Spin Orbit Torque memory for non-volatile microprocessor caches

F.Oboril[1], R.Bishnoi[1], M.Ebrahimi[1], and M.Tahoori[1] , G. Di Pendina[2], K.Jabeur[2], G.Prenat[2]
[1)]*Karlsruhe Institute of Technology, Germany*
[2)]*Univ. Grenoble Alpes, CNRS, CEA, INAC-SPINTEC, F-38000 Grenoble, France*

Abstract: *Magnetic spin-based memory technologies are a promising solution to overcome the incoming limits of microelectronics. Nevertheless, the long write latency and high write energy of these memory technologies compared to SRAM make it difficult to use these for fast microprocessor memories, such as L1-Caches. However, the recent advent of the Spin Orbit Torque (SOT) technology changed the story: indeed, it potentially offers a writing speed comparable to SRAM with a much better density as SRAM and an infinite endurance, paving the way to a new paradigm in processor architectures, with introduction of non-volatility in all the levels of the memory hierarchy towards full normally-off and instant-on processors. This paper presents a full design flow, from device to system, allowing to evaluate the potential of SOT for microprocessor cache memories and very encouraging simulation results using this framework.*

I. Introduction

For decades, microelectronic trends have been following Moore's law, stating that the performance and density of integrated circuits would double almost every 2 years. But today, the extreme miniaturization of devices leads to physical limits, like power consumption and heating issues, making this trend reach a plateau. Several solutions are investigated to continue the miniaturization of logic circuits. While the "more Moore" approach involves technology innovation to allow still scaling the CMOS technology, the "more than Moore" approach aims at using new devices beside or in replacement of CMOS transistors to sustain the miniaturization of circuits. Among the new technologies which are investigated today, the use of Non-Volatile (NV) devices appears as a promising solution to contribute to reduce the power consumption of logic circuits. Compared to other NV technologies, magnetic random access memory (MRAM) combines a set of advantages. It is intrinsically NV, immune to radiations, with high writing and reading speed, low-power consumption and a quasi-infinite endurance. The constitutive device of MRAM is the Magnetic Tunnel Junction (MTJ), a NV magnetic device whose electrical resistance depends on the magnetic state [1]. Several generations of MRAM have been proposed so far, the most relevant being based on the Spin Transfer Torque (STT) technology [2], which relies on the use of a spin-polarized current applied through the MTJ to switch the magnetic state. This technology allows a writing speed of a few nanoseconds, a density close to that of DRAM, combined with an intrinsic non-volatility. However, it suffers from two main weaknesses, due to the fact that it is a two-terminal device which uses the same path for reading and writing. First, there is a risk of accidental writing when reading, if the reading current is not perfectly controlled or if the process variations are too important. Second, the application of a relatively high current for writing can affect the endurance of the device and leads to a high energy consumption during write accesses, especially in the case of fast switching that need high writing current. Recently, a new technology has appeared, based on the SOT (Spin Orbit Torque) effect [3]. It allows conceiving new devices with three terminals, with separate reading and writing paths, solving the issues mentioned above. Moreover, the physical effect governing the switching and the use of a dedicated writing path allows a very fast writing speed, well below 1ns and so comparable to those of SRAM. So far, STT-MRAM was seen as a good candidate to replace some parts of the memory hierarchy of processors [4], down to cache level two (due to its limited writing speed). Thanks to its very high writing speed and unlimited endurance, the SOT technology can even allow replacing first level caches of processors without impairing the performance. In this paper, we present a full design flow, from device to system, which enables designers to evaluate the potential of SOT, and we highlight the potential of SOT-MRAM for caches in a system-level study.

The reminder of this paper is organized as follows. In Section II, the circuit level design tools and the design of an SOT-MRAM array are presented. In Section III, a comprehensive evaluation of SOT for processor caches is performed. Section IV concludes this paper.

II. DESIGN OF SOT-MRAM

The MTJ is the basic element of MRAMs. It is composed of two Ferro-Magnetic (FM) layers separated by an insulator. The reference layer magnetization is pinned. The storage layer magnetization is programmable (up or down). For

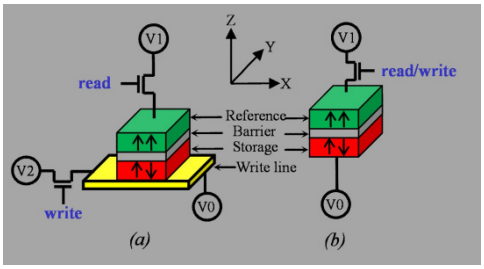

Fig. 1: 1-bit cell structure (a) 3-terminal SOT device with two independent paths for write and read operations (b) 2-terminal STT device with common read and write path

SOT-MTJ devices the storage layer is on the top of a conductor (metal electrode write line). An in-plane current injection through the write line (x-axis) induces the perpendicular switching of the storage layer according to the current direction. Generally for MTJs, when the reference and the storage layers have a parallel magnetization state, the resistance of the MTJ is Rmin (logic 0). When the reference and the storage layers have an anti-parallel magnetization state, the resistance of the MTJ is Rmax (logic 1). Fig. 1(a) shows the SOT-MTJ with a 3-terminal architecture alleviating the stress on the barrier by separating the read path from the write path. Fig. 1(b) shows the 2-terminal architecture of the STT-MTJ with a common read and write path.

In order to explore the potential of the SOT-based MTJs in real IC applications, a hybrid CMOS/MTJ MRAM is designed. To do so, it is necessary to integrate the SOT device in standard microelectronics design suites. This is performed by means of the so called process design kit (PDK) for the hybrid CMOS/magnetic technology, as presented in [5]. The magnetic PDK contains a compact model of the MTJ for electrical simulations, technology files for layout and physical verifications, and standard cells for the design of complex logic circuits. Moreover, this PDK is compatible with standard design suites. We developed the physical compact model by using the simulator-compatible Verilog-A language. While the model is described in details with an exhaustive study of different parameters variations in [6], Fig. 2 shows the behavior of the SOT-MTJ model which validates the theoretical expectations. If a negative pulse is applied, the magnetization of the storage layer (Mz) is reversed downward (Mz=-1) while a positive pulse switches the magnetization upward (Mz=1). Fig. 2(c) shows the symbol view of the SOT device which will be connected with other CMOS devices to build a full schematic of the hybrid Magnetic/CMOS circuit. Simulations are carried out to check the correct functionality of the designed circuit. Fig. 2(d) shows the layout view of the SOT device which corresponds to the successive steps of the manufacturing process. This layout is checked by several tools such as verification of the design rules of the manufacturer (Design Rule Check) and the layout is compared to the simulated schematic (Layout Versus Schematic). All these verifications rely on technology files which are part of the PDK for a given design suite.

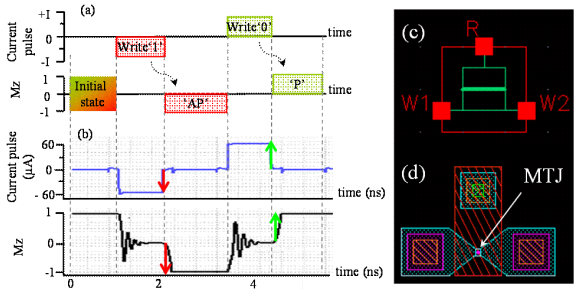

Fig. 2: SOT-MTJ device (a) Theoretical behavior (b) model behavior (c) Symbol view (d) Layout view

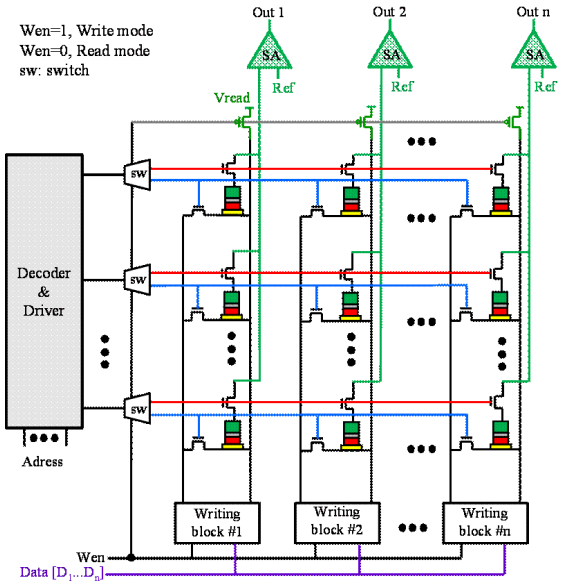

Fig. 3: Architecture of SOT-MRAM memory arrays

Fig. 3 shows the architecture of a SOT-MRAM memory array. It has a lot in common with the SRAM memory architecture. The peripheral circuits almost remain the same (this is similar in bit-cell-based memory arrays in general). For instance, a word-line decoder is required for the activation of the word-line indicated by the memory address. However, in our architecture, we designed a write block and a read block (SA, Sense Amplifier which can be preceded by pre-amplification blocks) for each column. Also, a switch is used for each word line to have access to bit-cells according to a write mode or a read mode depending on the value of the Wen signal as shown in Fig. 3. Once a word-line is chosen by the decoder, it is possible to write or read the n bits of the chosen word-line at the same time. It is conceivable to use only one write block and one read block for the whole architecture, then add another column decoder with multiplexer circuits. However, such a design choice adds more transistors in the write and read paths leading to more complex considerations at the design level (voltage drop, transistor sizing, reliability) but increases the reliability. Besides, the choice of the SOT-MRAM architecture in Fig. 3 eases the testing phase once the circuit is fabricated and enables the test of different circuit blocks independently.

In order to explain the operation of SOT-MRAM, we present in Fig. 4 one bit cell structure connected with the read and write blocks. Since the SOT-MTJ is a 3-terminal device, we can consider the whole SOT-based bit cell as a 5-terminal structure (t1..t5), as shown in Fig. 4. While the source line is common, it is clear that the write and read paths are completely separated thanks to the Read select and Write select transistors. Thus the reliability is noteworthy increased and the stress on the MTJ barrier is widely decreased.

For a write operation, the write enable signal (Wen) is activated. Indeed, the write operation in SOT-MRAM is bidirectional, i.e. the data stored in the bit-cell depends on the direction of the current which in-turn is

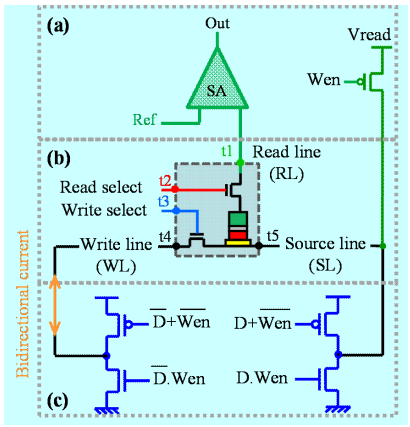

Fig. 4: SOT bit cell structure with write and read blocks (a) Read circuit (b) 1 Bit cell structure (c) Write circuit

determined by the input data value D. As a result, the write circuitry can be designed in such a way that the high resistance state of the MTJ cell represents either a logic 1 or a logic 0. Based on the SOT-MTJ model shown in Fig. 2, it is assumed that the anti-parallel state (high resistance) represents a logical value of 1. To perform a read operation, the Wen signal is low which means that transistors of the write circuit are inactivated as well as the Write select transistor. The read line is connected to a current sense amplifier while the source line is connected to the read-voltage Vread. Thus, the current sensed on the read line is compared with a reference value to determine the value stored in the cell.

III. Evaluation for Caches

Based on the device-level evaluation platform described in the previous Section we built a multi-level analysis framework depicted in Fig. 5 to be able to explore SOT-MRAM for large memory arrays and its feasibility for microprocessor caches. The circuit-level evaluation is performed with NVSim [7] tool, which contains circuit level performance, energy, and area models for various memory technologies such as SRAM, NAND-Flash and MRAM. Moreover, we modified NVSim to support SOT-MRAM and also the asymmetric write behavior (set vs. reset) of STT-MRAM. Using the device-level parameters as input for NVSim, we are able to obtain information

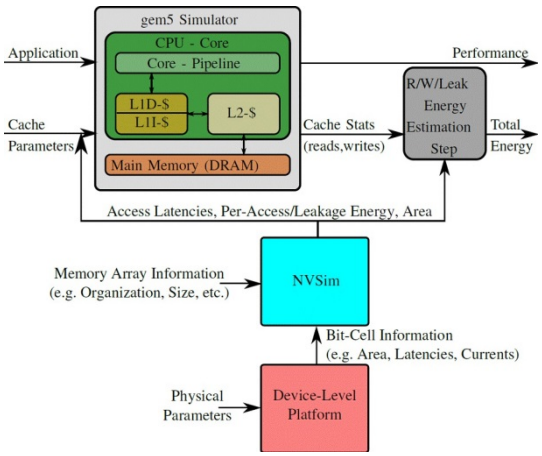

Fig. 5: Cross-Layer Evaluation Platform

Table 1: Experimental Setup for the System-Level Analysis

Processor	Single-core or 8-core @ 3 GHz
L1-Cache (Data & Instr.)	32KB, 2-way, 64B line, (SRAM: 0.7/0.7ns (read/write), SOT: 1.0/1.1ns, STT: 1.0/4.5ns)
L2-Cache	512KB, 8-way, 64B line, (SRAM: 2.1/2.1ns, SOT: 1.1/1.4ns, STT: 1.1/4.7ns)
Shared L3-Cache (multicore only)	16MB, 8-way, 64B line, (SRAM: 4.2/4.2ns, SOT: 3.8/2.8ns, STT: 3.8/6.2ns)
Benchmarks	MiBench+ SPEC

about an entire memory array such as access latencies, per- access energy, leakage power and area. This data is then used by the next tool in our framework to evaluate the implications of different memory technologies at system-level, if these memory technologies are used for processor caches at different levels. Therefore, we employ the cycle-accurate performance-simulator gem5 [8] which supports various memory configurations and allows to configure all relevant cache parameters such as capacity, associativity, latency, block size and policy. In addition, we modified gem5 to also support the asymmetric read and write behavior of STT-MRAM. Using this cross-layer analysis framework with the microprocessor setup detailed in Table 1 we conducted various experiments to compare SOT with SRAM and perpendicular STT. As the Table shows SRAM is faster for small caches, while SOT-MRAM is superior for larger caches. This is due to the fact that in SOT-MRAM the delay of the periphery circuits (e.g. sense amplifier, write circuitry) is dominant, while for SRAM it is the routing delay of the bit-cell array. Thus, the delay of an SRAM array increases considerably with increasing memory size, while for SOT-MRAM the delay remains on a very similar level. Moreover, SOT does not suffer from long write access latencies as it is the case for STT. Hence, for large memories SOT is very promising to be even faster than SRAM. To evaluate this statement 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 single- as well as multi-core systems. This consists of replacing either L1, L2 and/or L3-cache memories by emerging MRAMs, either STT or SOT.

The results for the single-core analysis with focus on the L1- and L2-cache are presented in Fig. 6. According to our results, replacing SRAM for the L2-cache with SOT provides significant area savings, because the bit-cell size is significantly smaller. However, an SOT based L1-cache is larger, due to the periphery circuit overhead that is dominating for small cache sizes. In terms of runtime performance SOT 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 can be used without considerably affecting the performance. The biggest advantage of SOT is its lower energy consumption. If it is employed for both cache-levels, the average energy consumed by the caches is

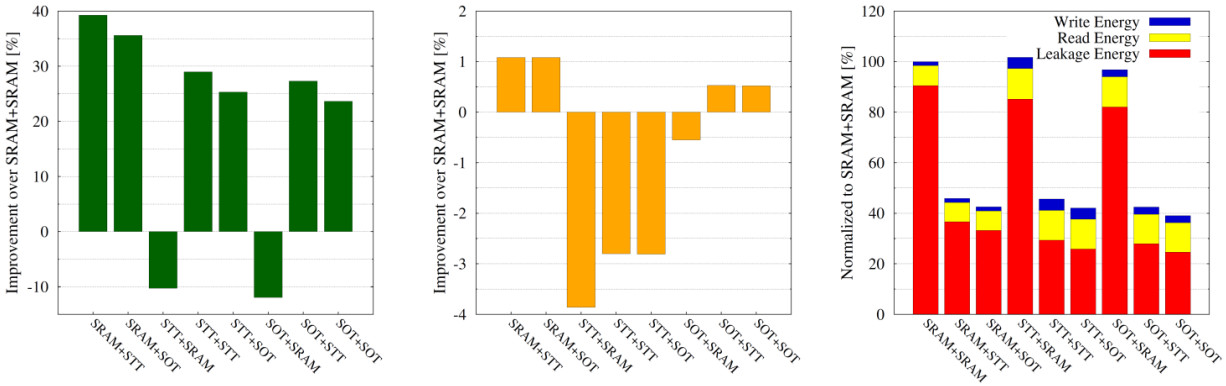

Fig. 6: Comparison of various cache configurations in terms of occupied area, average application runtime and average energy consumption (normalized to the standard configuration, i.e. SRAM for L1- and L2-cache)

reduced by ≈60% compared to an SRAM-only solution. Hence, in summary, SOT 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 has also an edge over STT, which means that it can be also applied for fast and small caches, for which STT is not an option. On average, the energy consumption can be reduced by additional 5% compared to STT and also the performance can benefit up to 3%. However, due to the additional bit-cell terminal, SOT requires approximately 4% more area than STT.

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 L1and L2-caches implemented with SRAM for each core as well as an shared L3-cache. The results are depicted in Fig. 7. As it can be seen, the advantage of SOT over SRAM is considerable, while the differences between STT and SOT are on a similar level as for the L1- and L2-cache. Still SOT offers a slightly better performance and energy consumption, at the expense of an slightly increased area. Nevertheless, a large L3-cache implemented in SOT consumes approximately just half of the area of an SRAM-based solution, which means that the cache size can be doubled if SOT is used. As a result, the performance improves considerably (by more than 4%), while still 40% less energy is consumed compared to an SRAM solution.

IV. Conclusions

A full design flow has been developed from device to system level, to investigate the use of SOT-MRAM in the memory hierarchy of processors. The study shows very encouraging results. It appears that the SOT technology can compete with SRAM in terms of performance and offers a much higher density, and could become a universal memory for some applications. This would allow a new paradigm contributing to push forward the limits of microelectronics in the way of "more than Moore".

V. Acknowledgement

The work and results reported were obtained on the framework of the spOt project (grant agreement n318144) funded by the European Commission under the Seventh Framework Programme.

VI. References

[1] S. Ikeda et al., "Tunnel magnetoresistance of 604suppression of ta diffusion in cofeb/mgo/cofeb pseudo-spin-valves annealed at high temperature," Appl. Phys. Lett. 93, 082508.
[2] H. Zhao et al., "A scaling roadmap and performance evaluation of in-plane and perpendicular mtj based stt-mrams for high-density cache memory," IEEE JSSC, Feb. 2013.
[3] P. Gambardella et al., "Current-induced spinorbit torques," Phil. Trans. R. Soc. A 369, 3 10.1098/rsta.2010.0336 (2011).
[4] International Technology Roadmap for Semicoductors 2010.
[5] G. Di Pendina et al., J. Appl. Phys. 111, 07E350 (2012).
[6] K. Jabeur et al., IEEE TMAG, July 2014.
[7] X. Dong et al., "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," IEEE TCAD, Jul. 2012.
[8] N. Binkert et al., "The m5 simulator: modeling networked systems," IEEE Micro 26, 5260 (July 2006).

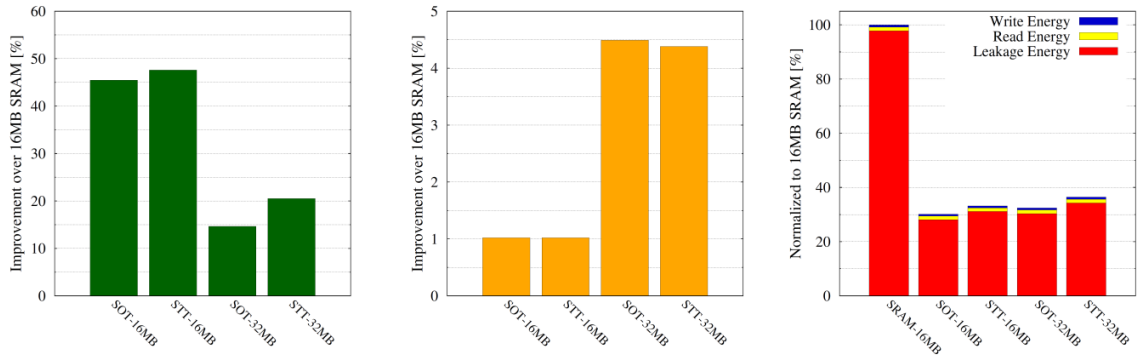

Fig. 7: Comparison of various L3-cache configurations in terms of occupied area, average application runtime and average energy consumption (normalized to the standard configuration, i.e. SRAM for all cache levels)