Architectural Aspects in Design and Analysis of SOT-based Memories

Rajendra Bishnoi, Mojtaba Ebrahimi, Fabian Oboril and Mehdi B. Tahoori
Karlsruhe Institute of Technology (KIT)
Chair of Dependable Nano Computing (CDNC)
Karlsruhe, Germany
e-mails: {rajendra.bishnoi, mojtaba.ebrahimi, fabian.oboril, mehdi.tahoori}@kit.edu

Abstract—Magnetic Random Access Memory (MRAM) is a very promising emerging memory technology because of its various advantages such as non-volatility, high density and scalability. In particular, Spin Orbit Torque (SOT) MRAM is gaining interest as it comes along with all the benefits of its predecessor Spin Transfer Torque (STT) MRAM, but is supposed to eliminate some of its shortcomings. Especially the split of read and write paths in SOT-MRAM promises faster access times and lower energy consumption compared to STT-MRAM. In this work, we provide a very detailed analysis of SOT-MRAM at both circuit- and architecture-level. We present a detailed evaluation of performance and energy related parameters and compare the novel SOT-MRAM with several other memory technologies. Our architecture-level analysis shows that with a *hybrid*-combination of SRAM for the L1-cache and SOT-MRAM for the L2-cache the energy consumption can be reduced by 63 % in average while the performance can be increased by 1 %. In addition, the memory area is 43 % lower compared to an SRAM-only configuration.

I. Introduction

As the continuous downscaling of CMOS technology becomes more and more challenging, there has been a great deal of efforts to find feasible alternatives. For random access memory (RAM), *nano-magnetic storage devices (MRAM)* are very promising candidates to replace the traditional CMOS-based memory solutions. Especially the non-volatility of MRAM is a major advantage, which minimizes static power consumption and paves the way towards normally-off/instant-on computing. In particular, MRAM based on *Magnetic Tunnel Junction*[1] *(MTJ)* [26, 27] storage devices is one of the most interesting candidates as identified by the ITRS [12]. Among these memory technologies, *Spin Transfer Torque MRAM (STT-MRAM)* [1] gains a lot of attention as it is non-volatile, scalable, and has a low read access time [4, 10, 27]. In addition, due to the high resistance of the MTJ storage elements, STT-MRAM is compatible with the CMOS process. Furthermore, the magnetization of the storage layer, and hence the stored data, can be switched without requiring an external magnetic field. Instead, a spin polarized current flowing through the MTJ device is employed.

Despite all these advantages, STT-MRAM also faces various challenges. First, although the write current is much lower than in many other MRAM technologies [10], it is still very high, leading to a high energy consumption (10x more energy per write operation than SRAM) [6, 24]. In addition, the high current through the MTJ imposes a severe stress for the memory cell. As a result, it leads to a time dependent degradation of the MTJ performance parameters such as tunneling magneto resistance, write current, and write latency. Moreover, also the lifetime is reduced, as the MTJ oxide is threatened by time dependent dielectric breakdown [20, 28]. Second, beside the high write current, also the write path itself is a challenge. In STT-MRAM, the read and write operations share the same access path (through the junction) which can impair the reliability, i.e. a read operation can by mistake lead to a bit flip (magnetization of the storage layer is switched). Third, the long write latencies usually prohibit the use of STT-MRAM in first level caches [4].

To mitigate these issues, *Spin Orbit Torque MRAM (SOT-MRAM)* has been recently proposed [7, 13, 18]. SOT-MRAM uses a three terminal MTJ-based concept to isolate the read and the write path compared to the two terminal concept of STT-MRAM. As a result, in SOT-MRAM the read and the write path are perpendicular to each other which significantly improves the read stability [13]. Moreover, the write current is much lower and also the write access is supposed to be much faster, as the write path can now be optimized independently.

In this paper, we provide a detailed circuit- and architecture-level analysis of the SOT-MRAM in both memory array design and its implications for a hybrid memory hierarchy in an advanced computing system. As we will show, the read and write latencies of SOT-MRAM are comparable to those of SRAM. In addition, SOT-MRAM offers a much higher density, lower energy consumption, is radiation immune and non-volatile. All of these aspects make SOT-MRAM a viable candidate for on-chip memory, not only for the last-level cache, but also for lower levels of cache to replace SRAM. To illustrate these benefits, we perform both circuit-level and architecture-level evaluations in which we compare SOT-MRAM with SRAM and STT-MRAM as L1- and L2-cache memory. This analysis shows that a *hybrid*-combination of SRAM for the L1-cache and SOT-MRAM for the L2-cache can reduce the energy consumption 63 % in average, while it even increases the performance slightly by 1 %. In addition, the area occupied by the memory units is 43 % lower compared to an SRAM-only solution. Even more energy savings are possible, if SOT-MRAM is used in both cache levels. However, this incurs a small performance penalty (up to 2 %).

The rest of this paper is organized as follows. In Section II, the basics of SOT-MRAM are introduced. Section III explains the details of the memory architecture using SOT-MRAM and the resulting memory characteristics such as access latencies,

[1] memory component consisting of the two magnetic layers and a barrier oxide in between storing a logic value in form of a resistance state (see Fig 1)

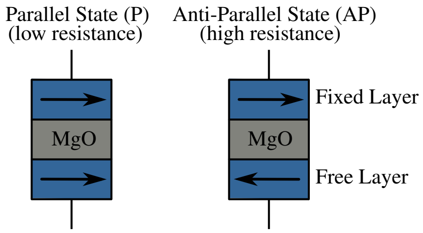


Fig. 1. MTJ resistance according to the magnetization of the free layer

energy consumption and density. Furthermore, the extracted data is compared with various other memory technologies. In addition, this information is used in Section IV to analyze the advantages and disadvantages of SOT-MRAM as a possible replacement of SRAM inside a classical memory hierarchy. Finally, Section V concludes the paper.

II. Background

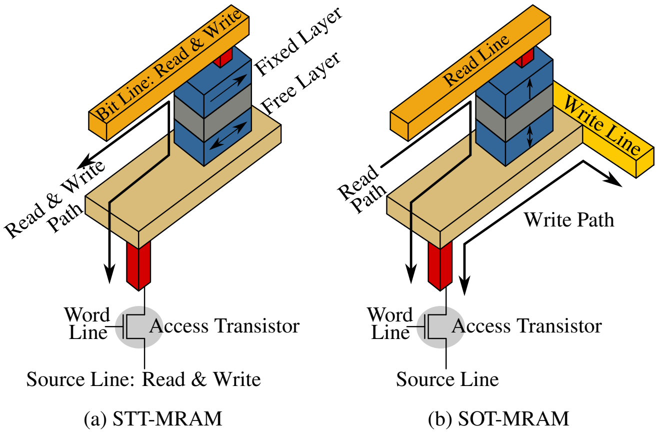
A. Magnetic Tunnel Junction Device

The storage devices in Spin Orbit Torque memories are Magnetic Tunnel Junction (MTJ) cells in which data is stored as a resistance state value. An MTJ device, as shown in Figure 1, consists of two independent ferromagnetic layers (e.g. CoFeB) separated by a very thin (≈ 1 nm) barrier oxide layer such as magnesium oxide (MgO) [13]. One of the two ferromagnetic layers has a fixed magnetization, i.e. the orientation of its magnetic field is fixed. Hence, this layer is known as *fixed* or *reference layer*. In contrast, in the second magnetic layer the magnetization can be freely rotated based on the direction of the current (i.e. spin of the electric particles) flowing through the MTJ device. Therefore, this layer is referred to as *free layer*.

When the direction of the magnetic field of the free layer is *parallel (P)* to the fixed layer, i.e. the magnetic field orientations in both layers are the same, the MTJ cell has a low resistance value. On contrary, when the magnetization of the free layer is opposite or *anti-parallel (AP)* to the fixed layer, the MTJ cell has a high resistance value. This high and low resistance values are used to represent logic ’1’ and ’0’ values.

B. SOT-MRAM Structure

The MTJ cell is the core part of a bit-cell in SOT-based memories as well as in STT-MRAM as shown in Figure 2. However, to eliminate the shortcomings of STT-MRAM, the SOT-MRAM bit-cell has an additional terminal to separate the (unidirectional) read and the (bidirectional) write path which are perpendicular to each other. The terminals comprise a *read line*, a *write line*, a *source line* and a *word line*. The word line is used to access the required bit-cell during memory accesses via the NMOS-based access transistor. If such an access is a read operation, the source line is connected to the ground and the read line is used to measure the MTJ resistance by sensing the current flowing through the MTJ cell. During the write operation the current flows between the source line and the write line. In fact, The current direction is determined by the potentials of the source line and the write line (i.e. the write path is bidirectional). The current direction in turn affects the magnetization of the free layer and hence the value stored in the bit cell. If the current flows from the source line to the write line, the MTJ resistance will be low. To achieve a high MTJ resistance, i.e. anti-parallel state, the current needs to flow from write to source line (high potential for the write line). However, the underlying physical relation between the current and the magnetic field orientation is still under discussion. On the one hand, the *Rashba effect* is said to be responsible for the current-induced magnetization switch [7, 19]. On the other hand, many people explain this phenomenon with the *Spin Hall Effect* [18]. Due to this reason, some also refer to SOT-MRAM as “Giant Spin Hall Effect” MRAM. Nevertheless, in both cases the spin-orbit-torque is responsible for the status of the free layer magnetization, which is the origin of the name SOT-MRAM.



(a) STT-MRAM

(b) SOT-MRAM

Fig. 2. Comparison of a standard bit-cell for STT-MRAM and SOT-MRAM

Since the read and write paths are independent of each other in SOT-MRAM, they can be also optimized separately. This is used to reduce the write current and write latency in SOT-MRAM compared to STT-MRAM. As we will show later, this is the reason why SOT-MRAM can achieve access times similar to SRAM, while STT-MRAM suffers from high write latencies. In addition, also the asymmetry between read and write operations can be significantly reduced, such that in SOT-MRAM read and write operations have the same access times, while in STT-MRAM a write access requires considerably more time.

It can be inferred from Figure 2 that a bit-cell consists of two different technologies, namely CMOS for the transistor and a nano-magnetic technology for the MTJ device. Therefore, the MTJ cells require additional layers in the layout and more processing steps during the fabrication process.

III. Circuit-Level Evaluation of SOT-MRAM

A. Details of the SOT-MRAM Architecture

The architecture of an SOT-MRAM memory array is shown in Figure 3. As it can be seen, similar to the SRAM memory architecture, it has a decoder which is responsible for the activation of the word line indicated by the memory address. The major difference with SRAM is in the write and read circuitry. As mentioned in Section II, the SOT bit-cell is a four terminal device which has different paths for write and read operations. In case the write enable signal is inactive, a read operation is performed by connecting the read line of the desired bit-cell to a current sense amplifier. The current sensed on the read line is

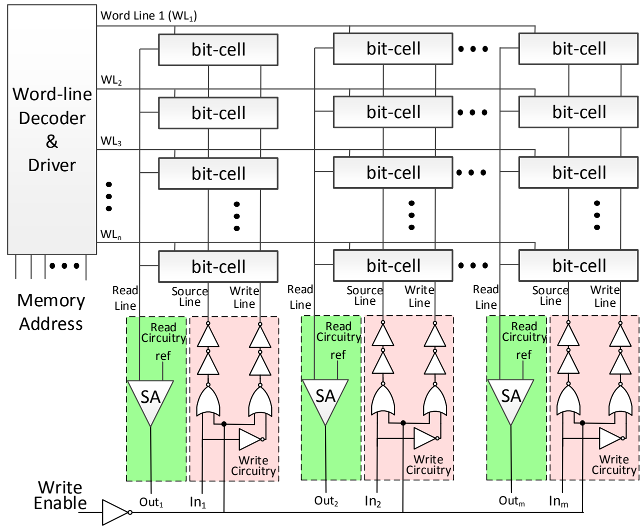


Fig. 3. Read and write operation using SOT-based bit-cell

compared with a reference value to distinguish the value stored in the bit-cell.

For the write operation, the write enable signal has to be activated. In fact, 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 determined by the input data value. 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'. For the write circuitry shown in Figure 3, it is assumed that the anti-parallel state (high resistance) represents a logical value of '1'. When the write enable signal is active and the input data has a logical value of '1', the current flows from the write line to the source line in the MTJ cell resulting in high resistance.

B. Comparison with other Memory Technologies

To investigate the SOT based memory architecture and compare it with other memory technologies, we use a multi-level approach. First, we analyze the behavior for a single bit-cell only. Afterwards, this information is used to extract the data for an entire memory array.

B.1 Circuit-Level Memory Evaluation Platform

For the bit-cell analysis of SOT-MRAM, we use the framework proposed in [13] in combination with the TSMC 65 nm general purpose library for the CMOS elements. For STT-MRAM we apply the model from [8], which employs in-plane magnetization, whereas the model for SOT-MRAM uses a perpendicular magnetic anisotropy. For both technologies, the magnetic switching dynamics for the free layers are described by the Landau-Lifshitz-Gilbert model [16].

The results of this analysis are summarized in Table I and underline the benefits of SOT-MRAM over STT-MRAM. In SOT-MRAM, the write access latency for a single bit-cell is similar to that of a read operation. However, SOT-MRAM has the same latency for the two possible write operations, i.e. write '1' (set) and write '0' (reset), while there is a huge difference for STT-MRAM. Hence, the significant asymmetry of STT-MRAM is no longer an issue for SOT-MRAM. This is mainly due to the fact that the write path in SOT-MRAM can be optimized separately as explained in Section II. Moreover, also the per-access energy and the write current of SOT-MRAM is much lower. Therefore, the access transistor of an SOT-MRAM bit-cell can be designed much smaller. This in turn leads to a lower leakage power for SOT-MRAM.

	SOT-MRAM	STT-MRAM
Read Latency [ps]	221	226
Write Latency [ps]	266	10,500 (reset) / 3,700 (set)
Write Current [uA]	100	525 (reset) / 616 (set)
Read Energy [pJ]	1.8	1.8
Write Energy [pJ]	0.1	3.9 (reset) / 3.4 (set)

TABLE I
COMPARISON OF SOT-MRAM AND STT-MRAM FOR A SINGLE BIT-CELL

Based on the results obtained from a single bit-cell, we extracted the area, read and write latency, per access energy and leakage power for a complete memory array using NVSim [5]. NVSim contains circuit-level performance, energy, and area models for various non-volatile memory technologies such as SRAM, PC-RAM, R-RAM, NAND-Flash and in particular STT-MRAM. However, the standard models used in thus tool for STT-MRAM do not consider its asynchronous write behavior (set vs. reset). Therefore, we modified NVSim to support this effect. Beside these necessary modifications for STT-MRAM, we adapt this model also for SOT-MRAM, which is possible as both technologies are very related. Moreover, we assume that the additional terminal of SOT-MRAM does not affect the bit-cell footprint. Therefore, the actual area numbers for SOT-MRAM could be higher than those reported here. For all memory technologies expect MRAM (i.e. SOT-MRAM and STT-MRAM) the default parameters of NVSim, which are based on the ITRS data, are applied for this study. For STT- and SOT-MRAM we use the previously extracted bit-cell information to feed the modified NVSim models as these are more accurate than the data provided by NVSim for STT-MRAM. All NVSim evaluations use the latency-optimized parameter set.

B.2 Comparison of SOT-MRAM with other Memory Technologies

To compare various memory technologies, we use a 512 KByte memory as a case study for which the results are summarized in Table II. For NAND-Flash, we consider the size of one page as 256 Byte and the write access energy number is reported per page. Furthermore, we report only the worst-case write latency and energy for all memories. As the results show, SOT-MRAM is competitive with SRAM in terms of performance and is even superior when it comes to energy consumption and cell density. In addition, unlike SRAM, SOT-MRAM does not have scalability limitations [1] and is also radiation immune. Although PC-RAM and R-RAM are comparable to SOT-RAM in terms of area and read latency, these memory technologies suffer significantly from their high write latency and write energy [11]. NAND-Flash has the smallest area and leakage, however it has problems with a high write energy, scalability and endurance.

Please note that for every memory technology different ways of implementation are possible, e.g. low-power, high-performance or high-density optimized versions. As a consequence, also the absolute numbers presented in Table II would change for other implementations. However, the major trends

	6T-SRAM [25]	NAND-FLASH [17,23]	STT-MRAM [15,21]	SOT-MRAM [7,13,18]	PC-RAM [22,29]	R-RAM [14]
Data Storage	Latch	Floating Gate Device	Magnetization	Magnetization	Resistance	Resistance
Non-Volatility	no	yes	yes	yes	yes	yes
Area [mm^2]	2.78	0.17	1.63	1.51	0.31	0.66
Read Latency [ns]	2.17	565.365	1.2	1.13	0.55	1.15
Write Latency [ns]	2.07	2×10^5	11.22	1.36	150.4	20.66
Read Access Energy [pJ]	587	3921	260	247	363.4	193
Write Access Energy [pJ]	355	6902	2337	334	63670	592
Leakage Power [mW]	932	77	387	254	153	115
Process	CMOS	Floating Gate Device	CMOS + STT-MTJ	CMOS + SOT-MTJ	CMOS + GST[2]	CMOS + MIM[3]
Features (based on ITRS [12])	(–) Scalability (++) Endurance (-) Radiation vulnerable	(-) Scalability (–) Endurance (-) Radiation vulnerable	(+) Scalability (+) Endurance (+) Radiation immune (-) Bit Failure Rate	(+) Scalability (+) Endurance (+) Radiation immune (-) Bit Failure Rate	(±) Scalability (-) Endurance (+) Radiation immune	(+) Scalability (-) Endurance (+) Radiation immune (-) Bit Failure Rate (-) Retention

TABLE II

COMPARISON OF VARIOUS MEMORY TECHNOLOGIES FOR A 512 KBYTE MEMORY BASED ON THE FLOW FROM SECTION III.B.1

will remain the same. Therefore, the main purpose of this analysis, as summarized in Table II, is a comparative analysis of the trends for several memory technologies and their usabilities for the on-chip memory hierarchy, rather than the actual numbers.

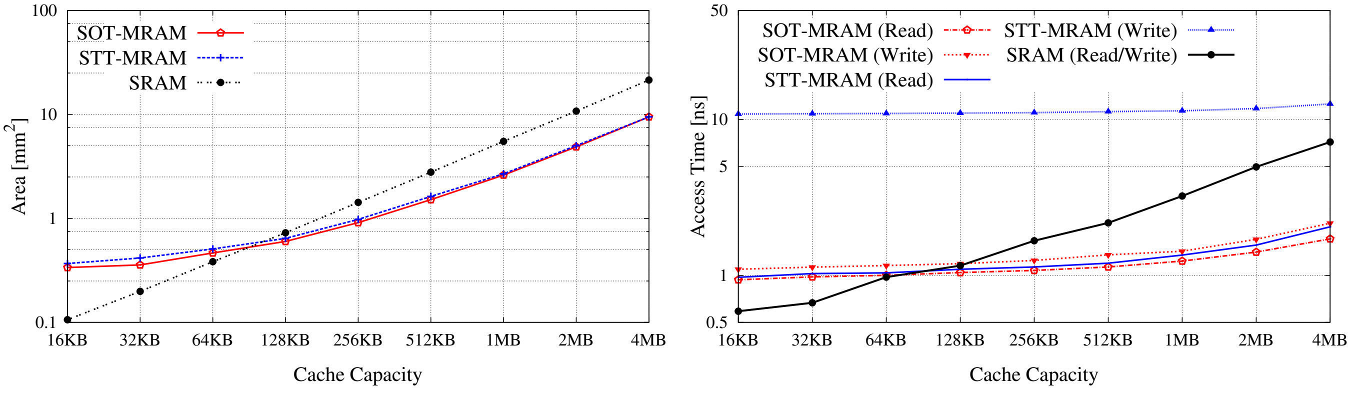
B.3 SOT-MRAM Scaling for Various Memory Sizes

Beside the analysis for a single memory size of 512 KByte, we also evaluated the most important memory parameters for SRAM (6T), STT-MRAM and in particular SOT-MRAM for various other memory sizes in the range between 16 KByte and 4 MByte using the same methodology as in the previous subsection. The results are summarized in Figure 4 as well as Figure 5 and are discussed in the following paragraphs. Please note that the actual numbers can differ based on the particular memory architecture, but the overall trends discussed here will remain the same.

Area: The first interesting observation of our analysis is the scaling behavior of the area occupied by the memory (Figure 4(a)). As it can be seen, for large memory capacities all three memory technologies show the same trend, i.e. with duplicated memory capacity also the area increases by a factor of almost 2. However, for sizes smaller than 512 KByte, the area of STT-MRAM and SOT-MRAM increases slower than the capacity. In contrast, SRAM still scales with the same trend. As a result, SRAM offers better area usage for small memory capacities, while SOT-MRAM is superior for larger sizes (here starting from 256 KByte).

To explain this phenomenon it is necessary to decompose the memory area into the total bit-cell area and the area of the periphery (i.e. write circuitry, decoder, sense amplifier). In this regard, the bit-cells for SOT-MRAM and STT-MRAM are much smaller than those for SRAM. In contrast the periphery for MRAM is larger, due to the higher write current. Both aspects together lead to the fact that, in case of MRAM, the size of the periphery dominates or is similar to the total bit-cell area for memory capacities below 64 KByte. Furthermore, the size of the periphery does not scale linearly with the memory capacity, while the total bit-cell area does. Hence, for small memory capacities, the scaling of SOT-MRAM and STT-MRAM is limited by the size of the memory periphery, while for SRAM the total bit-cell area is the limiting factor and thus it scales better.

Please note that the actual area numbers for SOT-MRAM could be higher than those reported here, since we assume that the additional bit-cell terminal does not increase the bit-cell footprint compared to STT-MRAM. In fact, the overhead due to the additional terminal depends on various aspects, e.g. design rules and size of the access transistor. Therefore, as SOT-



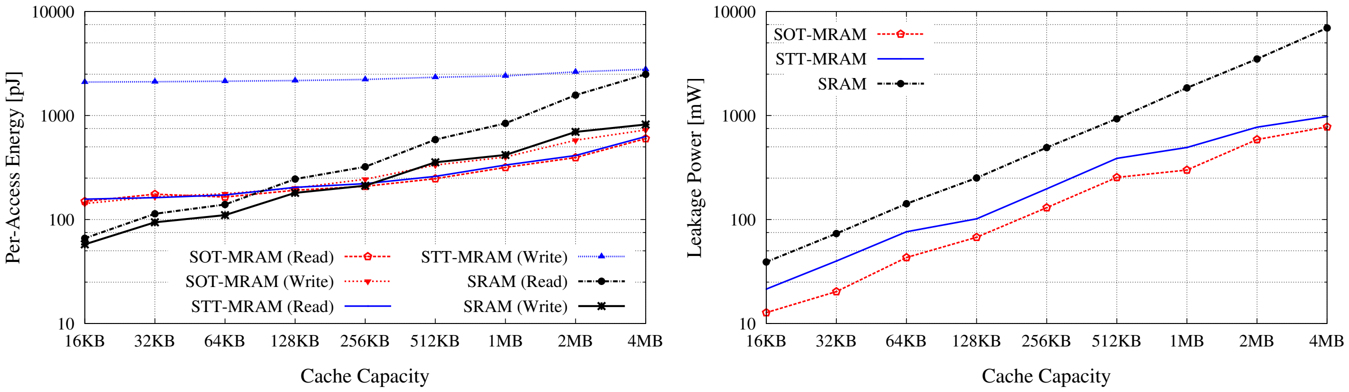
(a) Area

(b) Read & Write Latency

Fig. 4. Area and latency scaling behavior for SRAM, STT-MRAM and SOT-MRAM for various memory sizes

[2] GST: An alloy for Phase change material $Ge_2Sb_2Te_5$

[3] MIM : Metal-Insulator-Metal component



(a) Per-Access Energy

(b) Leakage Power

Fig. 5. Access energy and leakage scaling behavior for SRAM, STT-MRAM and SOT-MRAM for various memory sizes

MRAM is not yet in production, it is not possible to quantify the additional area required due to the fourth terminal.

Access Latencies: Another interesting phenomenon can be observed for the scaling behavior of the access latencies (see Figure 4(b)). Since the load capacitance of an SRAM-based bit-cell is much higher than that of an MTJ-based bit-cell, as the latter is much smaller, the access latencies of SRAM are stronger correlated to the number of bit-cells than those of SOT-MRAM or STT-MRAM. For MRAM memories, in the evaluated size range, the major contributor is the latency of the periphery circuitry and the routing delay. Thus, the access latencies of SOT-MRAM and STT-MRAM do not increase as much as those of SRAM with increasing memory size. As a result, although SRAM is the fastest memory technology for very small memory sizes, it is slower than SOT-MRAM for both read and write operations for larger memory sizes. While STT-MRAM is comparable to SOT-MRAM in terms of read latency, it suffers from its very long write latency. This underlines how effective the separation of read and write paths and hence their independent optimization in SOT-MRAM is. As a result, the asynchronous access behavior (almost) disappears for SOT-MRAM.

Per-Access Energy: The per-access energy shows a similar behavior as the access latencies as shown in Figure 5(a). Thereby, the reasons are the same as explained in the previous paragraph. As a result, SRAM is again the best choice for very small memories, but for larger memories (here: starting with 256 KByte) SOT-MRAM starts to become the better solution. In contrast, STT-MRAM has a very high write-access energy, due to the high write current required [3].

Leakage Power: In terms of leakage power SOT-MRAM is superior compared to STT-MRAM and SRAM. The reason for the high leakage power of SRAM is its CMOS nature. For STT-MRAM it is the larger access transistor compared to SOT-MRAM which is the reason why its leakage power is worse than that of SOT-MRAM.

Summary: In summary, based on our observations, SOT-MRAM is a very good replacement for SRAM in cache memories. However, its suitability for an L1-cache compared to SRAM strongly depends on the size of this cache and the clock frequency. For slower clock frequencies or larger cache sizes SOT-MRAM could be a viable choice even for L1 cache. However, the real cache performance depends not only on these parameters but also the application and its characteristics, e.g. read to write ratio or hit rate. Therefore, in the following section, we present a detailed study of SOT-MRAM as a candidate in various levels of the cache hierarchy in a real system.

IV. Evaluation of SOT-MRAM as Cache Memory

Based on the comparison of various memory technologies presented in Section III.B, SOT-MRAM is a promising candidate to (partially) replace SRAM as the memory technology for caches in microprocessors. Therefore, we analyze the advantages and disadvantages of SOT-MRAM as L1- and L2-cache memory technology in terms of performance, energy consumption as well as area. For this reason, various "*hybrid*" cache configurations are evaluated in which different memory technologies (SRAM, STT-MRAM, and SOT-MRAM) are used for different levels of cache hierarchy.

A. Hybrid-Memory Evaluation Platform

Our evaluation uses gem5 [2], a full-system, cycle-accurate performance simulator that supports various memory configurations and allows to configure all relevant cache parameters such as capacity, associativity, latency, block size and policy. However, to model the asymmetric behavior of STT- and SOT-

Processor	Single-core @ 3 GHz, out-of-order, 4-issue
L1-Cache	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/10.9 ns)
L2-Cache	512 KByte, 16-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/11.2 ns)
Execution Units	2x ALU, 2x CALU, 2x FPU
MiBench applications	BasicMath, BitCount, QSort, Dijkstra, Patricia, StringSearch, SHA, CRC, FFT

TABLE III
Configuration details for the experiments

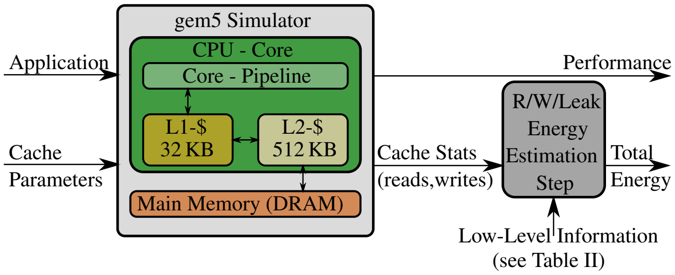


Fig. 6. Analysis flow to obtain performance & energy consumption for different cache configurations

MRAM we had to extend gem5 to support different read and write latencies for each cache.

The baseline configuration for our study is summarized in Table III. It is based on a single-core processor with a clock frequency of 3 GHz and an out-of-order pipeline based on the Alpha 21264 processor. Furthermore, the processor has an L1-cache with a capacity of 32 KByte and its L2-cache is 512 KByte large. For each memory technology we extracted the read and write access latencies for the L1- and L2-cache according to the methodology presented in Section III.B. Please note that due to the chosen clock frequency of 3 GHz the latencies correspond to 3 cycles and 7 cycles for SRAM, 3 (4) and 4 (5) cycles for SOT-MRAM for read (write) accesses, and 3 (33) and 4 (34) cycles for STT-MRAM for read (write accesses), for L1 and L2 caches respectively. This indicates already that the final performance does not only depend on the memory technology for each cache level, but also the clock frequency for each cache.

To evaluate the benefits and shortcomings of SOT-MRAM as cache memory, we use nine workloads out of the MiBench benchmark suite [9] as detailed in Table III. All workloads are simulated completely, including the initialization phase, to be as close as possible to the real world. Afterwards, the performance and cache statistics obtained from gem5 are used to estimate the dynamic (read & write) and static energy (leakage) for each memory configuration. Therefore, for each memory technology the per access energy and leakage power are taken into consideration. By considering also the number of read (write) accesses and the runtime, the total energy consumption for every application can be estimated as shown in Figure 6.

B. Main Results

The main results of our analysis are summarized in Figure 7. For this figure and all further discussions a configuration such as SRAM+SOT means that SRAM is used for the L1-cache, while SOT-MRAM is used for the L2-cache.

To explain these, we first focus on the area, afterwards on the performance and then discuss the energy consumption.

Area: As expected the usage of SOT-MRAM or STT-MRAM significantly reduces the cache area, which is due to the fact that both technologies have much smaller bit-cells than SRAM. In this regard, the major savings can be achieved, if SOT-MRAM or STT-MRAM are used for the L2-cache, since it occupies much more area due to the higher capacity. If SOT-MRAM instead of SRAM is employed for the L2-cache, area can be reduced by more than 40 %. As the L1-cache in our study is quite small, the phenomenon discussed in Section B.3 occurs, i.e. for this cache size SRAM is smaller than SOT-MRAM. Hence, if the L1-cache uses SOT-MRAM as memory technology, the size increases by 5 %.

Performance: The results also show that SOT-MRAM can replace SRAM in terms of performance, while STT-MRAM suffers from its long write latency as expected based on the analysis presented in Section III.B.3. However, the benefits strongly depend on the cache-level. For the L1-cache, SRAM offers in average 1 % more performance than SOT-MRAM, while for the L2-cache SOT-MRAM is slightly faster, which results in a slight performance increase (i.e. runtime reduction) of 1 % in average. Since the write access latency of the L2-cache is not so important, even STT-MRAM can be used for this cache-level [4]. However, for the L1-cache it is not feasible (i.e. runtime increase of 30 % in average).

Energy: To analyze the energy consumption, let us first focus on the L2-cache. As Figure 7 shows, SRAM is not competitive

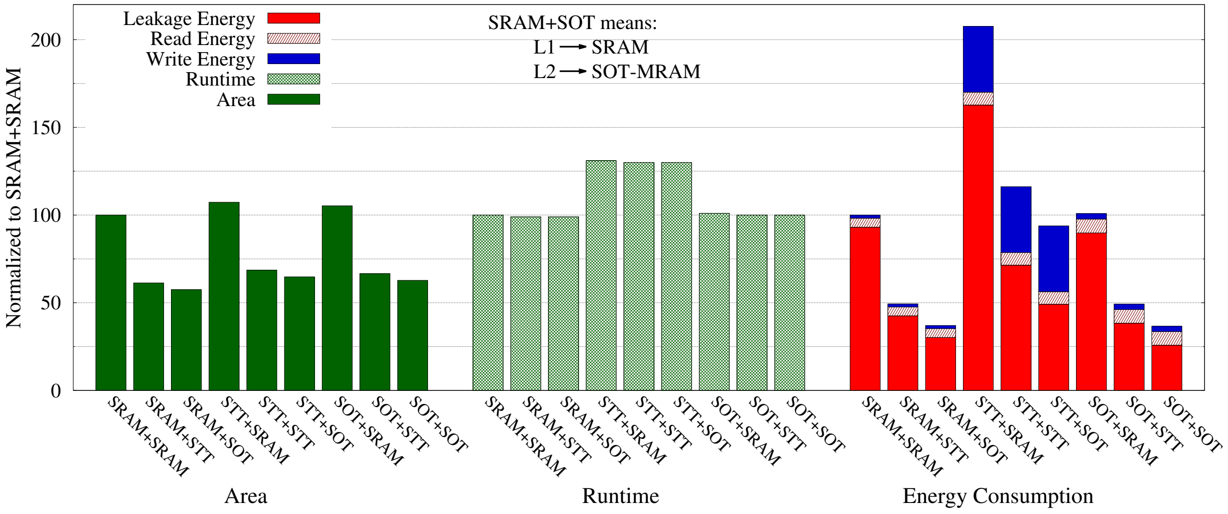


Fig. 7. Comparison of various cache configurations in terms of occupied area, average application runtime and average energy consumption (normalized to the standard configuration, i.e. SRAM for L1- and L2-cache)

	Runtime [ms]				Energy [mJ]			
	SRAM+SRAM	SRAM+STT	SRAM+SOT	SOT+SOT	SRAM+SRAM	SRAM+STT	SRAM+SOT	SOT+SOT
BasicMath	61.4	**59.8**	**59.8**	60.5	66.4	31.6	23.6	**22.8**
BitCount	130.1	**130.1**	**130.1**	130.1	133.8	63.0	45.6	**40.4**
CRC	998.8	**998.8**	**998.8**	1025.5	1075	531.5	398.1	**395.7**
Dijkstra	62.7	**62.4**	**62.4**	62.6	75.5	41.2	**32.9**	36.8
FFT	176.1	175.4	**175.3**	176.1	191.9	95.5	72	**71.6**
Patricia	49.1	46.7	**46.7**	47.6	54.6	25.8	19.5	**19.4**
QSort	35.2	34.9	**34.9**	34.9	36.7	17.6	12.7	**11.6**
SHA	23.3	**23.3**	**23.3**	23.3	26.1	13.4	**10.3**	10.7
StringSearch	1.5	**1.5**	**1.5**	1.5	1.7	0.9	0.7	**0.7**
Average	170.9 (100 %)	170.0 (99 %)	**170.0 (99 %)**	173.3 (101 %)	184.6 (100 %)	91.2 (49 %)	68.4 (37 %)	**67.7 (36 %)**

TABLE IV
PER BENCHMARK ANALYSIS OF DIFFERENT "HYBRID" CACHE CONFIGURATION (BOLD NUMBERS REPRESENT THE BEST VALUE)

with STT-MRAM or SOT-MRAM for this cache-level. This is due to the high leakage power of SRAM and the memory capacity of 512 KByte. If instead MRAM is used, the energy consumption can be reduced by more than 50 %, in average. Furthermore, as explained in Section B.3 the leakage power of SOT-MRAM is also smaller than that of STT-MRAM, due to smaller access transistors. As a result, SOT-MRAM offers the least power hungry solution for the L2-cache.

In contrast, for the small L1-cache (just 32 KByte), SRAM requires less energy than STT-MRAM. This is due to two facts. First, the per-access energy of STT-MRAM is much higher than that of SRAM, especially due to the high write current required by STT-MRAM. Second, leakage power is less important for this small memory. Both aspects together lead to a 40 % increase energy consumption compared to SRAM in average (if SOT-MRAM is used for the L2-cache). Since SOT-MRAM has a much lower per-access energy consumption than STT-MRAM, it eliminates a major shortcoming and thus allows to even reduce the energy consumption of the L1-cache compared to SRAM.

Summary: In summary, SOT-MRAM is a viable candidate to replace SRAM as memory technology for some levels of the cache hierarchy. It does not only offer a higher density and lower energy consumption but has also a similar performance. However, for smaller cache sizes such benefits reduces accordingly. As a consequence, the per-access energy gains importance and in turn SOT-MRAM looses advantages. Based on our observations, SOT-MRAM is a viable SRAM replacement for the L2-cache and in some cases even for the L1-cache, if it is large enough (in our setup, at least 64 KBytes). In other words, when the L1-cache size is small enough, SRAM is still a better choice. Moreover, for register files, due to their small sizes, SOT-MRAM is not a suitable choice.

C. In-Depth Evaluation

In Table IV the results per benchmark for the hybrid cache-configurations SRAM+SRAM, SRAM+STT, SRAM+SOT and SOT+SOT are shown. As it can be seen, SOT+SOT is in average the solution with the lowest energy consumption and hence is the best choice for low power systems. However, for some applications the combination of SRAM for the L1-cache and SOT-MRAM for the L2-cache offers a slightly better energy efficiency (e.g. Dijkstra or SHA). This is due to the fact that the combination of SRAM and SOT-MRAM is often faster and hence has the advantage of a lower runtime. In addition, the per-access energy for SRAM as L1-cache is lower than that of SOT-MRAM as memory technology for the L1-cache.

Furthermore, it can be seen that often SRAM+STT and SRAM+SOT deliver the same performance. This is the case for applications that have a low write access rate to the L2-cache (e.g. StringSearch or BasicMath). If the ratio of write access to the L2-cache is higher (e.g. FFT) SRAM+SOT is a better solution in terms of performance as STT-MRAM has much higher write access times. In terms of energy consumption SRAM+SOT is always much better than SRAM+STT.

The combination of SRAM+SRAM is neither the fastest nor the most energy saving solution for any benchmark. Hence, this configuration is, at least for our setup, not a viable choice. Instead a hybrid solution or SOT-MRAM-only is favorable. However, considering all aspects, i.e. performance, energy consumption and area, the hybrid solutions offers the best trade-off for our processor configuration.

V. CONCLUSIONS

For shrinking technologies, non-volatile memories are promising storage technologies due to their low static power. In this paper, we evaluated a novel nano-magnetic memory technology called Spin Orbit Torque (SOT-MRAM). It is related to Spin Transfer Torque MRAM (STT-MRAM), but has independent read and write paths. As a result SOT-MRAM can achieve access latencies similar to SRAM which makes SOT-MRAM a viable candidate for on-chip memory, not only for the last-level cache, but also for lower levels of cache to replace SRAM. Depending on the cache size, SOT-MRAM can even replace SRAM as memory technology for the L1-cache. In fact, our detailed architecture-level analysis shows that an SOT-only solution is the best choice for low power systems. We also found out that for very small memory blocks, such as register files or small L1-caches, SRAM is still superior to SOT-MRAM in terms of area and performance. Therefore, the best combination of performance, energy efficiency and area cost is offered by a "hybrid" solution composed of SRAM for the small L1-cache (32 KByte) and SOT-MRAM for the larger L2-cache (512 KByte). Compared to an SRAM-only configuration this allows to reduce the energy consumption by 63 %, the area by 43 % and in addition the performance will increase by 1 %.

VI. Acknowledgement

This work was partly supported by the European Commission under the Seventh Framework Program as part of the spOt project (http://www.spot-research.eu/). Furthermore, the authors would like to thank Gregory di Pendina, Kotb Jabeur and Guillaume Prenat from SPINTEC for their valuable feedback and discussion during this work.

References

[1] D. Apalkov, A. Khvalkovskiy, S. Watts, V. Nikitin, X. Tang, D. Lottis, K. Moon, X. Luo, E. Chen, A. Ong, A. Driskill-Smith, and M. Krounbi, "Spin-transfer torque magnetic random access memory (STT-MRAM)," *ACM Journal on Emerging Technologies in Computing Systems*, pp. 13:1–13:35, May 2013.

[2] N. L. Binkert, R. G. Dreslinski, L. R. Hsu, K. T. Lim, A. G. Saidi, and S. K. Reinhardt, "The M5 Simulator: modeling networked systems," *IEEE Micro*, pp. 52–60, Jul. 2006.

[3] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. Tahoori, "Asynchronous asymmetrical write termination (AAWT) for a low power STT-MRAM," in *Design, Automation and Test in Europe*, Mar. 2014.

[4] M.-T. Chang, P. Rosenfeld, S.-L. Lu, and B. Jacob, "Technology comparison for large last-level caches (L^3Cs): low-leakage SRAM, low write-energy STT-RAM, and refresh-optimized eDRAM," in *High Performance Computer Architecture*, Feb. 2013, pp. 143–154.

[5] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "NVSIM: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, pp. 994–1007, 2012.

[6] X. Dong, X. Wu, G. Sun, Y. Xie, H. Li, and Y. Chen, "Circuit and microarchitecture evaluation of 3D stacking magnetic RAM (MRAM) as a universal memory replacement," in *Design Automation Conference*, Jun. 2008, pp. 554–559.

[7] D. A. P. Gambardella and I. M. Miron, "Current-induced spin-orbit torques," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, pp. 3175–3197, 2011.

[8] W. Guo, G. Prenat, V. Javerliac, M. E. Baraji, N. de Mestier, C. Baraduc, and B. Diny, "SPICE modelling of magnetic tunnel junctions written by spin-transfer torque," *Journal of Physics D: Applied Physics*, p. 215001, May 2010.

[9] M. R. Guthaus, J. S. Ringenberg, D. Ernst, T. M. Austin, T. Mudge, and R. B. Brown, "MiBench: a free, commercially representative embedded benchmark suite," in *Workshop on Workload Characterization*, 2001, pp. 3–14.

[10] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto *et al.*, "A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM," in *International Electron Devices Meeting*, 2005, pp. 459–462.

[11] J. Hutchby and M. Garner, "Assessment of the potential & maturity of selected emerging research memory technologies," in *Workshop and ERD/ERM working group meeting*, Apr. 2010.

[12] International Technology Roadmap for Semiconductors, http://www.itrs.net, 2012.

[13] K. Jabeur, L. D. Buda-Prejbeanu, G. Prenat, , and G. D. Pendina, "Study of two writing schemes for a magnetic tunnel junction based on spin orbit torque," *International Journal of Electronics Science and Engineering*, pp. 501–507, 2013.

[14] D. S. Jeong, R. Thomas, R. S. Katiyar, J. F. Scott, H. Kohlstedt, A. Petraru, and C. S. Hwang, "Emerging memories: resistive switching mechanisms and current status," *Reports on Progress in Physics*, p. 076502, 2012.

[15] A. Khvalkovskiy, D. Apalkov, S. Watts, R. Chepulskii, R. Beach, A. Ong, X. Tang, A. Driskill-Smith, W. Butler, P. Visscher *et al.*, "Basic principles of STT-MRAM cell operation in memory arrays," *Journal of Physics D: Applied Physics*, pp. 74 001–74 020, 2013.

[16] L. D. Landau and E. Lifshitz, "On the theory of the dispersion of magnetic permeability in ferromagnetic bodies," *Phys. Zeitsch. der Sow.*, pp. 153–169, 1935.

[17] Y. Li and K. N. Quader, "NAND Flash memory: challenges and opportunities," *Computer*, pp. 23–29, 2013.

[18] L. Liu, C.-F. Pai, Y. Li, H. W. Tseng, D. C. Ralph, and R. A. Buhrman, "Spin-torque switching with the giant spin hall effect of tantalum," *Science*, pp. 555–558, 2012.

[19] I. M. Miron, K. Garello, G. Gaudin, P. Zermatten, M. Costache, S. Auffret, S. Bandiera, B. Rodmacq, A. Schuhl, and P. Gambardella, "Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection," *Nature*, pp. 189–193, 2011.

[20] G. Panagopoulos, C. Augustine, and K. Roy, "Modeling of dielectric breakdown-induced time-dependent STT-MRAM performance degradation," in *Device Research Conference*, 2011, pp. 125–126.

[21] D. Ralph and M. D. Stiles, "Spin transfer torques," *Journal of Magnetism and Magnetic Materials*, pp. 1190–1216, 2008.

[22] S. Raoux, G. W. Burr, M. J. Breitwisch, C. T. Rettner, Y.-C. Chen, R. M. Shelby, M. Salinga, D. Krebs, S.-H. Chen, H.-L. Lung *et al.*, "Phase-change random access memory: a scalable technology," *IBM Journal of Research and Development*, pp. 465–479, 2008.

[23] N. Shibata, H. Maejima, K. Isobe, K. Iwasa, M. Nakagawa, M. Fujiu, T. Shimizu, M. Honma, S. Hoshi, T. Kawaai *et al.*, "A 70 nm 16 Gb 16-level-cell NAND flash memory," *Solid-State Circuits, IEEE Journal of*, pp. 929–937, 2008.

[24] G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen, "A novel architecture of the 3D stacked MRAM L2 cache for CMPs," in *High Performance Computer Architecture*, 2009, pp. 239–249.

[25] N. Weste and D. Harris, *CMOS VLSI Design: a circuits and systems perspective*, 4th ed. USA: Addison-Wesley Publishing Company, 2010.

[26] S. A. Wolf, D. D. Awschalom, R. A. Buhrman, J. M. Daughton, S. von Molnr, M. L. Roukes, A. Y. Chtchelkanova, and D. M. Treger, "Spintronics: a spin-based electronics vision for the future," *Science*, pp. 1488–1495, 2001.

[27] S. A. Wolf, J. Lu, M. R. Stan, E. Chen, and D. M. Treger, "The promise of nanomagnetics and spintronics for future logic and universal memory," *Proceedings of the IEEE*, pp. 2155–2168, 2010.

[28] C. Yoshida, M. Kurasawa, Y. M. Lee, K. Tsunoda, M. Aoki, and Y. Sugiyama, "A study of dielectric breakdown mechanism in CoFeB/MgO/CoFeB magnetic tunnel junction," in *International Reliability Physics Symposium*, 2009, pp. 139–142.

[29] P. Zhou, B. Zhao, J. Yang, and Y. Zhang, "A durable and energy efficient main memory using phase change memory technology," in *ACM SIGARCH Computer Architecture News*, vol. 37, no. 3, 2009, pp. 14–23.