Evaluation of Hybrid Memory Technologies using SOT-MRAM for On-Chip Cache Hierarchy

Fabian Oboril, Rajendra Bishnoi, Mojtaba Ebrahimi and Mehdi B. Tahoori

***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 a *hybrid*-combination of SRAM for the L1-Data-cache, SOT-MRAM for the L1-Instruction-cache and L2-cache can reduce the energy consumption by 60 % while the performance increases by 1 % compared to an SRAM-only configuration. Moreover, the retention failure probability of SOT-MRAM is 27x smaller than the probability of radiation-induced Soft Errors in SRAM, for a 65 nm technology node. All of these advantages together make SOT-MRAM a viable choice for microprocessor caches.**

***Index Terms*—spin orbit torque, non-volatile memory, cache, hybrid, magnetic memory, reliability, failure rate**

I. Introduction

As the continuous downscaling of CMOS technology becomes more and more challenging, the microelectronic industry makes huge 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(MTJ)* [1, 2] storage devices is one of the most interesting candidates as identified by the ITRS [3]. Among these memory technologies, *Spin Transfer Torque MRAM (STT-MRAM)* [4] gains a lot of attention as it is non-volatile, scalable, and has a low read access time [2, 5, 6]. In addition, due to the high resistance of the MTJ storage elements, STT-MRAM is compatible with the CMOS process [7]. 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 [6], it is still very high, leading to a high energy consumption (10x more energy per write operation than SRAM) [8, 9]. 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 [10]. Moreover, the lifetime is reduced, as the MTJ oxide is threatened by time dependent dielectric breakdown [11, 12]. Second, beside the high write current, the write path itself is also a challenge. In STT-MRAM, the read and write operations share the same access path (through the junction) which can impair the reliability (*read disturb*), i.e. a read operation can by mistake lead to a bit flip (magnetization of the storage layer is switched) [13]. Third, the long write latencies usually prohibit the use of STT-MRAM in first level caches [5].

To mitigate these issues, *Spin Orbit Torque MRAM (SOT-MRAM)* has been recently proposed [7, 14, 15]. 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 [7, 16]. 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.

To evaluate the concept of SOT-MRAM and its implications at various design levels, we provide in this paper a detailed circuit- and architecture-level analysis of 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.

A preliminary version of this work was published in [17]. In this paper, we extend our preliminary work with a quantitative reliability analysis. Therefore, we compare the radiation-induced soft error rate of an SRAM-based cache with the error rate due to retention failures as well as read disturb faults for SOT-MRAM and STT-MRAM. In addition, we provide a more

Manuscript received July 25, 2014; revised September 12, 2014; accepted December 2, 2014.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/).

The authors are with Karlsruhe Institute of Technology (KIT), Germany, 76131 Karlsruhe. E-mail: {fabian.oboril,rajendra.bishnoi,mojtaba.ebrahimi, mehdi.tahoori}@kit.edu

detailed performance and energy analysis for various hybrid cache configurations and evaluate the impact of SOT-MRAM on the instruction and data caches.

To illustrate the benefits of SOT-MRAM, 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. The main results of this analysis can be summarized as follows:

1) A *hybrid*-combination of SRAM for the L1-Data-cache, SOT-MRAM for the L1-Instruction-cache and L2-cache is 1 % faster compared to an SRAM-only solution. In addition, it reduces the energy consumption by 60 % and the area by 30 %.
2) An SOT-MRAM implementation is the most energy efficient solution saving up to 71 %.
3) Using the area advantage of SOT-MRAM one can double the size of the L2-cache, which results in 6 % more performance, while still saving 56 % of energy compared to an SRAM-only approach with small L2-cache.
4) Due to the performance advantage of SOT-MRAM over STT-MRAM, retention failures are less likely in SOT-MRAM. In a 65 nm technology node the failure probability is 27x lower than the probability of radiation-induced soft errors in SRAM. However, the scaling projections indicate that in future technology nodes, the retention failure rate in SOT-MRAM will be comparable with the soft error rate in SRAM. Therefore, designers need to find suitable means to keep the retention failure probability on an acceptable level. In this case, SOT-MRAM is a fast and very reliable technology.

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. Furthermore, the extracted data is compared with various other memory technologies. In addition, this information is used in Section IV to analyze SOT-MRAM as a possible replacement of SRAM inside a classical memory hierarchy. Afterwards, a quantitative reliability comparison of SRAM and SOT-MRAM is presented in Section V. Finally, Section VI concludes the paper.

II. Background

A. Magnetic Tunnel Junction Device

The storage device in Spin Orbit Torque memories is a Magnetic Tunnel Junction (MTJ) cell 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 (a few nm) barrier oxide layer such as magnesium oxide (MgO) [7]. 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 current direction (i.e. spin of the electric particles) flowing through the MTJ device. Therefore, this layer is referred to as *free layer*.

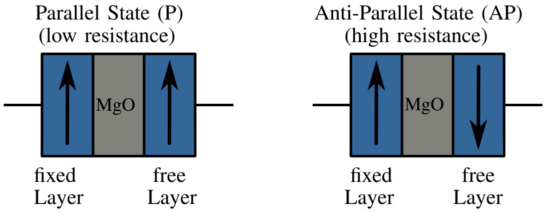
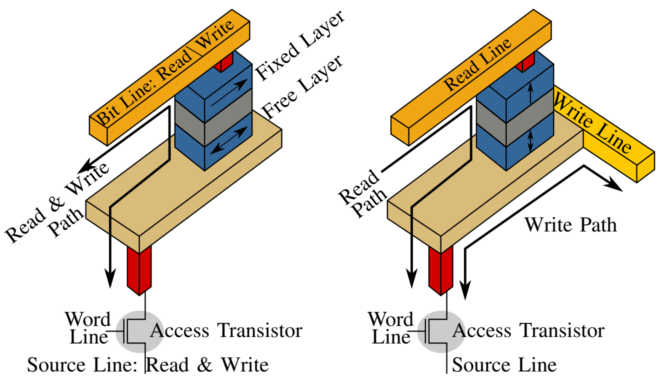


Fig. 1. MTJ resistance according to the magnetization of the free layer

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. Instead, 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 Bit-Cell 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, 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 [14, 18]. On the other hand, many people explain this phenomenon with the *Spin Hall Effect* [15], and hence refer to SOT-MRAM as "Giant Spin Hall Effect" MRAM. Nevertheless, in both cases the spin-



(a) STT-MRAM (b) SOT-MRAM

Fig. 2. Bit-cell for STT-MRAM and SOT-MRAM

orbit-torque is responsible for the free layer magnetization, which is the origin of the name SOT-MRAM.

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, it requires additional layers in the layout and more processing steps during fabrication.

C. Comparison of SOT-MRAM and STT-MRAM

The main difference between STT-MRAM and SOT-MRAM is that SOT-MRAM has separate paths for read and write operations. Hence, these paths can be also optimized independently. This can be 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, the asymmetry between read and write operations can be significantly reduced, such that in SOT-MRAM read and write operations have similar access times, while in STT-MRAM a write access requires considerably more time.

Furthermore, the probability of *read disturb*, i.e. that a read operation accidentally flips the bit-cell value, is negligible in SOT-MRAM [7, 16] due to the separated read and write paths, as well as the MTJ design, while it is an important source of unreliability in STT-MRAM [19]. In fact, in STT-MRAM the current for writing '0' ('P' state) and the read current share the same path and are in the same direction, which can cause read disturb. In contrast, in SOT-MRAM, the write path is always perpendicular to the read path, which avoids read disturb faults. Read disturb is also the reason, why in STT-MRAM read and write paths cannot be optimized independently, while this is possible in SOT-MRAM. In STT-MRAM it is very important to maintain a certain ratio between the read and the write current, to avoid high read disturb rates. If the read current is increased to achieve a better readability (reduced read error rate), the write current has to be increased as well, as otherwise the read disturb rate would increase. The same problem arises, if the write current should be reduced to reduce the write energy. In this case, the read current needs to be lowered as well, which however impairs the readability. In contrast, since in SOT-MRAM read and write paths are separated, the read and write currents can be optimized independently to co-optimize readability, access latencies and energy consumption. This tuning of the paths is achieved by designing the read and write circuitries accordingly (i.e. use smaller/larger transistors to reduce/increase the current)

A common reliability challenge of SOT-MRAM and STT-MRAM is the *retention failure*, which is due to an inherent thermal instability of the MTJ cells [19]. This thermal instability can lead to data loss which reduces the retention time.

Note that in this work the in-plane STT-MRAM technology is employed for our experimental analysis. Beside this implementation, there is also a perpendicular STT-MRAM solution. The difference between these two approaches is that in in-plane STT-MRAM the magnetic orientations of the MTJ layers are orthogonal to the current direction (read or write), while these are in parallel to the current direction (read or write) in the perpendicular implementation. Therefore, the latter version typically requires less write current and is more energy efficient [4, 20]. In fact, the reported current data and switching latencies vary a lot in literature. For in-plane STT-MRAM the typical write current is in the range of 100 uA-1000 uA with switching latencies between 2 ns and 12 ns [11, 21–27], while for perpendicular STT-MRAM the write current is usually between 30 uA-300 uA with write latencies in the range of 0.4 ns-45 ns [4, 20, 21, 28–31]. In this regard, always more current leads to lower write latencies for all STT-MRAM and SOT-MRAM approaches. The model for in-plane STT-MRAM used in this work is based on real silicon data [23] and requires a write current of 525 uA with a switching time of 10.5 ns (see Section III-B1 for more details), which is in the range of the previously mentioned data.

In both STT-MRAM approaches there is a considerable asymmetry between read and write delays as well as between the switching time from P→AP and AP→P [27, 28], which is not the case for SOT-MRAM. In addition, the read and write paths are shared and hence, even the perpendicular STT-MRAM version cannot achieve the same efficiency as SOT-MRAM, where read and write paths can be optimized independently [7]. Moreover, in SOT-MRAM the write path has a much lower resistance as the write path in STT-MRAM, which is going through the MTJ cell [7]. Thus, higher currents can be used for SOT-MRAM to allow faster write operations. In contrast, increasing the write current in STT-MRAM may lead to accelerated wearout of the MTJ cell. In addition, the switching behavior in SOT-MRAM is free of incubation as it is a surface effect relying on spin orbit torque, whereas in STT-MRAM the incubation time can significantly increase the switching speed [32]. As a result, if perpendicular STT-MRAM is used instead of in-plane, the STT-MRAM results shown in Section III and Section IV can improve, but not reach to the level of SOT-MRAM.

III. Circuit-Level Analysis 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

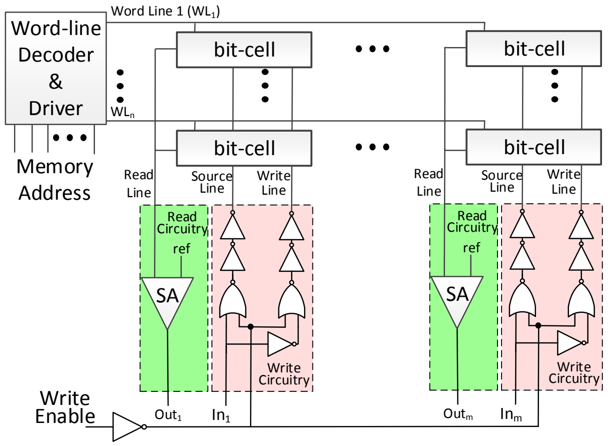


Fig. 3. Read and write operation using SOT-based bit-cell

	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 with one access transistor

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 the corresponding sense amplifier. The current sensed on the read line is 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.

1) Circuit-Level Memory Evaluation Platform: For the bit-cell analysis of SOT-MRAM, we use the framework proposed in [7] in combination with the TSMC 65 nm general purpose library for the CMOS elements. For STT-MRAM we apply the model from [23], which employs in-plane magnetization. For both technologies, the models are based on real silicon data [7, 22, 23, 41], and the switching dynamics for the free layers are described by the Landau-Lifshitz-Gilbert model [42].

The results of this analysis are summarized in Table I and underline the benefits of SOT-MRAM over (in-plane) STT-MRAM. In SOT-MRAM, the write access latency for a single bit-cell is similar to that of a read operation. Furthermore, SOT-MRAM has almost the same latency for the two possible write operations, i.e. write '1' (set) and write '0' (reset) [7], while there is a huge difference for STT-MRAM. Hence, the significant asymmetry of STT-MRAM is not 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 before. Moreover, the per-access energy and the write current of SOT-MRAM are much lower. Therefore, the access transistor and the write circuitry of an SOT-MRAM bit-cell can be designed much smaller. This in turn leads to a lower leakage power for SOT-MRAM. Note that for both technologies the write current can be reduced at the cost of an increased write latency [21].

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 [43] with its latency-optimized parameter set. 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 this tool for STT-MRAM do not consider its asymmetric 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. For all memory technologies except 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.

2) Comparison of SOT with Other 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

	6T-SRAM	NAND-FLASH	In-plane STT-MRAM	SOT-MRAM	PC-RAM	R-RAM
	[33]	[34, 35]	[23, 36, 37]	[7, 14, 15]	[38, 39]	[40]
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	1.80	0.31	0.66
Read Latency [ns]	2.17	565.37	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†	CMOS + MIM‡
Features (based on ITRS [3])	(–) 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-B1
(†: GST is an alloy for phase change material $Ge_2Sb_2Te_5$, ‡: MIM stands for Metal-Insulator-Metal component)

latency and energy (with respect to set/reset operations and location of the bit-cell). As the results show, SOT-MRAM is comparable to 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 [4] and can be considered as radiation immune. Although PC-RAM and R-RAM are comparable to SOT-MRAM in terms of area and read latency, these memory technologies suffer from their high write latency and write energy [44]. 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 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.

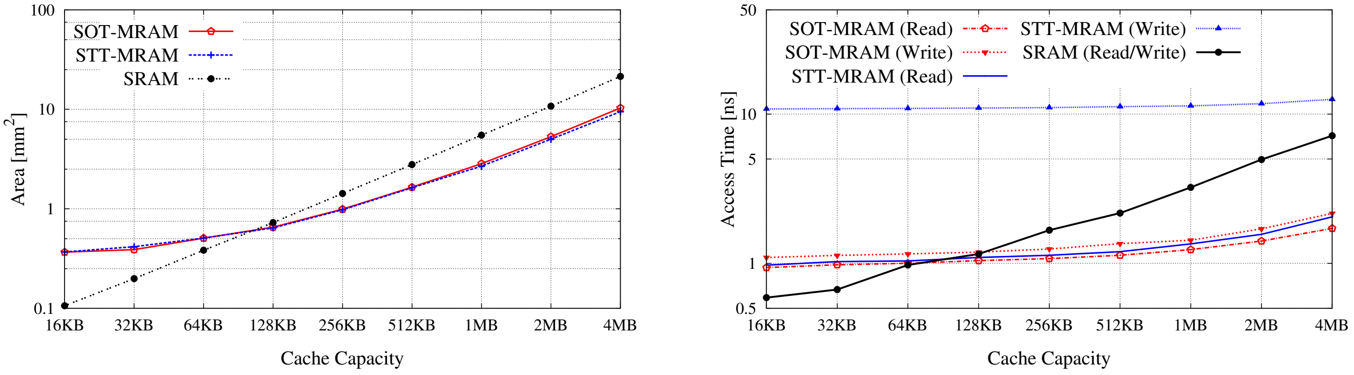
In terms of cost, STT-MRAM and hence also SOT-MRAM should reach a similar price per GBit compared to DRAM, when these technologies enter the mass production phase [45]. Consequently, both MRAM technologies will be cheaper than SRAM, although additional processing steps are required. However, these will be compensated by the higher density.

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), in-plane 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 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 and 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 for sizes of at least 16 KByte and thus it scales better.

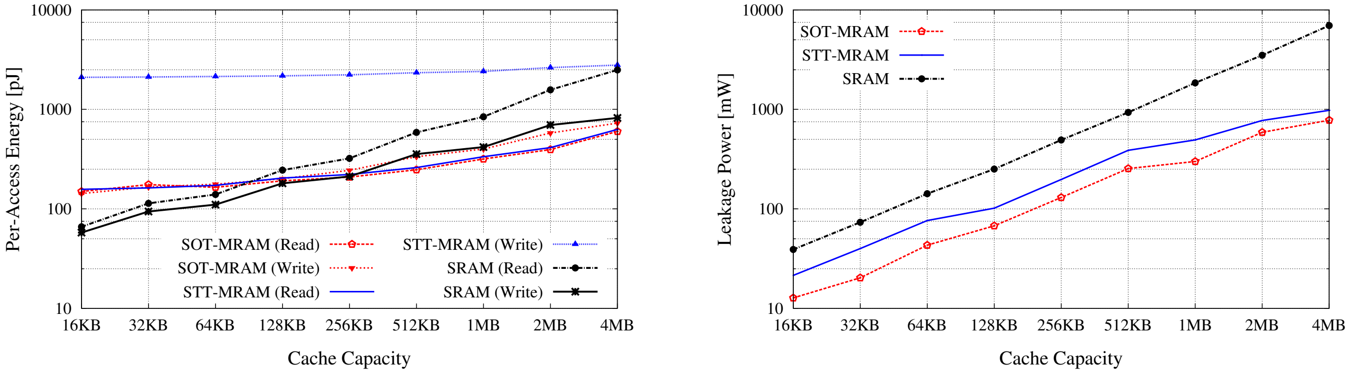
The second interesting aspect is the area comparison of SOT-MRAM and STT-MRAM. Although the access transistor and the periphery circuitry for SOT-MRAM are smaller than the counterparts for STT-MRAM (due to the lower currents), the overall area used by SOT-MRAM is slightly larger. This is due to the additional bit-cell terminal required by SOT-MRAM. In fact, this overhead depends on various aspects, e.g. design rules and size of the access transistor. However, as SOT-MRAM is not yet in production, it is not possible to exactly quantify the overhead. For an estimation, we decomposed the memory area into the bit-cell area (containing just the bit-cells and access transistors) and the remaining area (containing the decoders, read and write circuitry, etc.). While the latter is not affected by the additional terminal, the bit-cell area is. As the MTJ cells are placed in a different layer than the CMOS gates, each terminal requires a via. Since the metal pitch for the vias is the dominating aspect for the MTJ cell placement [46], the bit-cell area increases by 33 % due to the additional terminal (three vs. two). Using NVSim we obtained the ratio of the bit-cell area to the total memory area and found out that the bit-cell footprint is just 27 % of the overall memory area for



(a) Area

(b) Read & Write Latency

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



(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

a memory size of 512 KByte. Consequently, the additional terminal increases the total memory area by 9 %. Please note that this is just a very rough, first order approximation. In reality, the overhead can be considerably smaller.

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 asymmetric access behavior (almost) disappears.

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 [27].

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 access transistor and the periphery circuitry, that are designed for higher currents compared to SOT-MRAM, which are the reason why its leakage power is worse than that of SOT-MRAM. Note that in MRAM-based memories leakage power is only consumed by the periphery circuitry and access transistors, but not by the bit-cells themselves.

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-caches. 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.

IV. Evaluation of SOT-MRAM for Caches

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 the cache hierarchy.

A. Hybrid-Memory Evaluation Platform

Our evaluation uses gem5 [47], 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-MRAM we had to extend gem5 to support different read and write latencies for each cache.

Processor	Single-core @ 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/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
SPEC2000 applications	Bzip2, Equake, Gzip, MCF, VPR, Twolf
SPEC2006 applications	Hmmer, LBM, Sjeng

TABLE III
Configuration details for the experiments

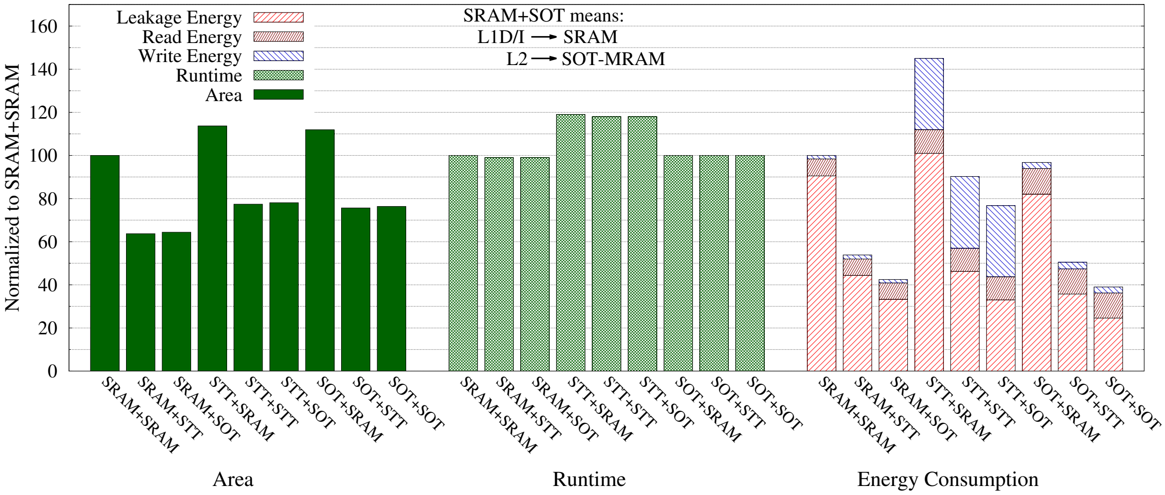


Fig. 7. Comparison of various cache configurations in terms of occupied area, average application runtime and average energy consumption (normalized to the standard configuration, i.e. SRAM for L1- and L2-cache)

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 employs a Harvard-architecture, which means it has separate L1-caches for instructions and data, respectively. Both have a capacity of 32 KByte and the 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 that the final performance does not only depend on the memory technology for each cache level, but also the clock frequency.

To evaluate the benefits and shortcomings of SOT-MRAM as cache memory, we use various workloads out of the MiBench benchmark suite [48] and SPEC2000 as well as SPEC2006 applications (see Table III) to show the system behavior under different workload conditions (small/large data sets, simple codes vs. complex codes). All MiBench workloads are simulated completely, including the initialization phase, to be as close as possible to the real world. Since it is not feasible to simulate an entire SPEC benchmark, only the first five billion instructions were simulated in case of the SPEC workloads. 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 as illustrated in Equation (1).

$$\begin{aligned} E_{total} &= E_l + E_w + E_r \\ &= P_l \cdot t_{runtime} + E_{WA} \cdot N_w + E_{RA} \cdot N_r \end{aligned} \quad (1)$$

Therefore, for each memory technology the per access energy (E_{WA} and E_{RA} for write and read, respectively) and leakage power P_l are taken into consideration. By considering also the number of read (write) accesses, N_r (N_w), and the runtime, the total energy consumption for every application can be estimated as shown in Figure 6.

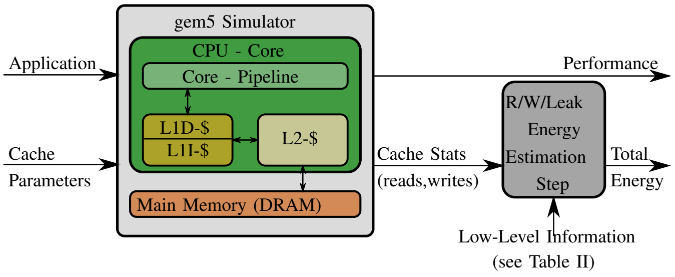


Fig. 6. Analysis flow to obtain performance & energy consumption for different cache configurations

B. Analysis of Performance, Area & Energy

The main results in terms of performance, area and energy consumption of our hybrid cache evaluation are summarized in Figure 7. For this figure[1] and all further discussions a configuration such as SRAM+SOT means that SRAM is used for both L1-caches and SOT-MRAM for the L2-cache.

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 is 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 36 %. As the L1-cache in our study is quite small, the phenomenon discussed in Section III-B3 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 about 11 %.

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-B3. However, the benefits strongly depend on the cache-level. For the L1-cache,

[1] Note that for Figure 7 all data was first normalized to the SRAM+SRAM configuration results and then the average values based on the normalized data were calculated. This way it is ensured that all benchmarks have similar influence on the average numbers reported here regardless of their runtime.

	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**	**(97)**	**59.8**	**(97)**	60.5	(99)	74.5	39.6	(53)	31.5	(42)	**29.5**	**(40)**
BitCount	130.1	**130.1**	**(100)**	**130.1**	**(100)**	130.1	(100)	146.8	76.0	(52)	58.6	(40)	**48.4**	**(33)**
CRC	998.8	**998.8**	**(100)**	**998.8**	**(100)**	1025.5	(103)	1175	631.9	(54)	498.5	(42)	**458.5**	**(39)**
Dijkstra	62.7	**62.4**	**(100)**	**62.4**	**(100)**	62.6	(100)	83.3	49.0	(59)	**40.6**	**(49)**	42.9	(52)
FFT	176.1	175.4	(100)	**175.3**	**(100)**	176.1	(100)	216.8	120.4	(56)	96.8	(45)	**93.7**	**(43)**
Patricia	49.1	46.7	(95)	**46.7**	**(95)**	47.6	(97)	60.6	31.6	(52)	25.3	(42)	**24.1**	**(40)**
QSort	35.2	34.9	(99)	**34.9**	**(99)**	34.9	(99)	40.0	20.9	(52)	16.0	(40)	**13.4**	**(33)**
SHA	23.3	**23.3**	**(100)**	**23.3**	**(100)**	23.3	(100)	29.4	16.8	(57)	**13.6**	**(46)**	13.7	(46)
StringSearch	1.5	**1.5**	**(100)**	**1.5**	**(100)**	1.5	(101)	1.9	1.0	(56)	0.8	(45)	**0.8**	**(45)**
Bzip2	2382	2369	(99)	**2369**	**(99)**	2372	(100)	2908	1605	(55)	1287	(44)	**1233**	**(42)**
Gzip	1826	1810	(99)	**1810**	**(99)**	1814	(99)	2211	1212	(55)	962	(44)	**905**	**(41)**
Twolf	5189	5130	(99)	**5126**	**(99)**	5148	(99)	5944	3086	(52)	2392	(40)	**2046**	**(34)**
VPR	3754	3709	(99)	**3706**	**(99)**	3715	(99)	4281	2204	(51)	1703	(40)	**1430**	**(33)**
Equake	5070	5017	(99)	**5015**	**(99)**	5025	(99)	5724	2927	(51)	2255	(39)	**1865**	**(33)**
MCF	3213	3174	(99)	**3172**	**(99)**	3179	(99)	3520	1749	(50)	1323	(38)	**1018**	**(29)**
Hmmer	1395	1370	(98)	**1369**	**(98)**	1378	(99)	1752	976	(56)	788	(45)	**781**	**(45)**
LBM	5303	5275	(99)	**5273**	**(99)**	5275	(99)	5909	3168	(54)	2301	(39)	**1834**	**(31)**
Sjeng	2219	2196	(99)	**2195**	**(99)**	2201	(99)	2733	1515	(55)	1209	(44)	**1160**	**(42)**
Average	100 %	99 %		**99 %**		100 %		100 %	54 %		42 %		**39 %**	

TABLE IV

Per benchmark analysis of different “hybrid” cache configuration (bold numbers represent the best value, numbers in () represent percentage values based on SRAM+SRAM)

SRAM offers on 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 % on average. Since the write access latency of the L2-cache has less effect on the overall system performance (lower write access rate, data is no written directly from the processor to the L2-cache as the L1-cache is not a write-through cache), even STT-MRAM can be used for this cache-level. However, for the L1-cache it is not feasible (i.e. runtime increase of 19 % on average).

***Energy**:* To analyze the energy consumption, let us first focus on the L2-cache. As Figure 7 shows, SRAM is not competitive 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 45 %, on average. Furthermore, as explained in Section III-B3 the leakage power of SOT-MRAM is also smaller than that of STT-MRAM. 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 35 % increase energy consumption compared to SRAM on 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. If both L1-caches (for data and instructions) employ SOT-MRAM instead of SRAM, the energy savings on average are 3 %.

***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 reduce accordingly. As a consequence, the per-access energy gains importance and in turn SOT-MRAM loses 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, the current SOT-MRAM implementation is not a suitable choice.

Please note that in a low-power processor running at a much lower clock frequency than the one used in this study, the access time differences between SRAM, SOT-MRAM and STT-MRAM are less significant. Therefore, the performance differences are also less pronounced.

1) 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 most situations the solution with the lowest energy consumption, regardless of the application runtime. Hence it 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 per-access energy for SRAM as L1-cache is lower than that of SOT-MRAM as memory technology for the L1-cache. As a consequence, SRAM is the better technology for the L1-cache, if the access energy is the main energy contributor, i.e. for high access rates. Whenever leakage energy is more important, for example due to a long runtime or low access rates, SOT-MRAM is the better choice.

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, Twolf or VPR) SRAM+SOT is faster, as STT-MRAM has much higher write access times. In terms of energy consumption SRAM+SOT is always 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 and area, the hybrid solutions offers the best trade-off for our processor configuration.

C. Using SOT-MRAM Advantages for Larger Caches

As we have shown in the previous parts, SOT-MRAM is a very promising memory technology for L2-caches, due to superior performance in combination with an extremely low energy consumption compared to SRAM. In addition an L2-cache using SOT-MRAM is also smaller than an SRAM-based L2-cache. In this section, we will demonstrate that this area advantage can be used to further improve the performance of the overall system, as illustrated in Figure 8.

For memory sizes larger than 256 KByte, SOT-MRAM offers an area advantage of roughly 2.3x compared to SRAM (see Figure 4(a)). Hence, if SOT-MRAM is used, one can double the cache size (1024 KByte instead of 512 KByte), and the 1024 KByte SOT-MRAM cache has a size comparable to the 512 KByte SRAM cache (2.8 mm^2 vs. 2.7 mm^2). In addition, the access latencies are almost the same for a 512 KByte and 1 MByte large cache based on SOT-MRAM (see Figure 4(b)). As a result, the performance with the 1024 KByte SOT-MRAM-based L2-cache is on average 6 % better than that achieved with a 512 KByte SRAM-based L2-cache. Especially, the larger SPEC benchmarks significantly benefit from the larger L2-cache (up 35 % improvement). Compared to an implementation with a 512 KByte SOT-MRAM L2-cache the performance benefit is 5 %, on average.

Beside performance, the increased L2-cache size has also an effect on energy. However, if SOT-MRAM is used for the L2-cache the disadvantage of the larger L2-cache is just 2 %, i.e. the energy savings compared to a processor with a 512 KByte SRAM-based L2-cache are still more than 55 %. In addition, for some benchmarks (e.g. Twolf, VPR), the energy consumption with the larger L2-cache is even lower than with the smaller L2-cache (both based on SOT-MRAM). This is due to the fact that in these situations the significant runtime advantage of the larger cache results in less leakage energy and hence in a lower overall energy consumption. If the larger cache does not improve the runtime, the energy increases by approximately 4 % (compared to an SOT-MRAM L2-cache with 512 KByte).

Of course, the same principal can be applied to higher cache-levels (e.g. L3-cache) as well. However, the performance advantages are typically less significant. We investigated this aspect using a SRAM+SOT+SRAM (32 KByte+512 KByte+64 MByte) and a SRAM+SOT+SOT (32 KByte+512 KByte+128 MByte) cache combination. On average, the last configuration can improve the performance by 2 % over the first setup. In contrast, the energy savings offered by SOT-MRAM increase considerably with higher cache-levels, due to the increase in memory capacity (see Fig. 5(a) and Fig. 5(b)).

D. SOT-MRAM for L1-Cache: I-cache vs. D-cache

A disadvantage of SOT-MRAM for small cache sizes are the longer read and write access latencies compared to SRAM. However, in a Harvard-architecture the access profiles for the instruction cache (I-cache) and the data cache (D-cache) are considerably different. For example, there are no write accesses from the CPU-side to the I-cache and the I-cache access patterns are usually much more regular. Hence, the effect of employing SOT-MRAM instead of SRAM for these two caches is considerably different.

Table V summarizes the comparison of SOT-MRAM and SRAM for different L1-caches. As expected, the D-cache is more sensitive to the access latency than the I-cache. Therefore, using SRAM for the D-cache is the better choice. However, for the I-cache SOT-MRAM can be used without affecting the performance. This combination, SOT-MRAM for the I-cache and SRAM for the D-cache, also helps to reduce the energy consumption. On average, the energy consumption is 7 % lower using this configuration compared to a standard SRAM implementation, at the cost of a small area increase (2.24 mm^2 vs. 2.05 mm^2). Nevertheless a pure SOT-MRAM-based configuration is on average the most energy efficient solution, but it is also the most area demanding approach (2.43 mm^2). However, for applications with a very high access rate (e.g. Dijkstra, SHA, Hmmer, Sjeng) SRAM for the D-cache is slightly better.

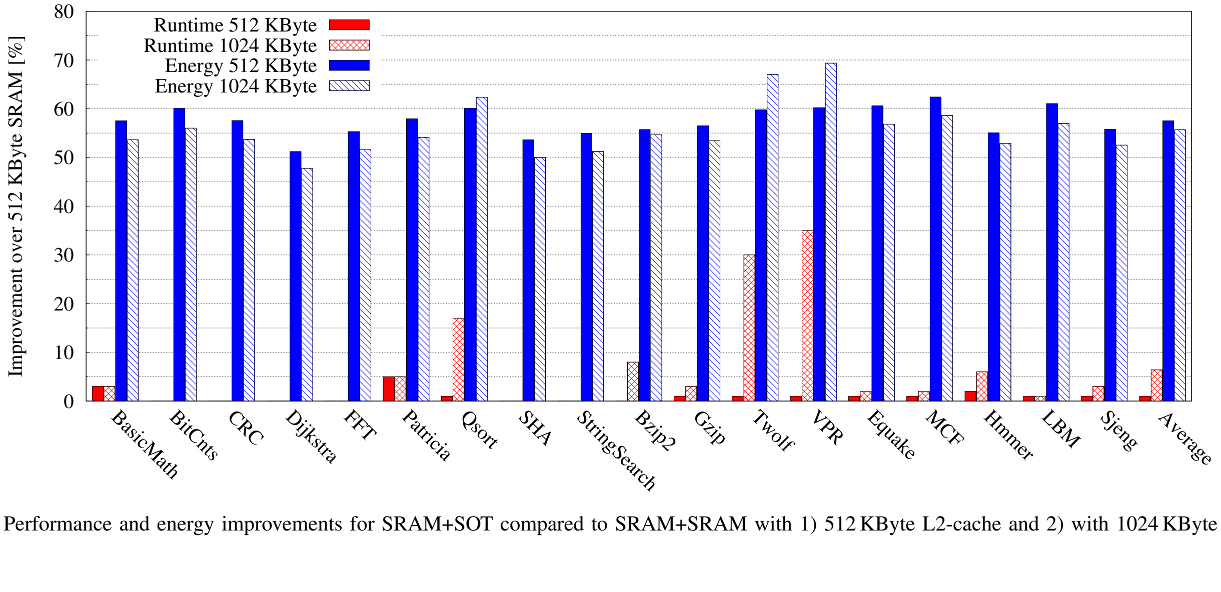


Fig. 8. Performance and energy improvements for SRAM+SOT compared to SRAM+SRAM with 1) 512 KByte L2-cache and 2) with 1024 KByte L2-cache

	Runtime [ms]							Energy [mJ]						
	DL1: SRAM IL1: SRAM	DL1: SOT IL1: SRAM		DL1: SRAM IL1: SOT		DL1: SOT IL1: SOT		DL1: SRAM IL1: SRAM	DL1: SOT IL1: SRAM		DL1: SRAM IL1: SOT		DL1: SOT IL1: SOT	
BasicMath	**59.8**	60.0	(100)	60.3	(101)	60.5	(101)	31.7	30.8	(97)	30.8	(97)	**29.8**	**(94)**
BitCnts	**130.1**	130.1	(100)	**130.1**	**(100)**	130.1	(100)	58.6	53.4	(91)	53.6	(91)	**48.4**	**(83)**
CRC	**998.8**	1025.5	(103)	**998.8**	**(100)**	1025.5	(103)	498.5	498.1	(100)	460.4	(92)	**458.5**	**(92)**
Dijkstra	**62.4**	62.6	(100)	**62.4**	**(100)**	62.6	(100)	40.6	44.5	(110)	**39.0**	**(96)**	42.9	(106)
FFT	**175.3**	175.9	(100)	175.6	(100)	176.1	(100)	96.9	96.5	(100)	94.3	(97)	**93.8**	**(97)**
Patricia	**46.7**	46.8	(100)	47.4	(102)	47.6	(102)	25.6	25.3	(99)	24.9	(97)	**24.6**	**(96)**
Qsort	**34.9**	34.9	(100)	**34.9**	**(100)**	34.9	(100)	16.0	14.9	(94)	14.5	(91)	**13.4**	**(84)**
SHA	**23.3**	23.3	(100)	**23.3**	**(100)**	23.3	(100)	13.6	14.0	(103)	**13.3**	**(97)**	13.7	(100)
StringSearch	**1.5**	1.5	(101)	**1.5**	**(100)**	1.5	(101)	0.8	0.9	(104)	0.8	(95)	**0.8**	**(99)**
Bzip2	**2370**	2372	(100)	**2370**	**(100)**	2372	(100)	1287	1320	(103)	**1200**	**(93)**	1233	(96)
Gzip	**1810**	1814	(100)	**1810**	**(100)**	1814	(100)	962	959	(100)	909	(94)	**905**	**(94)**
Twolf	**5126**	5146	(100)	**5126**	**(100)**	5148	(100)	2392	2274	(96)	2164	(90)	**2046**	**(86)**
VPR	**3706**	3715	(100)	**3706**	**(100)**	3715	(100)	1703	1586	(93)	1548	(91)	**1430**	**(84)**
Equake	**5015**	5024	(100)	**5015**	**(100)**	5025	(100)	2255	2095	(93)	2026	(90)	**1865**	**(83)**
MCF	**3172**	3179	(100)	**3172**	**(100)**	3179	(100)	1323	1174	(89)	1168	(88)	**1018**	**(77)**
Hmmer	**1369**	1378	(101)	**1369**	**(100)**	1378	(101)	788	833	(106)	**736**	**(93)**	782	(99)
LBM	**5273**	5276	(100)	**5273**	**(100)**	5276	(100)	2301	2080	(90)	2055	(89)	**1834**	**(80)**
Sjeng	**2195**	2197	(100)	2199	(100)	2202	(100)	1209	1215	(100)	**1155**	**(96)**	1160	(96)
Average	**100 %**	100 %		100 %		100 %		100 %	98 %		93 %		**91 %**	

TABLE V

COMPARISON OF SRAM AND SOT-MRAM FOR THE I-CACHE AND D-CACHE; EACH IS 32 KBYTE LARGE AND SOT-MRAM IS USED FOR THE 512 KBYTE LARGE L2-CACHE (NUMBERS IN () REPRESENT PERCENTAGE VALUES BASED ON SRAM/SRAM)

E. SOT-MRAM in Tag Arrays

So far, only the data part (set) of caches was studied. However, beside the set(s) there is also another important part: the tag array. In this section the implications of MRAM on these cache tag arrays will be discussed.

A cache tag array has always fewer entries than the cache itself, as there is only one tag entry for each cache block (here, 16 entries) in the cache sets. Hence, the tag array is also considerably smaller, has shorter read/write access latencies and also the read/write energy is smaller compared to the cache sets. In more detail, in our experiments, the L1 and L2 tag arrays have a size of 1 KB and 15 KB, respectively. For these tag arrays, Table VI summarizes the area, performance, and energy (extracted with the setup detailed in Section III). As it can be seen, SRAM has the best area, access latency, and dynamic energy while SOT-MRAM is superior in terms of leakage. However, the per-access energy plays a more important role for the tag arrays than it does for the cache itself. This is due to the fact, that the access rate for tag arrays is higher. Consequently, SOT-MRAM requires on average 25 % more energy for the L1-Data-tag than an SRAM-based solution. For the L1-Instruction-tag SOT-MRAM still requires 6 % more and even for the L2-tag it consumes 1 % more energy than SRAM, on average. Hence, it is best to employ SRAM for the tag arrays, as SRAM combines the smallest area and with the lowest overall energy consumption.

V. RELIABILITY ANALYSIS

In addition to performance, area and energy constraints, also reliability is a major design aspect. As already mentioned in Table II, SRAM is very susceptible to radiation-induced soft errors that can lead to bit flips in the bit-cells. In contrast, STT-MRAM as well as SOT-MRAM are less vulnerable to radiation, but suffer from retention failures due to an inherent thermal instability. In addition, STT-MRAM reliability is also impaired by read failures due to read disturb faults. In this section we compare the radiation-induced failure probability of an SRAM cache with the probability of failures in SOT-MRAM and STT-MRAM.

	L1 Tag = 1KB			L2 Tag = 15KB		
	STT	SOT	SRAM	STT	SOT	SRAM
Area [μm^2]	18896	21307	**8012**	118495	139485	**90030**
Read Latency [ns]	0.895	0.894	**0.421**	1.020	1.019	**0.951**
Write Latency [ns]	10.754	1.054	**0.421**	10.898	1.179	**0.951**
Read Energy [pJ]	5.52	4.97	**2.98**	34.91	33.30	**19.58**
Write Energy [pJ]	126.62	4.51	**0.99**	957.68	44.38	**5.86**
Leakage Power [mW]	2.22	**2.07**	2.96	18.27	**18.14**	33.43

TABLE VI

AREA, PERFORMANCE, AND ENERGY CHARACTERISTICS FOR L1 AND L2 TAGS; BOLD NUMBERS SHOW THE BEST OPTION

A. Soft Error Analysis in SRAM

When SRAM technology is employed for on-chip memory units, cache units are the major contributors to the overall *Soft Error Rate* (SER) of the microprocessor [49]. In order to estimate each cache unit SER, its *Architectural Vulnerability Factor* (AVF) is computed according to the life time analysis method presented in [50] and multiplied with the intrinsic alpha and neutron-induced FIT rates of a 65 nm technology SRAM cell working in terrestrial environment. The FIT rate is obtained from an industrial soft error analysis tool [51].

After obtaining the radiation-induced failure rate for each workload, the failure probability for one execution of that workload is computed according to:

$$F_{rad}(t) = 1 - e^{-\lambda t} \qquad (2)$$

where λ and t are the failure rate and workload execution time, respectively.

B. Retention Failure Analysis in MRAM

Memory units based on SOT-MRAM or STT-MRAM are less susceptible to radiation-induced soft errors, since radiation has no effect on the MTJ cells. In addition the size of the access transistor is rather large to supply a large enough current. Hence, only particles with a very high energy can influence the bit-cell state and this is only during write operations. Therefore, the probability of radiation-induced failures in MRAM is negligible.

However, MTJ cells suffer from an inherent thermal instability, which is an important source of unreliability for MRAM

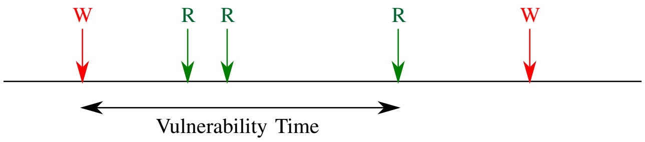


Fig. 9. Illustration of data vulnerability time in a cache memory (W = write, R = read)

technologies [19]. This thermal instability can lead to data loss and hence negatively affects the retention time of SOT-MRAM and STT-MRAM. To estimate the retention failure rate, F_{ret}, in MRAM we employ the probabilistic model presented in [19], which is similar to the radiation-induced failure probability of SRAM. According to this model

$$F_{ret} = 1 - e^{-\frac{n \cdot t_{vul}}{\tau_0 \cdot e^{\Delta}}} \tag{3}$$

where n is the number of bit-cells in the memory unit, t_{vul} is the average vulnerability time of a bit-cell, τ_0 is a constant equal to 1 ns and $\Delta \approx 60$ is the thermal stability factor, which is a function of the MTJ cell size. In this regard, as depicted in Figure 9, the vulnerability time of a bit-cell is the time period between a write operation and the last read access before the next write operation occurs.

Using the power series definition of the exponential function and the fact that in the L1-cache the vulnerability times are very small, Eq. (3) can be simplified to

$$F_{ret} = \frac{n \cdot t_{vul}}{\tau_0 \cdot e^{\Delta}} - \sum_{i=2}^{\infty} \frac{1}{i!} \left(\frac{-n \cdot t_{vul}}{\tau_0 \cdot e^{\Delta}} \right)^i \approx \frac{n \cdot t_{vul}}{\tau_0 \cdot e^{\Delta}}. \tag{4}$$

Hence the retention failure probability is exponentially dependent on Δ and has a linear dependence with the average vulnerability time of the bit-cells.

To extract the failure probability for various applications, for each workload the average vulnerability time t_{vul} was calculated with the lifetime analysis method employed for the soft error evaluation. Therefore, during the benchmark execution the vulnerability periods for all bits were obtained according to Figure 9 using the simulation platform described in Section IV. Afterwards, the failure probabilities were estimated using Equation (4).

C. Read Disturb Analysis in STT-MRAM

Beside retention failures also read disturb is a major source of unreliability in STT-MRAM, due to the shared read and write paths. Because of that, it can happen that the applied read current flips the bit-cell value accidentally [52]. However, since the read current is unidirectional, read disturb can happen only in one direction. In this work, read disturb can occur if and only if the bit-cell state is AP (i.e. ‘1‘ is stored). The model for the read disturb probability is given in [52]:

$$F_{rd} = 1 - e^{-\frac{t_{read}}{\tau_0 e^{\Delta(1 - I_{read}/I_{C0})}}} \tag{5}$$

Again $\Delta = 60$ is the thermal stability factor, t_{read} is the read period and I_{read} is the read current. Furthermore, I_{C0} is the critical current required to flip the bit-cell state from $AP \rightarrow P$.

D. Discussion

A comparison of the three different fault types for various cache configurations for a 65 nm technology node is shown in Figure 10 and for some selected configurations more details can be found in Table VII. In this regard, the main results can be summarized as follows:

- Read disturb faults in STT-MRAM are only a problem in the L1-cache, while they can be neglected for higher cache-levels. This is due to the fact that in average much more read operations per bit are performed to the L1-cache than to the L2-cache. In this regard, the L1-Data-cache is in average more susceptible than the L1-Instruction-cache, as most applications perform more read operations to the L1-Data-cache than to the L1-Instruction-cache.
- For all caches SRAM is the most vulnerable technology, due to its high soft error susceptibility. In particular, this applies for the L1-Data-cache and the L2-cache, while it is less severe for the L1-Instruction-cache. This is because the data lifetime (i.e. vulnerability time) in the L1-Instruction-cache is typically much shorter than in the other caches.
- Retention failures in MRAM are least likely compared to radiation-induced upsets in SRAM or read disturb faults in STT-MRAM. Accordingly, SOT-MRAM is the most reliable memory technology for caches in this comparison. As shown in Table VII, an SOT-MRAM-only solution has typically a 27 times smaller failure probability than an SRAM-only solution. If SOT-MRAM is used for the L2-cache only, the advantage over SRAM varies from application to application between 2 % and

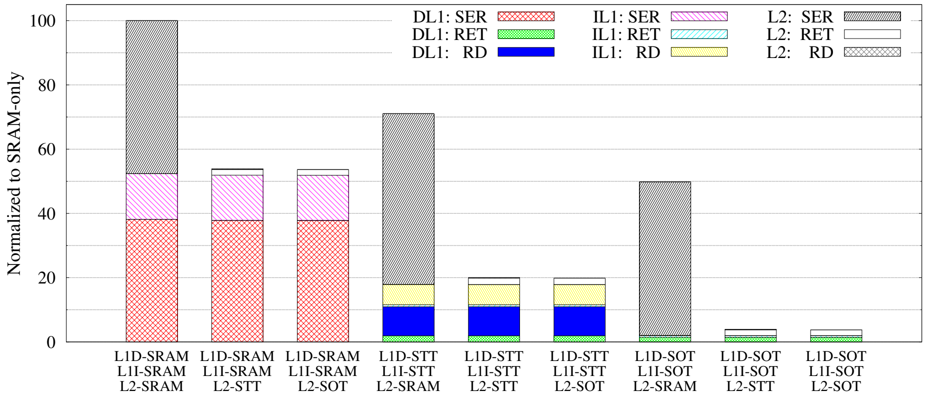


Fig. 10. Comparison of the average failure probabilities over all benchmarks for different hybrid cache configurations due to soft errors (SER), retention failures (RET) and read disturb (RD)

	DL1: SRAM + IL1: SRAM L2: SRAM	DL1: SRAM + IL1: SRAM L2: STT		DL1: SRAM + IL1: SRAM L2: SOT		DL1: SOT + IL1: SOT L2: SOT		DL1: SRAM + IL1: SOT L2: SOT	
BasicMath	2.618	1.703	(65.0)	1.693	(64.7)	0.097	(3.7)	1.459	(55.7)
BitCnts	0.849	0.731	(86.1)	0.731	(86.1)	0.032	(3.8)	0.347	(40.9)
CRC	21.605	21.255	(98.4)	21.255	(98.4)	0.836	(3.9)	20.297	(93.9)
Dijkstra	2.896	2.513	(86.8)	2.512	(86.7)	0.109	(3.8)	2.372	(81.9)
FFT	5.723	1.239	(21.6)	1.235	(21.6)	0.215	(3.8)	0.67	(5.9)
Patricia	1.012	0.509	(50.3)	0.496	(49.0)	0.037	(3.7)	0.301	(29.7)
Qsort	0.461	0.190	(41.2)	0.189	(41.0)	0.017	(3.7)	0.156	(33.8)
SHA	0.454	0.448	(98.7)	0.448	(98.7)	0.017	(3.7)	0.417	(91.9)
StringSearch	0.028	0.025	(89.3)	0.025	(89.3)	0.001	(3.6)	0.022	(78.6)
Bzip2	76.342	17.078	(22.4)	17.097	(22.4)	2.870	(3.8)	9.027	(11.8)
Gzip	74.313	22.299	(30.0)	22.207	(29.9)	2.783	(3.7)	16.178	(21.8)
Twolf	204.033	49.500	(24.3)	49.302	(24.2)	7.621	(3.7)	22.124	(10.8)
VPR	159.591	34.498	(21.6)	34.155	(21.4)	5.951	(3.7)	19.877	(12.5)
Equake	49.722	33.511	(67.4)	33.29	(67.0)	1.861	(3.7)	18.622	(37.5)
MCF	20.842	5.873	(28.2)	5.814	(27.9)	0.776	(3.7)	1.978	(9.5)
Hmmer	51.122	36.414	(71.2)	36.259	(70.9)	1.903	(3.7)	29.399	(57.5)
LBM	52.541	7.927	(15.1)	7.918	(15.1)	1.97	(3.7)	4.001	(7.6)
Sjeng	68.952	35.872	(52.0)	35.73	(51.8)	2.58	(3.7)	24.103	(35.0)
Average	100 %	53.9 %		53.7 %		3.7 %		40.1 %	

TABLE VII

PER BENCHMARK FAILURE PROBABILITIES FOR SELECTED HYBRID CACHE CONFIGURATIONS. NUMBERS IN () REPRESENT PERCENTAGE VALUES COMPARED TO SRAM BASELINE; FAILURE PROBABILITIES IN 10^{-12}

85 % depending on which cache-level is the dominating source of unreliability. As read disturb faults in the L2-cache have a very low probability, the main advantage of SOT-MRAM over STT-MRAM is for the L1-cache. Moreover, the probability of retention failures in SOT-MRAM is lower than in STT-MRAM, due to the fast execution times, if SOT-MRAM is used. This leads to vulnerability times (average and max) of SOT-MRAM that are comparable to SRAM, and thus significantly shorter compared to STT-MRAM.

In summary, SOT-MRAM provides the most reliable and least energy consuming solution. Although, its failure rate is not zero, it is much lower than that of STT-MRAM (due to read disturb susceptibility) or SRAM (due to radiation-induced soft errors). Therefore, if SOT-MRAM is employed, simpler ECC techniques to detect and correct errors can be used, which reduces the overall memory costs. For SOT-MRAM single-bit error correction is sufficient (lower error rate and random nature of retention failure), according to our results, while for SRAM multi-bit error correction is required to deal with multi-bit upsets in advanced technology nodes. This can be achieved by a combination of single error correction and an appropriate interleaving distance ($\geq$4) [53]. However, such a large interleaving distance might affect the memory aspect ratio and infer additional area and delay penalty for small memory arrays, hence, more sophisticated ECC schemes would be required.

Please note that beside read disturb and retention failures also write errors can impair the reliability of STT- and SOT-based memories, since the write behavior (switching of the magnetic orientation) is of stochastic nature. According to [54] the probability for a write error is given by:

$$F_{write} = e^{-\dfrac{t_{write}}{\tau_0 e^{\Delta(1-I_{write}/I_{C0})}}} \qquad (6)$$

Hence, I_{write}/I_{C0} is a driving factor, where I_{write} is the write current applied during the write period t_{write}. However, based on the physical data of our models, that current ratio is around 1.5 for both SOT-MRAM and STT-MRAM. As a result, the write error rate is negligible compared to the error rate due to read disturb or retention failure. Thus, it was not included in our analysis. Nevertheless, for ultra-low power memory devices, the write current will be significantly lower which can increase the write error rate such it has to be considered.

E. Impact of Technology Scaling

A problem of SOT-MRAM as well as STT-MRAM is the relation of the thermal stability factor Δ with the size of the MTJ cell. Since Δ has a linear relation with area [19], technology scaling tremendously weakens the thermal stability of the MTJ cells. In more detail, downscaling by one technology node reduces Δ by a factor of two. Hence, it is required to take counter measures in form of architectural techniques (e.g. ECC protection) or material improvements, in order to ensure the required reliability when SOT-MRAM is scaled down.

In contrast, technology scaling reduces the vulnerability of SRAM to radiation-induced soft errors. According to the radiation testing experiments results, as technology scales from 65 nm to 45 nm, the vulnerability to alpha and neutron-induced particles reduces by 20 %-30 %[2] [55, 56]. Furthermore, the FinFET technology, which is a promising solution for further downscaling, is 1.5-4X less vulnerable to soft errors [57]. Hence, even if Δ can be kept at 60 for newer technology nodes, the failure rate of SRAM will become comparable with MRAM-based technologies in two to three generations. Therefore, MRAM designers need to find ways to reduce the retention failure probability in future technology nodes to keep the failure rate at an acceptable level.

VI. CONCLUSION

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

[2] The probability of having a radiation-induced strike linearly decreases with the area reduction. However, in smaller nodes a particle has more chance to change the stored value due to smaller capacitance and operational supply voltage. These two competing factors, i.e. strike rate and error generation probability, determine the overall soft error vulnerability.

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 a 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-Data-cache (32 KByte) and SOT-MRAM for the L1-Instruction-cache as well as the larger L2-cache (512 KByte). Compared to an SRAM-only configuration this allows to reduce the energy consumption by 60 %, the area by 30 % and in addition the performance will increase by 1 %. These area gains can be used to double the size of the L2-cache to further improve the overall performance by 6 %, while the energy consumption is comparable to the system with the smaller L2-cache. Moreover, the retention failure probability of SOT-MRAM is 27x smaller than the probability of radiation-induced Soft Errors in SRAM, for a 65 nm technology node. All of these advantages make SOT-MRAM a viable choice for microprocessor caches.

References

[1] S. Wolf, D. Awschalom, R. Buhrman, J. Daughton, S. von Molnr, M. Roukes, A. Chtchelkanova, and D. Treger, "Spintronics: a spin-based electronics vision for the future," *Science*, vol. 294, no. 5546, pp. 1488–1495, Nov. 2001.

[2] S. Wolf, J. Lu, M. Stan, E. Chen, and D. Treger, "The Promise of Nanomagnetics and Spintronics for Future Logic and Universal Memory," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2155–2168, Dec. 2010.

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

[4] 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*, vol. 9, no. 2, pp. 13:1–13:35, May 2013.

[5] M. Chang, P. Rosenfeld, S. 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 *HPCA*, Feb. 2013, pp. 143–154.

[6] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano, "A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM," in *IEDM*, Dec. 2005, pp. 459–462.

[7] 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*, vol. 7, no. 8, pp. 501–507, 2013.

[8] 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 *DAC*, Jun. 2008, pp. 554–559.

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

[10] C.-H. Ho, G. Panagopoulos, S. Y. Kim, Y. Kim, D. Lee, and K. Roy, "A physical model to predict STT-MRAM performance degradation induced by TDDB," in *Device Research Conference*, June 2013, pp. 59–60.

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

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

[13] A. Raychowdhury, "Pulsed READ in spin transfer torque (STT) memory bitcell for lower READ disturb," in *Symposium on Nanoscale Architectures*, Jul. 2013, pp. 34–35.

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

[15] 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.

[16] M. Cubukcu, O. Boulle, M. Drouard, K. Garello, C. Onur Avci, I. Mihai Miron, J. Langer, B. Ocker, P. Gambardella, and G. Gaudin, "Spin-orbit torque magnetization switching of a three-terminal perpendicular magnetic tunnel junction," *Applied Physics Letters*, vol. 104, no. 4, pp. 1–5, 2014.

[17] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. Tahoori, "Architectural Aspects in Design and Analysis of SOT-based Memories," in *ASDPAC*, Jan. 2014, pp. 1–8.

[18] I. 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*, vol. 476, no. 7359, pp. 189–193, Aug. 2011.

[19] H. Naeimi, C. Augustine, A. Raychowdhury, S. Lu, and J. Tschanz, "STTRAM Scaling and Retention Failure," *Publisher Managing Editor Content Architect*, vol. 17, no. 1, pp. 54–75, May 2013.

[20] C. Xu, D. Niu, X. Zhu, S. Kang, M. Nowak, and Y. Xie, "Device-architecture co-optimization of STT-RAM based memory for low power embedded systems," in *ICCAD*, Nov. 2010, pp. 463–470.

[21] 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 Journal of Solid-State Circuits*, vol. 48, no. 2, pp. 598–610, Feb. 2013.

[22] G. Prenat, B. Dieny, G. Pendina, and K. Torki, "Hybrid CMOS/Magnetic Process Design Kit and application to the design of reliable and low-power non-volatile logic circuits," in *MEDIAN Workshop*, Jun 2012, pp. 13–16.

[23] W. Guo, G. Prenat, V. Javerliac, M. 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*, vol. 43, no. 21, p. 215001, May 2010.

[24] A. Nigam, C. Smullen, V. Mohan, E. Chen, S. Gurumurthi, and M. Stan, "Delivering on the promise of universal memory for spin-transfer torque RAM (STT-RAM)," in *ISLPED*, Aug. 2011, pp. 121–126.

[25] Y. Zhang, X. Wang, and Y. Chen, "STT-RAM cell design optimization for persistent and non-persistent error rate reduction: a statistical design view," in *ICCAD*, Nov. 2010, pp. 471–477.

[26] P. Wang, W. Zhang, R. Joshi, R. Kanj, and Y. Chen, "A thermal and process variation aware MTJ switching model and its applications in soft error analysis," in *ICCAD*, Nov. 2012, pp. 720–727.

[27] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. Tahoori, "Asynchronous asymmetrical write termination (AAWT) for a low power STT-MRAM," in *DATE*, Mar. 2014, pp. 1–6.

[28] D. Lee, S. Gupta, and K. Roy, "High-performance low-energy STT MRAM based on balanced write scheme," in *ISLPED*, Jul. 2012, pp. 9–14.

[29] T. Kishi, H. Yoda, T. Kai, T. Nagase, E. Kitagawa, M. Yoshikawa, K. Nishiyama, T. Daibou, M. Nagamine, M. Amano, S. Takahashi, M. Nakayama, N. Shimomura, H. Aikawa, S. Ikegawa, S. Yuasa, K. Yakushiji, H. Kubota, A. Fukushima, M. Oogane, T. Miyazaki, and K. Ando, "Lower-current and fast switching of a perpendicular TMR for high speed and high density spin-transfer-torque MRAM," in *EDM*, Dec. 2008, pp. 1–4.

[30] Y. Zhang, W. Zhao, G. Prenat, T. Devolder, J. Klein, C. Chappert, B. Dieny, and D. Ravelosona, "Electrical Modeling of Stochastic Spin Transfer Torque Writing in Magnetic Tunnel Junctions for Memory and Logic Applications," *IEEE Transactions on Magnetics*, vol. 49, no. 7, pp. 4375–4378, Jul. 2013.

[31] X. Zhu and J. Zhu, "Spin Torque and Field-Driven Perpendicular MRAM Designs Scalable to Multi-Gb/Chip Capacity," *IEEE Transactions on Magnetics*, vol. 42, no. 10, pp. 2739–2741, Oct. 2006.

[32] A. van den Brink, S. Cosemans, S. Cornelissen, M. Manfrini,

A. Vaysset, W. Van Roy, T. Min, H. Swagten, and B. Koopmans, "Spin-Hall-assisted magnetic random access memory," *Applied Physics Letters*, vol. 104, no. 1, pp. 1–3, 2014. [Online]. Available: http://scitation.aip.org/content/aip/journal/apl/104/1/10.1063/1.4858465

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

[34] N. Shibata, H. Maejima, K. Isobe, K. Iwasa, M. Nakagawa, M. Fujiu, T. Shimizu, M. Honma, S. Hoshi, T. Kawaai, K. Kanebako, S. Yoshikawa, H. Tabata, A. Inoue, T. Takahashi, T. Shano, Y. Komatsu, K. Nagaba, M. Kosakai, N. Motohashi, K. Kanazawa, K. Imamiya, H. Nakai, M. Lasser, M. Murin, A. Meir, A. Eyal, and M. Shlick, "A 70 nm 16 Gb 16-level-cell NAND flash memory," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 4, pp. 929–937, Apr. 2008.

[35] Y. Li and K. N., "NAND Flash memory: challenges and opportunities," *IEEE Computer*, vol. 46, no. 8, pp. 23–29, Aug. 2013.

[36] A. Khvalkovskiy, D. Apalkov, S. Watts, R. Chepulskii, R. Beach, A. Ong, X. Tang, A. Driskill-Smith, W. Butler, P. Visscher, D. Lottis, E. Chen, V. Nikitin, and M. Krounbi, "Basic principles of STT-MRAM cell operation in memory arrays," *Journal of Physics D: Applied Physics*, vol. 46, no. 7, pp. 74 001–74 020, Feb. 2013.

[37] D. Ralph and M. Stiles, "Spin transfer torques," *Journal of Magnetism and Magnetic Materials*, vol. 320, no. 7, pp. 1190–1216, Apr. 2008.

[38] P. Zhou, B. Zhao, J. Yang, and Y. Zhang, "A durable and energy efficient main memory using phase change memory technology," in *ISCA*, Jul. 2009, pp. 14–23.

[39] S. Raoux, G. Burr, M. Breitwisch, C. Rettner, Y. Chen, R. Shelby, M. Salinga, D. Krebs, S. Chen, H. Lung, and C. Lam, "Phase-change random access memory: a scalable technology," *IBM Journal of Research and Development*, vol. 52, no. 4.5, pp. 465–479, Jul. 2008.

[40] D. Jeong, R. Thomas, R. Katiyar, J. Scott, H. Kohlstedt, A. Petraru, and C. Hwang, "Emerging memories: resistive switching mechanisms and current status," *Reports on Progress in Physics*, vol. 75, no. 7, pp. 1–31, Jul. 2012.

[41] K. Garello, C. Onur Avci, I. Mihai Miron, O. Boulle, S. Auffret, P. Gambardella, and G. Gaudin, "Ultrafast magnetization switching by spin-orbit torques," *ArXiv e-prints*, October 2013.

[42] 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.

[43] X. Dong, C. Xu, Y. Xie, and N. Jouppi, "NVSIM: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE TCAD*, vol. 31, no. 7, pp. 994–1007, Jul. 2012.

[44] 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, pp. 6–7.

[45] A. Driskill-Smith, D. Apalkov, V. Nikitin, X. Tang, S. Watts, D. Lottis, K. Moon, A. Khvalkovskiy, R. Kawakami, X. Luo, A. Ong, E. Chen, and M. Krounbi, "Latest Advances and Roadmap for In-Plane and Perpendicular STT-RAM," in *International Memory Workshop*, May 2011, pp. 1–3.

[46] S. Gupta, S. Park, N. Mojumder, and K. Roy, "Layout-aware Optimization of STT MRAMs," in *DATE*. San Jose, CA, USA: EDA Consortium, 2012, pp. 1455–1458.

[47] 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, July 2006.

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

[49] G. Saggese, N. Wang, Z. Kalbarczyk, S. Patel, and R. Iyer, "An experimental study of soft errors in microprocessors," *IEEE Micro*, vol. 25, no. 6, pp. 30–39, Dec. 2005.

[50] A. Biswas, P. Racunas, R. Cheveresan, J. Emer, S. Mukherjee, and R. Rangan, "Computing architectural vulnerability factors for address-based structures," in *International Symposium on Computer Architecture*, Jun. 2005, pp. 532–543.

[51] E. Costenaro, D. Alexandrescu, K. Belhaddad, and M. Nicolaidis, "A Practical Approach to Single Event Transient Analysis for Highly Complex Design," *Journal of Electronic Testing*, pp. 301–315, Jun. 2013.

[52] Y. Huai, M. Pakala, Z. Diao, and Y. Ding, "Spin-transfer switching current distribution and reduction in magnetic tunneling junction-based structures," *IEEE Transactions on Magnetics*, vol. 41, no. 10, pp. 2621–2626, Oct. 2005.

[53] M. Ebrahimi, A. Evans, M. Tahoori, R. Seyyedi, E. Costenaro, and D. Alexandrescu, "Comprehensive Analysis of Alpha and Neutron Particle-induced Soft Errors in an Embedded Processor at Nanoscales," in *DATE*, 2014, pp. 1–6.

[54] E. Chen, D. Apalkov, Z. Diao, A. Driskill-Smith, D. Druist, D. Lottis, V. Nikitin, X. Tang, S. Watts, S. Wang, S. Wolf, A. Ghosh, J. Lu, S. Poon, M. Stan, W. Butler, S. Gupta, C. Mewes, T. Mewes, and P. Visscher, "Advances and Future Prospects of Spin-Transfer Torque Random Access Memory," *IEEE Transactions on Magnetics*, vol. 46, no. 6, pp. 1873–1878, June 2010.

[55] D. Alexandrescu, "A comprehensive soft error analysis methodology for SoCs/ASICs memory instances," in *International On-Line Testing Symposium*, Jul. 2011, pp. 175–176.

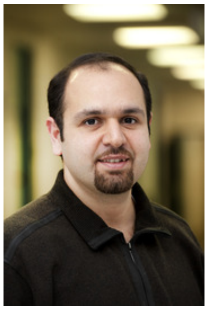
[56] E. Ibe, H. Taniguchi, Y. Yahagi, K. Shimbo, and T. Toba, "Impact of scaling on neutron-induced soft error in SRAMs from a 250 nm to a 22 nm design rule," *IEEE Transactions on Electron Devices*, vol. 57, no. 7, pp. 1527–1538, Jul. 2010.

[57] N. Seifert, B. Gill, S. Jahinuzzaman, J. Basile, V. Ambrose, Q. Shi, R. Allmon, and A. Bramnik, "Soft Error Susceptibilities of 22 nm Tri-Gate Devices," *IEEE Transactions on Nuclear Science*, vol. 59, no. 6, pp. 2666–2673, Dec. 2012.

Fabian Oboril received his diploma degree in mathematics techn. from the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany in 2010. Since 2010 he is a PhD student at the CDNC group of Prof. Tahoori at KIT. His research interests include the reliability issues of systems build in the nano era including transistor aging and low-power high-performance processors.

Rajendra Bishnoi received his MS degree from Manipal University, India in 2006 and wrote his thesis at Purple Vision Tech, Bangalore. He worked in Freescale as Design Engineer (from 2006 to 2012). In 2012 he joined KIT as PhD student in CDNC Group under the supervision of Prof. Tahoori.

Mojtaba Ebrahimi received his Master degree in Computer Engineering from Sharif university in 2010. He was a research assistant at Dependable System Laboratory of Sharif university until 2011. Since 2012, he is a PhD student at the CDNC group of Prof. Tahoori at KIT. His research is focused on the soft error rate estimation of microprocessors.



Mehdi B. Tahoori received his Ph.D. and M.S. in Electrical Engineering from Stanford University in 2003 and 2002, respectively, and B.S. in Computer Engineering from Sharif University of Technology, Tehran, Iran in 2000. He is a full professor and Chair of Dependable Nano-Computing (CDNC) at the Institute of Computer Science and Engineering (ITEC), Department of Computer Science, Karlsruhe Institute of Technology (KIT), Germany.