univ. Grenoble Alpes, INAC-SPINTEC, F-38000 Grenoble, France; CNRS, INAC-SPINTEC, F-38000 Grenoble, France; and CEA, INAC-SPINTEC, F-38000 Grenoble, France. E-mail: guillaume.prenat@cea.fr.*
- *F. Oboril, R. Bishnoi, M. Ebrahimi, and M. Tahoori are with Karlsruhe Institute of Technology (KIT), Kaiserstr. 12, D-76131 Karlsruhe, Germany. E-mail: {oboril, bishnoi, tahoori}@ira.uka.de.*
- *K. Garello and P. Gambardella are with the Department of Materials, ETH Zürich, Hönggerbergring 64, CH-8093, Zürich, Switzerland. E-mail: {kevingarello, pietro.gambardella}@mat.ethz.ch.*
- *J. Langer and B. Ocker are with Singulus Technologies, D-63796 Kahl, Germany.*

2 Spin Orbit Torque MRAM technology

The STT-MRAM (Spin Transfer Torque Magnetic Random Access Memory) has been identified as one of the most

promising candidates among emerging memories [1], [2]. It is an electrically addressable memory combining non-volatility, fast read and write operations and low writing energy [3], [4]. Most of the major microelectronics and memory companies are now developing STT-MRAM schemes for embedded as well as stand-alone applications. The 0 and 1 states of an STT-RAM are defined by the relative orientation (parallel or antiparallel) of the magnetization in two magnetic layers separated by a nonmagnetic material. One of the two magnetic layers is the storage layer (or the "free" layer) while the other one is called the reference layer. The read operation is performed using the magnetoresistance signal: the electrical resistance of the stack is higher (lower) if the magnetizations of the two layers are antiparallel (parallel). Larger signals are obtained with the tunnel magnetoresistance (TMR) using an insulating material in the non-magnetic layer, the tunnel barrier. The writing operation relies on the transfer of spin angular momentum from the reference layer to the free layer, mediated by an electric current. The spins of the conduction electrons are polarized by the reference layer, the magnetization of which is fixed, and transfer their magnetization to the storage layer. If the number of spins (that is the current density) is large enough, this magnetization of the storage layer can be reversibly switched between the two stable states defined by the magnetic anisotropy. In the MRAMs state-of-the-art , the tunnel barrier is a 1-2 nm thick MgO layer and the two magnetic layers are usually CoFeB films with thickness in the range 1-5 nm. The material stack constituting the memory, however, is often much more complicated than this, consisting of different layers that serve to optimize the structure as well as the magnetic and electric properties of the CoFeB/MgO/CoFeB tunnel junction. Today, the reading and the writing operations are very efficient, with more than 100-200 percent of resistance variation in production devices and current densities of the order of $10^{10} A/m^2$ for writing the state of the storage layer. The major limitations of the STT-MRAM come from the fact that the read and write current paths are identical. Undesired writing while reading can happen [5], [6], [7] and the read and write current distributions need to be well controlled. Moreover, the tunnel barrier must have a very low resistance to accommodate the large writing current density, which is achieved by making the barrier thinner. Since the RA product (resistance times the junction area) has to be lower than typically $1\ \Omega.\mu m^2$ for the writing process, the scalability of the STT-MRAM can be questioned for nodes beyond 22 nm. This will require complex material optimization to obtain a large TMR signal [8] and reproducible stack properties. Finally, the STT-MRAM can be switched very rapidly providing that very large current densities are used [9], [10], [11]. Since this current is injected through the tunnel barrier, reliability issues are expected due to the rapid aging of the tunnel barrier [5].

Fortunately, alternative writing schemes of the storage layer have been explored and shown to be very efficient. Among them, the SOT-MRAM [12] naturally solves these issues and is one of the most promising alternatives to STT technology. Here, while the reading mechanism is the same as in the STT-MRAM approach, the writing current is injected in the plane of the storage layer rather than perpendicular to. Hence, the read and write paths are decoupled: no more untimely aging of the tunnel barrier nor undesired writing. Indeed, since the writing current is not injected through the MTJ, the MTJ does not suffer from accelerated aging. This is particularly true for fast switching needed for SRAM replacement in cache memories that require very large current densities. Concerning the read disturb resulting in undesired writing, having two separate paths allows an additional degree of freedom on the choice of the MTJ resistance. In standard STT technology, the writing current is applied through the barrier. To avoid exceeding the breakdown voltage of the oxide barrier during the writing, the choice of the resistance. area (RA) product of the barrier is limited to low values around some $\Omega.m^2$. During the reading, the voltage applied across the barrier could result in a large current density and in consequence to read disturb. In SOT, the value of the RA product is not limited by the writing process since the read and write paths are decoupled. This RA can be increased, resulting, for a given read voltage, in a lower reading current and a better immunity to read disturb. Moreover, the relaxed constraints on the RA are beneficial for the reading. The loss of reading speed resulting from the reduced reading current is compensated by the increase of the resistance: increasing the value of the resistance of the MTJ allows, for the same value of the TMR, to increase the difference of the resistance in the reading path and so the reading signal, speeding up the reading. The geometry of the spin injection also solves the issue of non-negligible incubation time delay typical of STT mechanism (few nanoseconds), limiting ultrafast switching and inducing a broad switching time distribution[13]. Finally, if the magnetization can be switched with current density similar to those used in the STT-MRAM, the total current to be delivered by the transistor is much smaller since the cross section area of injection is now the width of the current line times its thickness (a few nanometers), to be compared with the lateral area of the junction for the STT approach. The writing principle of the SOT-MRAM relies on the transfer of angular momentum from the lattice to the magnetization via the spin-orbit interaction and no more from one magnetic layer to the other as in the STT-MRAM approach. Considering a trilayer with structural inversion asymmetry, typically a thin FerroMagnetic (FM) layer sandwiched between two different materials, one Half Metallic (HM) layer with a large spin orbit coupling and one insulating Oxide layer (Ox), e.g. Pt/Co/AlOx or a Ta/CoFeB/MgO trilayer, an electric current injected in the metallic layer will produce two torques on the magnetization of the magnetic layer. The first one, the field-like term [14], is equivalent to an effective magnetic field H_{FL} whose direction is fixed, in the plane of the layer, perpendicular to the current direction. The second one, the damping-like or antidumping term, is equivalent to a magnetic field H_{DL} whose direction is perpendicular to both the magnetization and H_{FL} [15], [16]. The origin of these torques has been widely discussed. They can both originate from the Inverse Spin Galvanic Effect (ISGE) [17], [18], [19] generated mainly at the FM/HM interface or from the Spin Hall Effect (SHE) [20], [21], [22] generated mainly in the bulk of the non-magnetic HM material [15], [23], [24], [25], [26], [27], [28].

Considering an out-of plane magnetized material, the damping-like torque, whose effective field H_{DL} is perpendicular to the magnetization, is effective in switching the magnetization. However, since H_{DL} is identical for a magnetization

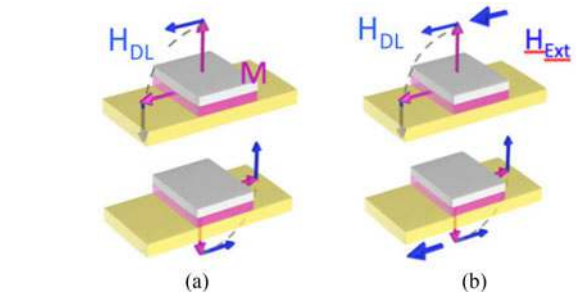


Fig. 1. Schematic of the switching mechanism driven by the effective field. a) Without any applied magnetic field a current flowing in the track (yellow) will destabilize both up and down magnetizations. b) An applied magnetic field Hext will either add or subtract to the effective field H_{DL} and eventually lead to the current induced magnetization switching towards the stable state (down here).

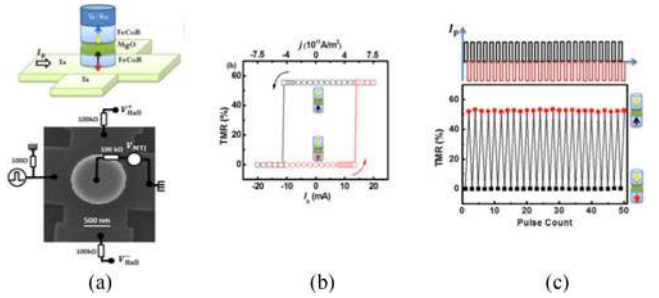


Fig. 2. Proof of concept of the SOT-MRAM. a) Schematic of a memory dot and SEM image of a 1 um dot on top of a 1.3 um Ta line with a schematic of the electrical measurement setup). b) TMR as a function of current pulse amplitude injected in the Ta electrode using 50 ns long pulses under an in-plane magnetic field H = 0.4 kOe. The arrows show the sweep direction of the current. c) (top) Schematic of the pulse sequence. (middle) The AHE resistance, proportional to the Mz component of the bottom FeCoB layer. and (down) TMR measured after the injection of positive (black squares) and negative (red circles) current pulses of amplitude 20 mA and 50 ns long under H = 0.4 kOe (from [39]).

pointing either up or down, both configurations are equally destabilized by this field. Therefore, a constant bias magnetic field applied along the direction of the current is required to break this symmetry, adding or subtracting to H_{DL} depending on the magnetization direction, as shown in Fig. 1, and eventually leading to determinist switching. This bias magnetic field is constant and can be generated by an additional magnetic layer acting like a permanent magnet, with no penalty in terms of power consumption or footprint of the device. This description is very simplified, as switching occurs in reality through a more complex motion of the magnetization or/ and by other mechanism involving the nucleation of a magnetic domain followed by the propagation of the domain wall if the device size exceeds the limit of a single magnetic domain [29], [30], [31].

This switching was first demonstrated in a Co/AlOx dot sitting on top of a Pt current line [15]. It was then evidenced in other HM/FM/Ox systems such as Ta/CoFeB/MgO [16], [32], [33], W/CoFeB [34] and Pt/Co/MgO [35]. The presence of strong SOT was evidenced as well in metallic systems such as Pt/Co/Ni and Ta/NiFe [36], CoPd multilayers [37] and asymmetrical Pt/Co/Pt tri-layers [38].

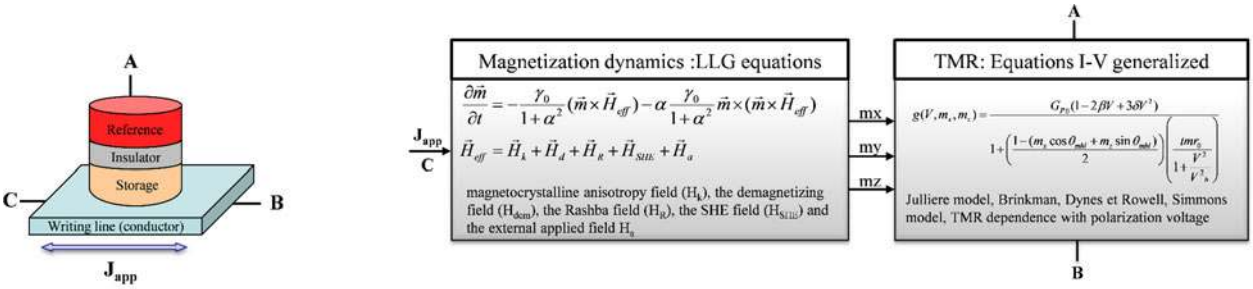
The HM/FM/Ox tri-layers are typically the bottom part of magnetic tunnel junction (MTJ) and need to be completed to form an SOT-MRAM. The difficulty here is to combine large spin orbit torque amplitudes for writing and high TMR for reading, while using materials with perpendicular magnetic anisotropy. We demonstrated the first proof of concept of a SOT-MRAM with out-of-plane magnetization [39] in the framework of the spOt european project [40]. A Ta/FeCoB/MgO/FeCoB/Ta/Ru stack deposited by SINGULUS was patterned into dots sitting on a Ta track and the bipolar switching induced by the injection of a current into the Ta track was monitored by the TMR signal (Fig. 2). TMR up to 90 percent were obtained with a RA product of $1.15\ k\Omega.\mu m^2$. The current density is still high in this first prototype device, around $5.10^{11} A/m^2$ for 20 ns long current pulses. However, the total current is compatible with corresponding CMOS technologies nodes. Moreover, this is a first demonstrator where many parameters can be optimized: thickness of the writing line, design of the line filling, materials used. Recent reports show that W-based or AFM-based stacks show very promising decrease of the critical switching current, but are still under investigation as well as material engineering. Current densities as low as 3-6 $\times\ 10^{10}\ A/m^2$ were reported for 2-3 nm thick Ta using a quasi-DC current [33]. Considering this current density and a 2 nm thick, 50 nm wide Ta line, currents as low as 3-6 μA could be used, already competing with the best results reported for the STT-MRAM [9], [41], [42]. Moreover, only few materials have been tested until now and there is certainly plenty of room for improvement on this side.

In addition to having a very large endurance, the writing of a SOT-MRAM can be very fast, well into the sub-ns regime. We performed a systematic study of the sub-ns switching using 90 nm to 100 nm Co/AlOx dots on top of a Pt line [29]. Deterministic switching, bipolar with respect to either field or current was obtained down to 180 ps (limitation of the experimental setup). These results confirmed that the incubation time is negligibly small [13] and that the switching is consistent with a nucleation process followed by the rapid propagation of the domain wall. SOT-MRAM appears to be a credible candidate for SRAM replacement in cache memory, where fast writing and reliability is required, as well as for logic applications since it combines non volatility, potentially infinite endurance and ultra-fast switching.

3 Compact Modeling of the SOT-MTJ Device

The integration of the SOT device into standard microelectronics design suites is a fundamental step toward the design of hybrid CMOS/MTJ circuits. Therefore an accurate and fast SPICE compact model of the MTJ, i.e. the elementary device, must be used for analog electrical simulations for the hybrid CMOS/magnetic technology, as presented in [43]. This compact model should be provided within the Process Design Kit (PDK), in addition to technology files for layout and physical verification, and standard cells for the design of complex logic circuits. Concerning accuracy and speed, we described in [44] the two possible strategies to efficiently model spintronics while highlighting pros and cons of the different modeling strategies.

In [45], we provided the first compact model written in Verilog-A of a SOT-MTJ. Our choice of the coding language is motivated by the capability of Verilog-A to afford a quick method of enhancing compact models to illustrate new physics of advanced processes. In addition, it is on the path to become the preferred compact modelling language for both



(a) Physical view (b) Block diagram of the compact model

Fig. 3. Modeling strategy of the three-terminal SOT-MTJ.

academic and industrial research groups thanks to its flexibility to run in numerous simulators (Spectre, HSpice, ADS, Eldo) and internal simulators of semiconductor companies. We aim to obtain a straightforward, high-speed and precise electrical representation of the physical behaviour of the SOT-MTJ device. First, we analyse the model equations along with some approximations. Then, thanks to close interactions with Spintec technologists and physicists, a number of associated parameters are fed into these equations. The simulator represents the established equations as equivalent circuit elements. In order to develop this model, we proceed under the macrospin approximation. We consider that the magnetization of each ferromagnetic layer is uniform (single domain). Hence, it can be described by a single macroscopic magnetic moment. This hypothesis significantly abridges the mathematical analysis. The smaller the sample used is, the more pertinent the macrospin assumption is.

Fig. 3 illustrates the strategy and the equations used to describe the SOT-MTJ device. The memory cell is described as a three-terminal logic device and includes the dynamic behavior described by the Landau-Lifshitz-Gilbert model (LLG) [46]. To follow the variation of the SOT-MTJ resistance, the Julieres model [47] as well as the Simmons model [48] were used in the expression describing the conductance through the junction (Tunnel Magneto-Resistance). Moreover, for an improved accuracy, we integrated the dynamic conductance given by the Brinkman's model [49] and we took in consideration the dependence of magneto-resistance on bias voltage. Finally, a special interest has been given to damping-like and field-like spin-orbit torques inside the LLG equation to highlight the impact of these two factors on the dynamic of magnetization switching intensively argued in [15] and [16]. Further details about the choice of parameters and the integration of the applied current (Japp) in the equations are available in [45]. Also, examples of the model validation for circuit design are presented in [45], [50] and [51].

Fig. 4a describes the theoretical switching of the magnetization m_z from parallel P to AP and vice versa depending on the current direction applied during the time. Fig. 4b shows the simulation results of the SOT-MTJ model which obviously corresponds to the theoretical behavior of the device and where we clearly observe the oscillations (from the LLG equations) during the switching (write phase).

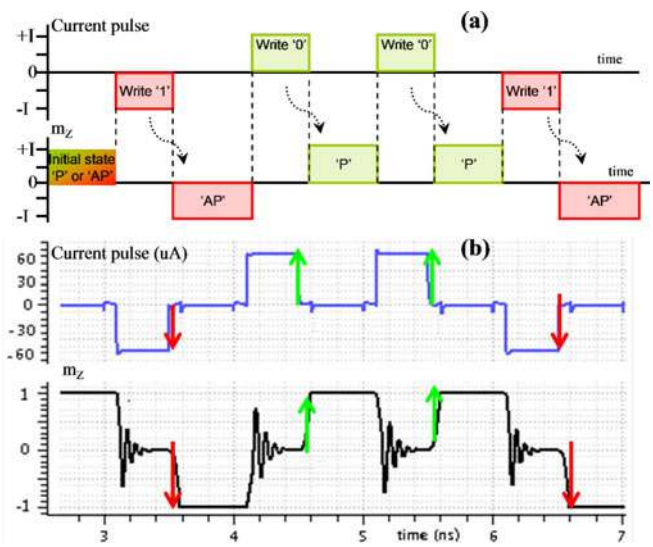


Fig. 4. SOT-MRAM model validation. (a) Theoretical behavior. (b) Model behavior.

4 Evalution of SOT-MRAM for Caches

In order to evaluate if SOT-MRAM can replace SRAM as memory technology for microprocessor caches, it is necessary to abstract the device-level information for a single MTJ cell all the way up to system-level. Therefore, in this section, we present a unique cross-layer MRAM analysis platform shown in Fig. 5, which allows us to analyze system-level implications of device-level changes. Moreover, a comprehensive system-level study is presented including the evaluation of single- and multi-core microprocessors.

4.1 Device-To-Memory Level Abstraction

Based on the device-level evaluation platform described in the previous section, we built a cross-layer analysis framework depicted in Fig. 5. This framework allows us to explore SOT-MRAM for various memory arrays and its feasibility for microprocessor caches. For this purpose, after a device-level evaluation is conducted for a single bit-cell

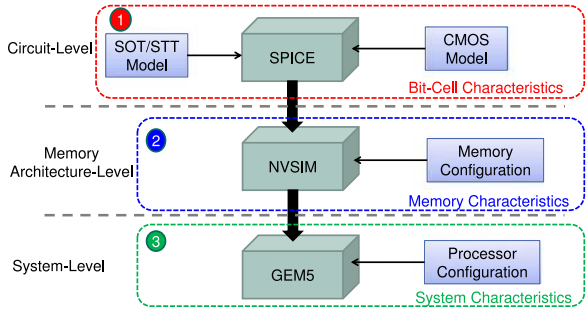


Fig. 5. Overview of the cross-layer analysis platform.

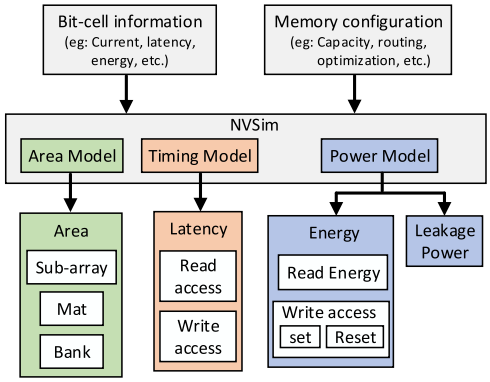


Fig. 6. Overview of the circuit-level part of our cross-layer analysis platform.

using SPICE simulations and the device-level models from the previous section, a circuit-level analysis is performed for the complete memory array. Therefore, we have chosen NVSim [52], which allows predicting circuit-level performance, energy, and area models for various memory technologies such as SRAM, NAND-Flash and STT-MRAM. Moreover, we modified NVSim to support SOT-MRAM and the asymmetric write behavior (set versus reset) of STT-MRAM. The inputs to NVSim are device-level parameters such as switching energy and latency as well as details about the memory organization such as memory capacity, routing, partitioning and optimization constraints. Based on these information NVSim extracts the read and write access latencies for the given memory architecture, the per-access read and write energy, the leakage power as well as the area of the memory array, as shown in Fig. 6.

The corresponding results normalized to a 6T-SRAM technology for various memory sizes for SOT-MRAM and perpendicular STT-MRAM are presented in Fig. 7, whereas a comparison with in-plane STT-MRAM can be found in [53]. The underlying bit-cell parameters that we employed in this study can be found in Table 1. We obtained these parameters by performing a comprehensive SPICE analysis. For this purpose, we implemented a single MRAM bit-cell with one MTJ cell and one access transistor as well as the read and write circuitry using the TSMC 65 nm library and the SOT model presented in Section 3. For STT-MRAM we employed the model presented in [3]. The read and write circuitry are presented in [53]. As it can be seen, for small memory sizes SRAM is the better choice compared to SOT-MRAM. It offers better access latencies, smaller area and lower access energies. However, with increasing memory capacity, SOT-MRAM becomes more efficient due to the following aspects: 1.) The SOT-MRAM bit-cell is smaller than an SRAM bit-cell, and thus, as

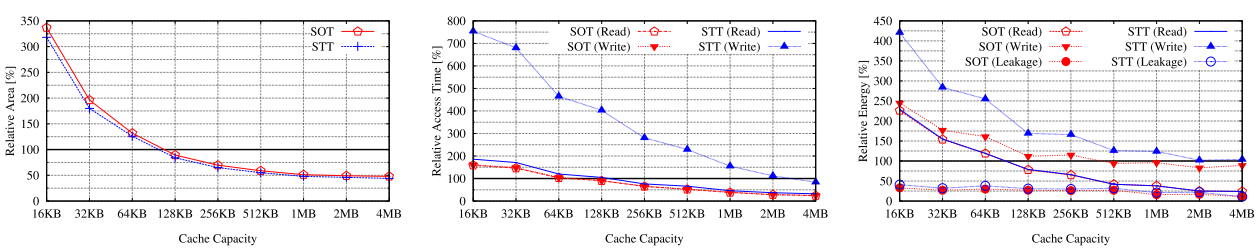
TABLE 1
Comparison of SOT-MRAM and Perpendicular STT-MRAM for a Single Bit-Cell

	SOT-MRAM	STT-MRAM
Read Latency [ps]	221	226
Write Latency [ps]	266	4140 (AP) / 2610 (P)
Write Current [uA]	100	150 (AP) / 93 (P)
Read Energy [pJ]	1.8	1.8
Write Energy [pJ]	0.1	0.3 (AP) / 0.3 (P)
CMOS Technology	TSMC 65 nm Typical	

soon as the bit-cell array dominates the total area and not the periphery circuitry, SOT-MRAM provides the smaller area. In contrast, for small memory sizes, the periphery (e.g., sense amplifier, write circuitry) dominates, and as a result SOT-MRAM consumes more area due to the sophisticated read and write circuitry. 2.) Due to the same reason, the interconnect (routing) delays are less important in SOT-MRAM memories, and consequently, the access latencies do not increase as much as it is the case for SRAM for increasing memory sizes. 3.) As the interconnects also influence the per-access energy, SOT-MRAM also offers a lower per-access energy than SRAM for larger memory sizes. In our analysis, the turning point is around 128 KByte which corresponds to the size of a small L2-cache. Please note that SOT-MRAM is always the best choice in terms of leakage power, as only the periphery circuits suffer from leakage, whereas in SRAM also the bit-cells contribute to the overall leakage power. In addition, it is worth to note that SOT-MRAM is also slightly better than perpendicular STT-MRAM in terms of access latencies and energy consumption as shown in Fig. 7. However, it requires slightly more area due to the additional terminal compared to STT-MRAM, as laid out in [53].

4.2 Memory-To-System Level Abstraction

The data from NVSim is then used by the next tool in our cross-layer framework to evaluate the implications of different memory technologies at system-level, where these memory technologies are used for microprocessor caches at different levels. Therefore, we employ gem5, a cycle-accurate performance-simulator [54], which supports various memory configurations and allows to configure all relevant cache parameters such as capacity, associativity, latency, block size and policy. Furthermore, gem5 enables an evaluation of various microprocessor architectures ranging from low-power single-core embedded processors to high-performance many-core solutions. In order to support the asymmetric read and write behavior of SOT-and STT-MRAM, we modified gem5 accordingly. The output of gem5 are the overall system



(a) Area (b) Access Latencies (c) Read/Write Energy & Leakage

Fig. 7. Scaling behavior of SOT-MRAM and STT-MRAM normalized to SRAM for various memory sizes.

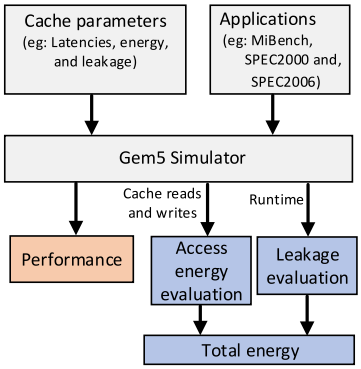


Fig. 8. Overview of the system-level part of our cross-layer analysis platform.

performance (e.g. runtime) and cache statistics such as the number of read and write access per cache. This is then used to calculate the total energy consumption as illustrated in Fig. 8.

4.3 Experimental Evaluation

Using our cross-layer analysis framework with the microprocessor setup detailed in Table 2, we conducted various experiments to compare SOT-MRAM with SRAM and perpendicular STT-MRAM. To evaluate the impact of SOT-MRAM at system-level, and to analyze the energy consumption of different cache technologies under realistic conditions, we run several applications in the performance simulator. For this purpose, we performed a two step evaluation including a single-core evaluation as well as a study of a multi-core processor, both depicted in Fig. 9. First, we analyzed the single-core architecture without L3-cache and replaced either the L1 and/or L2-cache memories by emerging MRAMs, either STT or SOT. For this setup we evaluated all benchmarks described in Table 2 and simulated their behavior for five billion instructions. The main results of this system-level study are presented in Fig. 10.

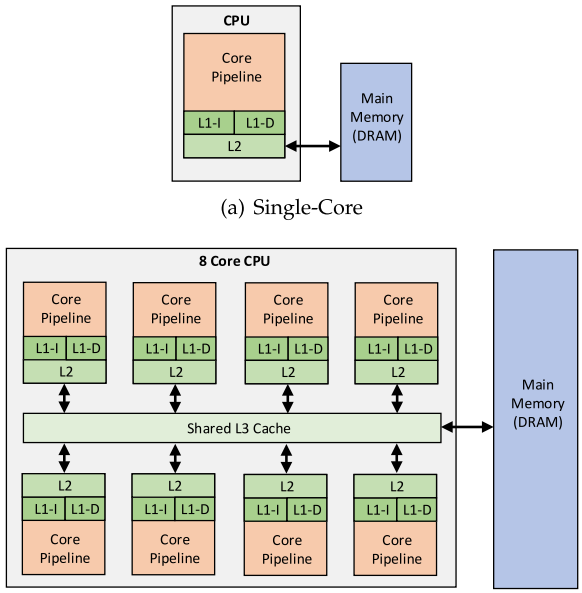


Fig. 9. Microprocessor configurations used for evaluation.

According to our results, replacing SRAM for the L2-cache with SOT-MRAM provides significant area savings, because the bit-cell size is significantly smaller. However, an SOT-MRAM based L1-cache is larger, due to the periphery circuit overhead that is dominating for small cache sizes. In terms of runtime performance SOT-MRAM is comparable to SRAM and offers even a small performance advantage, when it is employed for the L2-cache. Nevertheless, even for the L1-cache SOT-MRAM can be used without considerably affecting the performance. The biggest advantage of SOT-MRAM is its lower energy consumption. If it is employed for both cache-levels, the average energy consumed by the caches is reduced by ≈ 60 percent compared to an SRAM-only solution. Hence, in summary, SOT-

TABLE 2
Configuration Details for the Experiments

Processor	1-core or 8-cores @ 3 GHz, out-of-order, 4-issue
L1-Cache (Data & Instr.)	32 KByte, 2-way set associative, 64 B line size, 1 bank, MESI cache (SRAM: 0.7 ns, SOT: 1.0 ns/1.1 ns, STT: 1.0 ns/4.5 ns)
L2-Cache	512 KByte, 8-way set associative, 64 B line size, 1 bank, MESI cache (SRAM: 2.1 ns, SOT: 1.1 ns/1.4 ns, STT: 1.1 ns/4.7 ns)
Shared L3-Cache (for multi-core only)	16 MByte, 8-way set associative, 64 B line size, 1 bank, MESI cache (SRAM: 4.2 ns, SOT: 3.8 ns/2.8 ns, STT: 3.8 ns/6.2 ns)
Execution Units	2x ALU, 2x CALU, 2x FPU
MiBench applications [55]	BasicMath, BitCount, QSort, Dijkstra, Patricia, StringSearch, SHA, CRC, FFT
SPEC2000 applications	Bzip2, Equake, Gzip, MCF, VPR, Twolf
SPEC2006 applications	Hmmer, LBM, Sjeng
Multi-core workloads	Gzip+VPR+Bzip2+Twolf+Equake+Hmmer+LBM+Sjeng LBM+Equake+LBM+Equake+Bzip2+Equake+Sjeng+Hmmer Bzip2+Bzip2+Bzip2+Equake+VPR+Sjeng+Gzip+Twolf Twolf+LBM+LBM+Equake+Bzip2+Hmmer+Equake+Sjeng VPR+Twolf+Sjeng+Sjeng+Bzip2+Twolf+VPR+Sjeng

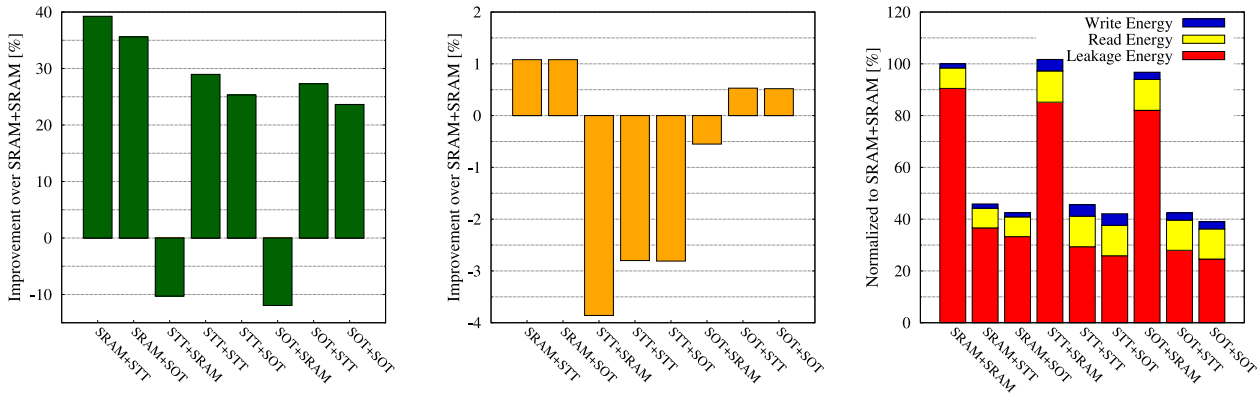


Fig. 10. Comparison of various cache configurations in terms of occupied area, average application runtime and average energy consumption (normalized to the standard configuration, i.e., SRAM for L1- and L2-cache).

MRAM caches offer a similar performance compared to SRAM caches, while the resulting energy is significantly lower and when used for higher level caches also the area is much smaller. Moreover, SOT-MRAM has often also an edge over STT-MRAM. On average, the energy consumption can be reduced by additional 5 percent compared to STT-MRAM and also the performance can benefit up to 3 percent. However, due to the additional bit-cell terminal, SOT-MRAM requires approximately 4 percent more area than STT-MRAM.

In addition to the single-core analysis, we also evaluated an 8-core processor with a shared L3-cache which is implemented either with SRAM or with MRAM. For this purpose we modified gem5 to support private L1- and L2-caches implemented with SRAM for each core as well as an shared L3-cache. The workloads are a mix of SPEC benchmarks (see Table 2) to fully utilize all available cores and also make use of the large L3-cache. Due to the increased simulation time, we simulated only eight billion instructions per workload mix. The corresponding results are depicted in Fig. 12.

As it can be seen, SOT-MRAM offers a considerable area (45 percent on average) and energy advantage (60 percent on average) over SRAM and also can slightly improve the overall system performance (by 1 percent on average). In this regard it is important to note that the tremendous energy savings are due to the fact that the major contributor to the overall energy consumption is the L3-cache, if it is implemented with SRAM, as shown in Fig. 11a. In contrast, if SOT-MRAM is used for the L3-cache, the SRAM-based L2-cache becomes the dominating part (see Fig. 11b). Therefore, for power-constrained systems, not only the L3-cache, but also the L2-cache should be implemented with SOT-MRAM to minimize the cache energy consumption. If performance is the major design constraint, some of the enormous savings offered by SOT-MRAM can be used to increase the size of the L3-cache, i.e. in our case to double the size from 16 MB to 32 MB. As a result, the performance improves by more than 4 percent on average compared to a 16 MB SRAM cache, while the area is still considerably smaller and the energy savings are still very impressive (around 58 percent). The reasons why the energy savings offered by a 16 MB and 32 MB SOT-MRAM cache are almost the same as twofold. First, the shorter runtime compensates partially the increase in leakage power, and consequently the energy consumption due to leakage only increases slightly, if the L3-cache size is increased from 16 MB to 32 MB. In fact, it is important to note again, that leakage power in SOT-MRAM is only consumed by the periphery circuitry. This is in contrast to SRAM, where also the bit-cells consume leakage power, and thus doubling the size also doubles the leakage power, if SRAM is employed. Second, the per-access energy plays only a minor role for such a last-level cache, and thus the increase in read and write energy is not noticeable.

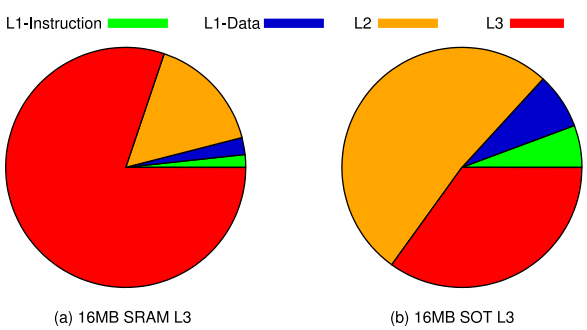
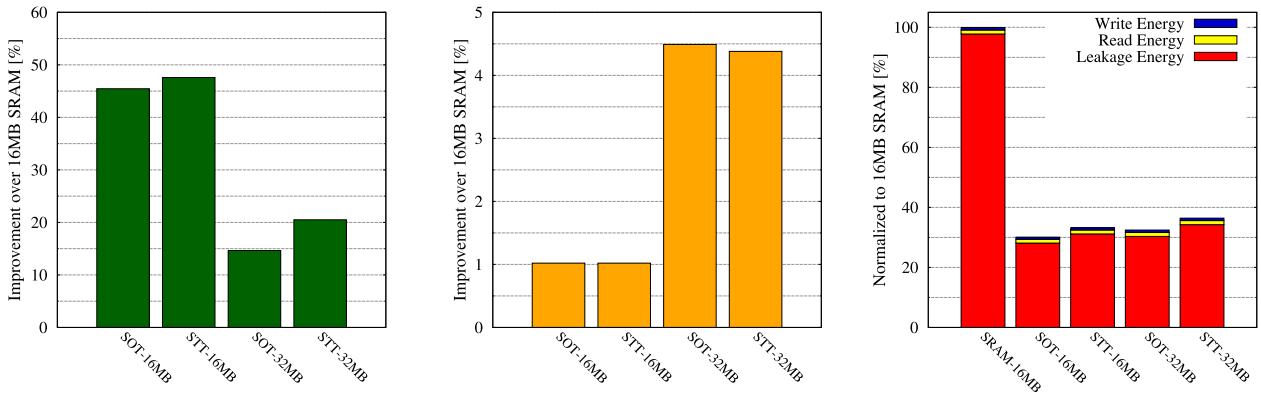


Fig. 11. Total cache energy breakdown.

Compared to a shared last-level cache implemented in STT-MRAM, SOT-MRAM shows similar benefits and disadvantages as before for a single-core. The area penalty is still around 5 percent due to the partitioning of the large L3-cache into multiple cache blocks. This cost is traded against an energy consumption, which is on average 5 percent lower than that of an STT-MRAM based L3-cache. Furthermore, the overall system performance is on the same level. In general, the farther away the cache is from a processor core, the closer becomes the performance of SOT-MRAM and STT-MRAM. The reason is that both have similar read latencies, and the write latencies are only important for the first and second cache level.

5 Non-Volatility for Power Saving Using the Power Gating Technique

Various design techniques exist to reduce the power consumption of complex SoCs like multicore processors: clock gating, power gating, multi-vdd design, etc. Power gating consists in cutting off the power supply of unused blocks of



(a) Area (b) Performance (c) Energy Consumption

Fig. 12. Comparison of various L3-cache configurations in terms of occupied area, performance (average application runtime) and average energy consumption (normalized to the standard configuration, i.e., SRAM for all cache levels).

a circuit to save leakage, which becomes an extremely important issue in modern ASICs, representing about half of the total power consumption in very advanced processes. This requires saving the content of the circuit in distant non-volatile or very low-leakage memories or registers. Introducing non-volatility in the circuit allows saving the context locally and so ease and improve the power gating technique. The combination of power gating and non-volatility could be used with various granularities to reduce power-consumption. It can also improve the circuit timing performances in so called normally-off and instant-on applications. In [56], a specifically designed microprocessor with non-volatility introduced by means of STT-based Non-Volatile Flip-Flops (NVFF) is presented. Since FFs are isolated memory elements, they don't suffer from high capacitive and resistive loads like the memory blocks, and it is expected that the backup and restore operations in the magnetic part can be done at a speed similar to the one of the magnetic device itself. In this case, the advantages of SOT in terms of speed and endurance could allow a much more frequent backup of the data in the magnetic devices, possibly at each clock cycle, easing or simplifying the backup process and allowing to still accelerate the power-off and power-on procedures. In order to validate the feasibility of integrating the SOT technology in complex systems and to measure its benefits, it should be interesting to consider an existing significantly complex circuit, to introduce non-volatile SOT-based elements and to fabricate it, using the most standard conception flow. The Leon3 microprocessor is a good vehicle as it is a 32-bit processor based on the SPARC V8 architecture which supports multiprocessing configurations (up to 16 CPU cores can be implemented). The LEON3 multiprocessor core is available in full source code under the GNU GPL license for evaluation, research and educational purposes [57]. In complex designs, replacing all the CMOS FFs by NV FFs implies an area overhead that might be too important compared to the overall benefits, especially since storing the contents of all the FF of the design might not be necessary. Moreover complex SoCs are composed of several blocks which are not active at the same time. In some cases, the activity of a dedicated block strongly depends on the application. For instance, in the Leon3 architecture, 32-bit divider and multiplier blocks are only enabled when a set of dedicated 32-bit divide and multiply instructions are employed. Making smartly chosen parts of a design non-volatile could reduce power consumption with no impact on the application. To summarize, to fabricate an hybrid CMOS/SOT circuit, it is necessary i) to develop a NVFF architecture ii) to integrate this FF as a Standard Cell in the digital design flow and iii) to adapt the flow to be able to select in which block the FF should be non-volatile.

An example of NVFF architectures proposed by Spintec is shown on Fig. 13. It includes two SOT-MTJs integrated in the master latch. Writing logic (Fig. 14) enables to save the output Q automatically when the writing of the MTJs is set (Wr signal). In order to ease the designer work, we propose to integrate the power gating transistors into the NVFF itself. So as to read correctly the MTJs the master latch is powered-on before the rest of the FF. The NVFF was

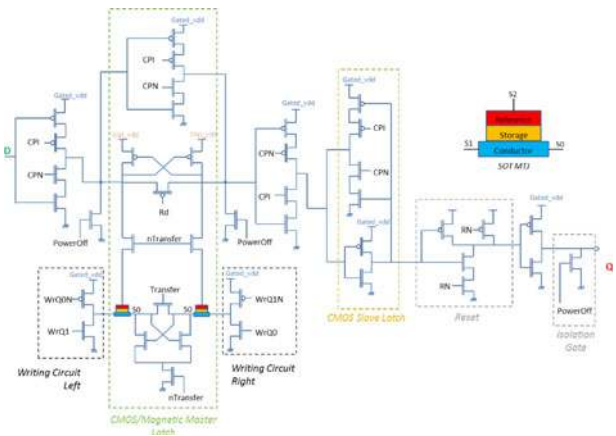


Fig. 13. Non-Volatile SOT-MTJs based Flip-Flop architecture.

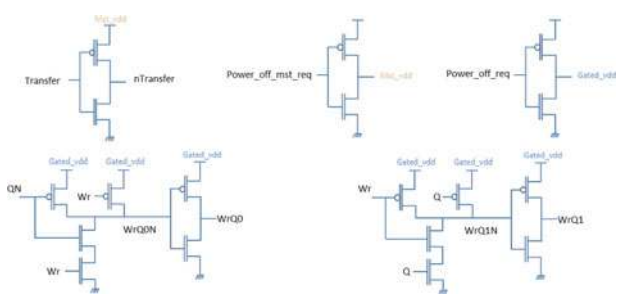


Fig. 14. Non-Volatile SOT-MTJs based Flip-Flop control logic.

compared in terms of surface, dynamic and static power consumption to a standard FF of the CMOS technology. The results are the following:

The surface is three times larger (due to the MTJs and in particular the writing circuit). The leakage power in twice the one of the standard FF, due to the additional transistors. And the dynamic power is almost the same, because the circuit to control the MTJs is not used in normal operation.

It is important to notice that we do not expect the NVFF by itself to have better characteristics than standard FF, because the purpose is not to replace all the FFs by NVFFs. The aim is to smartly choose which FF/registers has to be replaced by NVFF, depending on the application and the architecture of the circuit, for the gain in standby power consumption in power gating mode (with no power supply) to compensate the additional power consumption of the NVFFs. The future investigations will allow to choose the FFs to be replaced, depending on the architecture, application and performance of the magnetic part and even possibly to adapt the architecture of the processor to take advantage of the non-volatility.

Several important points have to be considered to allow using NVFF in a digital design flow. First of all, the NVFFs have to be carefully and accurately designed and characterized at electrical level. This is possible thanks to the compact model we developed [45]. CMOS process and magnetic post-process variations have to be considered, via parametric and Monte Carlo simulations. Second of all, all the features of these NVFFs, in terms of area, power consumption speed and so on, have to be given to the synthesis tool, in order to optimize this critical step in the flow. Third of all, the layout of these NVFFs has to be precisely mapped to the CMOS rules in order to be integrated in the automatic place and route flow. Last of all, the post-process has to be clearly defined to make it possible to extract its intrinsic parasitics for accurate post-layout simulations. From the design point of view, it is mandatory to consider the nominal operating clock frequency and the read and write durations of the MTJs. Indeed, if no additional logic is used to control the signals related to the MTJs, special attention must be given to the current pulses. In that case, the duration of read and write current pulses would be the duration of the period or at least half the period. The transistors must be tailored to obtain the minimum currents able to write and read the MTJs in the operation windows imposed by the global clock of the circuit. Concerning the design flow, it must allow the designer to choose where NV cells must be employed depending on its constraints in terms of application and requirements. The most critical phase is the synthesis. For the Leon3 case study, synthesizing each block targeting either CMOS or NV cells, and then assembling the synthesized blocks is not an easy task because of the high configurability of the processor. The complete design is thus synthesized once keeping the hierarchy, CMOS FFs are replaced by NVFFs and the modified netlist is finally recompiled and flattened. FFs replacement is made block by block by means of scripts, depending on the objectives.

If the objective is to make some blocks NV, the benefits depend on the NVFF power consumption itself, on the characteristics of the NV blocks and on the application. Indeed, power gating technique using NVFF has a cost: firstly, writing the content in the MTJs is energy consuming. For the technique to be efficient, the leakage saved should at least compensate this overhead. The ratio between the inactive and active periods of operation determines Secondly, making a FF NV can affect its performance, area or power consumption, even in standard operation. The NVFF power consumption must be as close as possible to the power consumption of the corresponding CMOS FF, on pain of degrading the full energy balance and so the technique efficiency. These considerations about MTJs read/write costs and energy savings has to be made block by block since the number of NVFF varies from a block to another. If the objective is to shut-down the complete processor by external means, the running application is stopped until the user commands the power-on. The number of FFs that must be NV (pipeline registers, program counter, etc) is then significant.

Future work will be carried-out on the power-off of the complete processor. This will enable us to fully detail the benefits and the costs of the hybrid CMOS/SOT approach for a complex SoC. So far, results show the importance of the non-volatility strategy as a function of the application.

6 Conclusion

In this paper, we have presented a new MRAM technology based on Spin Orbit Torque writing, whose advantages in terms of speed and endurance allow adressing very high performance applications. Indeed, ultra-fast switching of 186 ps was demonstrated as well as writing current densities of $3-6 \times 10^{10}\ A/m^2$. This allows expecting writing currents of 3-6 μA for advanced devices. A full design framework has been developed to evaluate the benefits that can be expected from using such devices in the memory hierarchy of complex systems. The results show that SOT-MRAM can be advantageoulsy introduced at any level of the cache of processors, including level 1, offering a strong reduction of the power consumption (up to 60 percent compared to SRAM only solutions) without affecting the performance. Compared to STT solutions, the gain in around 5 percent in terms of power consumption, with a slight area penalty, but great benefit in terms of endurance (required for high speed operating memories). Work is ongoing to evaluate the use of SOT-MRAM in the logic itself, to further reduce the power consumption and pave the way towards normally-off computing. We are confident that this technology can strongly contribute to the future of microelectronics.

Acknowledgments

This work has been partially funded by the European Commission under the spot project (grant agreement n318144) in the framework of the Seventh Framework Program.

References

[1] J. Hutchby and M. Garner, "Assessment of the potential & maturity of selected emerging research memory technologies," in *Proc. Workshop ERD/ERM Work. Group Meet.*, Apr. 2010, pp. 6–7.

[2] M. Kryder and C. S. Kim. (2009, Oct.). After Hard DrivesWhat Comes Next?" *IEEE Trans. Magn.* [Online]. *45(10)*, p. 3406. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5257331

[3] W. Guo, G. Prenat, V. Javerliac, M. Baraji, N. de Mestier, C. Baraduc, and B. Diny. (2010, May). SPICE modelling of magnetic tunnel junctions written by spin-transfer torque, *J. Phys. D: Appl. Phys.* [Online]. *43(21)*, p. 215001. Available: http://stacks.iop.org/0022-3727/43/i=21/a=215001

[4] D. Ralph and M. Stiles, "Spin transfer torques," *J. Magnetism Magn. Mater.*, vol. 320, no. 7, pp. 1190–1216, Apr. 2008.

[5] W. Zhao, Y. Zhang, T. Devolder, J. Klein, D. Ravelosona, C. Chappert, and P. Mazoyer, "Failure and reliability analysis of STT-MRAM," *Microelectron. Rel.* [Online]. *52(9–10)*, pp. 1848–1852. Available: http://linkinghub.elsevier.com/retrieve/pii/S0026271412002326

[6] J. Li, C. Augustine, S. Salahuddin, and K. Roy, "Modeling of failure probability and statistical design of spin-torque transfer magnetic random access memory (STT MRAM) array for yield enhancement," in *Proc. 45th Annu. Des. Autom. Conf.*, 2008, pp. 278–283.

[7] W. Zhao, T. Devolder, Y. Lakys, J.-O. Klein, C. Chappert, and P. Mazoyer, "Design considerations and strategies for high-reliable STT-MRAM," *Microelectron. Rel.*, vol. 51, no. 9, pp. 1454–1458, 2011.

[8] K. Chun, H. Zhao, J. Harms, T. Kim, J. Wang, and C. Kim, "A scaling Roadmap and performance evaluation of In-plane and perpendicular MTJ based STT-MRAMs for high-density cache memory," *IEEE J. Solid-State Circuits*, vol. 48, no. 2, pp. 598–610, Feb. 2013.

[9] M. Gajek, J. Nowak, J. Sun, P. Trouilloud, E. OSullivan, D. Abraham, M. Gaidis, G. Hu, S. Brown, Y. Zhu, et al., "Spin torque switching of 20 nm magnetic tunnel junctions with perpendicular anisotropy," *Appl. Phys. Lett.*, vol. 100, no. 13, p. 132408, 2012.

[10] M. Marins de Castro, R. C. Sousa, S. Bandiera, C. Ducruet, A. Chavent, S. Auffret, C. Papusoi, I. L. Prejbeanu, C. Portemont, L. Vila, U. Ebels, B. Rodmacq, and B. Dieny, "Precessional spin-transfer switching in a magnetic tunnel junction with a synthetic antiferromagnetic perpendicular polarizer," *J. Appl. Phys.* [Online]. *111(7)*, p. 07C912. Available: http://link.aip.org/link/JAPIAU/v111/i7/p07C912/s1&Agg=doi

[11] H. Zhao, B. Glass, P. K. Amiri, A. Lyle, Y. Zhang, Y.-J. Chen, G. Rowlands, P. Upadhyaya, Z. Zeng, J. a. Katine, J. Langer, K. Galatsis, H. Jiang, K. L. Wang, I. N. Krivorotov, and J.-P. Wang. (2012, Jan.). Sub-200 PS spin transfer torque switching in in-plane magnetic tunnel junctions with interface perpendicular anisotropy, *J. Phys. D: Appl. Phys.* [Online]. *45(2)*, p. 025001. Available: http://stacks.iop.org/0022-3727/45/i=2/a=025001?key=crossref.000fc7de61505e980bdbd0f445085454

[12] G. Gaudin, I. M. Miron, P. Gambardella, and A. Schuhl, "Magnetic memory element," ICN, CNRS, and ICREA, US Patent application 12/899,072 (06.10.2010). WO 2012/014131 (02/12/2012), vol. 14131, p. 2012, 2012.

[13] A. D. Kent, B. Özyilmaz, and E. del Barco. (2004). Spin-transfer-induced precessional magnetization reversal, *Appl. Phys. Lett.* [Online]. *84(19)*, p. 3897. Available: http://scitation.aip.org/content/aip/journal/apl/84/19/10.1063/1.1739271

[14] I. M. Miron, G. Gaudin, S. Auffret, B. Rodmacq, A. Schuhl, S. Pizzini, J. Vogel, and P. Gambardella. (2010, Mar.). Current-driven spin torque induced by the Rashba effect in a ferromagnetic metal layer, *Nature Mater.*, Mar. 2010. [Online]. *9(3)*, pp. 230–234. Available: http://www.ncbi.nlm.nih.gov/pubmed/20062047

[15] I. M. Miron, K. Garello, G. Gaudin, P.-J. Zermatten, M. V. Costache, S. Auffret, S. Bandiera, B. Rodmacq, A. Schuhl, and P. Gambardella. (2011, Aug.). Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection, *Nature* [Online]. *476 (7359)*, pp. 189–193. Available: http://www.ncbi.nlm.nih.gov/pubmed/21804568

[16] L. Liu, C. Pai, Y. Li, H. Tseng, D. Ralph, and R. Buhrman, "Spin-torque switching with the giant spin hall effect of tantalum," *Science*, vol. 336, no. 6081, pp. 555–558, 2012.

[17] P. Gambardella and I. M. Miron, "Current-induced spin-orbit torques. (2011, Aug.). *Philosoph. Trans. Series A, Math., Phys., Eng. Sci.* [Online]. *369(1948)*, pp. 3175–3197. Available: http://www.ncbi.nlm.nih.gov/pubmed/21727120

[18] S. Ganichev and W. Prettl, "Spin photocurrents in quantum wells," *J. Phys.: Condens. Matter*, vol. 15, pp. R935–R983, 2003.

[19] A. Manchon and S. Zhang, "Theory of nonequilibrium intrinsic spin torque in a single nanomagnet," *Phys. Rev. B*, vol. 78, no. 21, p. 212405, 2008.

[20] A. Hoffmann, "Spin hall effects in metals," *IEEE Trans. Magn.*, vol. 49, no. 10, pp. 5172–5193, Oct. 2013.

[21] T. Jungwirth, J. Wunderlich, and K. Olejník. (2012, May.). Spin hall effect devices, *Nature Mater.* [Online]. *11(5)*, pp. 382–390. Available: http://www.ncbi.nlm.nih.gov/pubmed/22522638

[22] J. Sinova, S. O. Valenzuela, J. Wunderlich, C. H. Back, and T. Jungwirth. (2014, Nov.). Spin hall effect, *arxiv 1411.3249v1* [Online]. p. 48. Available: http://arxiv.org/abs/1411.3249

[23] F. Freimuth, S. Blügel, and Y. Mokrousov, "Spin-orbit torques in Co/Pt(111) and Mn/W(001) magnetic bilayers from first principles," *Phys. Rev. B*, vol. 90, no. 17, p. 174423, Nov. 2014.

[24] P. M. Haney, H.-W. Lee, K.-J. Lee, A. Manchon, and M. D. Stiles, "Current induced torques and interfacial spin-orbit coupling: Semiclassical modeling," *Phys. Rev. B*, vol. 87, no. 17, p. 174411, May 2013.

[25] C. O. Pauyac, X. Wang, M. Chshiev, and A. Manchon. (2013). Angular dependence and symmetry of Rashba spin torque in ferromagnetic heterostructures, *Appl. Phys. Lett.* [Online]. *102*, p. 242403. Available: http://scitation.aip.org/content/aip/journal/apl/102/25/10.1063/1.4812663

[26] D. A. Pesin and A. H. MacDonald. (2012, Jul.). Quantum kinetic theory of Current-induced torques in Rashba ferromagnets," *Phys. Rev. B* [Online]. *86(1)*, p. 014416. Available: http://link.aps.org/doi/10.1103/PhysRevB.86.014416

[27] X. Wang and A. Manchon, "Diffusive spin dynamics in ferromagnetic thin films with a rashba interaction," *Phys. Rev. Lett.*, vol. 108, no. 11, p. 117201, Mar. 2012.

[28] K. Garello, I. M. Miron, C. O. Avci, F. Freimuth, Y. Mokrousov, S. Blügel, S. Auffret, O. Boulle, G. Gaudin, and P. Gambardella, "Symmetry and magnitude of spin-orbit torques in ferromagnetic heterostructures," *Nature Nanotechnol.*, vol. 8, no. 8, pp. 587–593, 2013.

[29] K. Garello, C. O. Avci, I. M. Miron, M. Baumgartner, A. Ghosh, S. Auffret, O. Boulle, G. Gaudin, and P. Gambardella. (2014, Nov.). Ultrafast magnetization switching by spin-orbit torques, *Appl. Phys. Lett.* [Online]. *105(21)*,p. 212402. Available: http://scitation.aip.org/content/aip/journal/apl/105/21/10.1063/1.4902443

[30] E. Martinez, L. Torres, N. Perez, M. A. Hernandez, V. Raposo, and S. Moretti. (2015). Universal chiral-triggered magnetization switching in confined nanodots, *Sci. Rep.* [Online]. *5*, p. 10156. Available: http://www.ncbi.nlm.nih.gov/pubmed/26062075

[31] N. Mikuszeit, O. Boulle, I. M. Miron, K. Garello, P. Gambardella, G. Gaudin, and L. D. Buda-Prejbeanu, "Spin-orbit torque driven chiral magnetization reversal in ultrathin nanostructures," *Phys. Rev. B*, vol. 92, no. 14, p. 144424, 2015.

[32] M. Hayashi, J. Kim, M. Yamanouchi, and H. Ohno, "Quantitative characterization of the spin-orbit torque using harmonic Hall voltage measurements," *Phys. Rev. B*, vol. 89, no. 14, p. 144425, Apr. 2014.

[33] C. Zhang, M. Yamanouchi, H. Sato, S. Fukami, S. Ikeda, F. Matsukura, and H. Ohno. (2014, May). Magnetization reversal induced by in-plane current in Ta/CoFeB/MgO structures with perpendicular magnetic easy axis, *J. Appl. Phys.* [Online]. *115(17)*,p. 17C714. Available: http://scitation.aip.org/content/aip/journal/jap/115/17/10.1063/1.4863260

[34] C.-F. Pai, L. Liu, Y. Li, H. W. Tseng, D. C. Ralph, and R. A. Buhrman. (2012). Spin transfer torque devices utilizing the giant spin Hall effect of tungsten, *Appl. Phys. Lett.* [Online]. *101(12)*, p. 122404. Available: http://scitation.aip.org/content/aip/journal/apl/101/12/10.1063/1.4753947

[35] C. Onur Avci, K. Garello, I. Mihai Miron, G. Gaudin, S. Auffret, O. Boulle, and P. Gambardella. (2012). Magnetization switching of an MgO/Co/Pt layer by in-plane current injection, *Appl. Phys. Lett.* [Online]. *100(21)*, p. 212404. Available: http://scitation.aip.org/content/aip/journal/apl/100/21/10.1063/1.4719677

[36] X. Fan, J. Wu, Y. Chen, M. J. Jerry, H. Zhang, and J. Q. Xiao. (2013, Jan.). Observation of the nonlocal spin-orbital effective field, *Nature Commun.* [Online]. *4*,p. 1799. Available: http://www.ncbi.nlm.nih.gov/pubmed/23653211

[37] M. Jamali, K. Narayanapillai, X. Qiu, L. M. Loong, A. Manchon, and H. Yang, "Spin-orbit torques in Co/Pd Multilayer Nanowires," *Phys. Rev. Lett.*, vol. 111, no. 24, p. 246602, Dec. 2013.

[38] P. P. J. Haazen, E. Murè, J. H. Franken, R. Lavrijsen, H. J. M. Swagten, and B. Koopmans. (2013, Apr.). Domain wall depinning governed by the spin Hall effect," *Nature Mater.* [Online]. *12(4)*, pp. 299–303. Available: http://www.ncbi.nlm.nih.gov/pubmed/23377291

[39] M. Cubukcu, O. Boulle, M. Drouard, K. Garello, C. Onur Avci, I. Mihai Miron, J. Langer, B. Ocker, P. Gambardella, and G. Gaudin. (2014, Jan.). Spin-orbit torque magnetization switching of a three-terminal perpendicular magnetic tunnel junction, *Appl. Phys. Lett.* [Online]. *104(4)*, p. 42406. Available: http://scitation.aip.org/content/aip/journal/apl/104/4/10.1063/1.4863407

[40] [Online]. Available: http://www.spot-research.eu

[41] G. Jan, Y.-j. Wang, T. Moriyama, Y.-j. Lee, M. Lin, T. Zhong, R.-y. Tong, T. Torng, and P.-k. Wang, "High spin torque efficiency of magnetic tunnel junctions with MgO/CoFeB/MgO free layer," *Appl. Phys. Exp.*, vol. 5, p. 93008, 2012.
[42] K. Yamane, Y. Higo, H. Uchida, Y. Nanba, S. Sasaki, H. Ohmori, K. Bessho, and M. Hosomi, "Spin torque switching of perpendicularly magnetized CoFeB-based tunnel junctions with high thermal tolerance," *IEEE Trans. Magn.*, vol. 49, no. 7, pp. 4335–4338, Jul. 2013.
[43] G. Di Pendina, G. Prenat, B. Dieny, and K. Torki, "A hybrid magnetic/complementary metal oxide semiconductor process design kit for the design of Low-power Non-volatile logic circuits," *J. Appl. Phys.*, vol. 111, no. 7, p. 07E350, 2012.
[44] K. Jabeur, F. Bernard-Granger, G. Di Pendina, G. Prenat, and B. Dieny, "Comparison of Verilog-a compact modelling strategies for spintronic devices," *Electron. Lett.*, vol. 50, no. 19, pp. 1353–1355, 2014.
[45] K. Jabeur, G. Di Pendina, G. Prenat, L. D. Buda-Prejbeanu, and B. Dieny, "Compact modeling of a magnetic tunnel junction based on spin orbit torque," *IEEE Trans. Magn.*, vol. 50, no. 7, pp. 1–8, Jul. 2014.
[46] L. Landau and E. Lifshitz, "On the theory of the dispersion of magnetic permeability in ferromagnetic bodies," *Phys. Zeitsch. Der Sow.*, vol. 8, no. 153, pp. 153–169, 1935.
[47] M. Julliere, "Tunneling between ferromagnetic films," *Phys. Lett. A*, vol. 54, no. 3, pp. 225–226, 1975.
[48] J. G. Simmons, "Generalized formula for the electric tunnel effect between similar electrodes separated by a thin insulating film," *J. Appl. Phys.*, vol. 34, no. 6, pp. 1793–1803, 1963.
[49] W. Brinkman, R. Dynes, and J. Rowell, "Tunneling conductance of asymmetrical barriers," *J. Appl. Phys.*, vol. 41, no. 5, pp. 1915–1921, 1970.
[50] K. Jabeur, G. Di Pendina, and G. Prenat, "Ultra-energy-efficient CMOS/magnetic nonvolatile flip-flop based on spin-orbit torque device," *Electron. Lett.*, vol. 50, no. 8, pp. 585–587, 2014.
[51] K. Jabeur, G. Di Pendina, F. Bernard-Granger, and G. Prenat, "Spin orbit torque non-volatile flip-flop for high speed and low energy applications," *IEEE Electron Device Lett.*, vol. 35, no. 3, pp. 408–410, Mar. 2014.
[52] X. Dong, C. Xu, Y. Xie, and N. Jouppi, "NVSIM: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Trans. Comput.-Aided Des. Integrated Circuits Syst.*, vol. 31, no. 7, pp. 994–1007, Jul. 2012.
[53] F. Oboril, R. Bishnoi, M. Ebrahimi, and M. Tahoori, "Evaluation of hybrid memory technologies using SOT-MRAM for on-chip cache hierarchy," *IEEE Trans. Comput.-Aided Des. Integrated Circuits Syst.*, vol. 34, no. 3, pp. 367–380, Mar. 2015.
[54] N. Binkert, R. Dreslinski, L. Hsu, K. Lim, A. Saidi, and S. Reinhardt, "The M5 simulator: Modeling networked systems," *IEEE Micro.*, vol. 26, no. 4, pp. 52–60, 2006.
[55] M. Guthaus, J. Ringenberg, D. Ernst, T. Austin, T. Mudge, and R. Brown, "MiBench: A free, commercially representative embedded benchmark suite," in *Proc. Workshop Workload Characterization*, Dec. 2001, pp. 3–14.
[56] H. Koike, T. Ohsawa, S. Ikeda, T. Hanyu, H. Ohno, T. Endoh, N. Sakimura, R. Nebashi, Y. Tsuji, A. Morioka, et al., "A Power-gated mpu with 3-microsecond entry/exit delay using Mtj-based nonvolatile flip-flop," in *Proc. Solid-State Circuits Conference (A-SSCC), 2013 IEEE Asian*. IEEE, 2013, pp. 317–320.
[57] [Online]. Available: http://www.gaisler.com/doc/leon3_product_sheet.pdf