Power efficient Thermally Assisted Switching Magnetic memory based memory systems

Sophiane Senni, Lionel Torres, Gilles Sassatelli, Anastasiia Butko, Bruno Mussard

Power efficient Thermally Assisted Switching Magnetic memory based memory systems

Sophiane Senni[1,2], Lionel Torres[1], Gilles Sassatelli[1] and Anastasiia Bukto[1]
[1]LIRMM – UMR CNRS 5506 – University of Montpellier 2
Montpellier, France
{name}@lirmm.fr

Bruno Mussard[2]
[2]Crocus technology
Rousset, France
{ssenni, bmussard}@crocus-technology.com

***Abstract*—With the increasing size of the memory system inside today's chips, memories are becoming a critical part of the design of modern embedded systems. SRAM, DRAM and FLASH, respectively used for cache, working memory and non-volatile storage, are the three main memory technologies of current memory hierarchies. But all are facing to manufacturing constraints in the most advanced node, which compromises further evolution. Magnetic RAM (MRAM) technology is a very attractive alternative offering simultaneously reasonable performance and power consumption efficiency, high density and non-volatility. Among the MRAM technologies, while Toggle MRAM suffers from scalability issue and Spin Transfer Torque MRAM (STT-MRAM) is facing to data retention failure, Thermally Assisted Switching MRAM (TAS-MRAM) uses a scheme allowing a fully scalable bit cell, low power writing and excellent data retention. This paper demonstrates how features of TAS-MRAM can lead to power efficient memory systems. A case study of a TAS-MRAM-based L2 cache shows significant power saving while keeping reasonable performance.**

Keywords—MRAM, Thermally Assisted Switching, NVM, Memory hierarhy, VLSI, SoC, Embedded Systems

I. INTRODUCTION

Because it is the fastest memory technology, SRAM is currently chosen to design the upper level of cache memories in order to reach the best performance, particularly for multiprocessor architecture. With advanced technology nodes, SRAM is facing to leakage current which can lead to penalty in power consumption. DRAM occupies a lower level of the memory hierarchy as it is slower, but has higher density than SRAM. This technology is also power consuming due to its refresh policy to avoid data loss. Finally, we may find FLASH as the last level of the memory hierarchy, used for its high storage and non-volatility capabilities. To overcome performance and power issues of future embedded systems, some non-volatile memory technologies (NVMs) emerged in the past years. Magnetic RAM (MRAM), Resistive RAM (RRAM) and Phase-Change RAM (PCRAM) are seen by ITRS as the most promising candidates for a new era of low cost, low energy, high density and non-volatile embedded/mobile devices. While being non-volatile, MRAM combines good scalability, low leakage and radiation hardness [1]. For a same die footprint, MRAM can be used instead of SRAM to get about four times larger memory, which can lead to significant improvement of overall system performance and power consumption. Although MRAM is presenting a lot of attractive properties, two main issues are still under severe investigation: latency and dynamic energy. Compared to SRAM, MRAM write latency is around three to ten times higher, as well as MRAM write energy due to the high current needed to switch the bit cell.

In this paper, we explore integration of TAS-MRAM into the memory hierarchy of multiprocessor architecture. Both performance and energy are evaluated using a processor architecture simulator (GEM5) and a circuit-level model simulator for NVMs (NVSIM). We will demonstrate that using TAS-MRAM can be an attractive alternative to optimize overall system power consumption keeping reasonable performance.

II. MRAM BASICS

MRAM bit is a Magnetic Tunnel Junction (MTJ) which consists of two ferromagnetic layers separated by a thin insulating barrier. The information is stored as the magnetic orientation of one of the two layers, called the Free Layer (FL). The other layer, called the Reference Layer or Fixed Layer (RF), provides a fixed reference magnetic orientation required for reading and writing. The Tunnel Magnetoresistance (TMR) effect [2] causes the resistance of the MTJ to depend significantly on the relative orientation of the two magnetic layers: the antiparallel state gives a resistance much larger than in the parallel state. It enables the magnetic state of the FL to be sensed thanks to a current flowing through the MTJ. Hence, stored information can be read. In order to switch the orientation of the FL, three methods have been proposed: Toggle [3], Spin Transfer Torque (STT) [4] and Thermally Assisted Switching (TAS) [5].

A. Toggle MRAM

Fig. 1 illustrates a typical Toggle MRAM bit cell. This memory uses one transistor and one MTJ (1T-1MTJ) for each bit cell. The transistor provides the current flow through the

978-1-4799-5810-8/14/$31.00 ©2014 IEEE

MTJ needed for the read operation. Each MTJ is located at the intersection of two conductive lines. The Toggle write scheme consists of a specific current pulse sequence through the conductive lines in order to generate a magnetic field to switch the magnetic orientation of the FL to its opposite direction. Combining use of a free layer synthetic antiferromagnetic layer (SAF) with the current pulse sequence avoids the half-select problem for better selectivity. Limitation of this technology is its scalability. Reducing the size of the MTJ will lead to data retention issue. Besides, compared to other MRAM technologies, Toggle MRAM consumes a significant amount of current which does not shrink proportionally to the bit cell size [6].

B. STT-MRAM

Fig. 2 describes the write scheme of STT-MRAM. This technology uses the STT effect to switch magnetic orientation of the FL. A highly spin polarized current flowing into the MTJ provokes a "torque" applied by the injected electron spins on the magnetization of the FL. Hence, applying a sufficient current through the MTJ will causes sufficient torque to switch the bit cell, thus information can be written. Although STT effect allows STT-MRAM to use fewer current than Toggle MRAM, its main limitation on dimensional scaling will be posed by retention time failure. When a STT-MRAM cell is scaled, the thermal stability factor scales down linearly with the area, and can cause unreliability due to retention failure [7].

C. TAS-MRAM

Fig. 3 displays a complete TAS write operation. TAS method allows TAS-MRAM to combine very low power writability with excellent data retention by adding an antiferromagnetic layer in order to block the FL magnetic orientation under a threshold temperature. To switch the bit cell, a select transistor provides a current flow to heat the MTJ above the blocking temperature enabling storage of new information thanks to application of a magnetic field. Besides, heating the FL allows TAS-MRAM to use lower current than Toggle MRAM for writing the bit cell. This write scheme leads simultaneously to very high scalability, good thermal stability and low power programming.

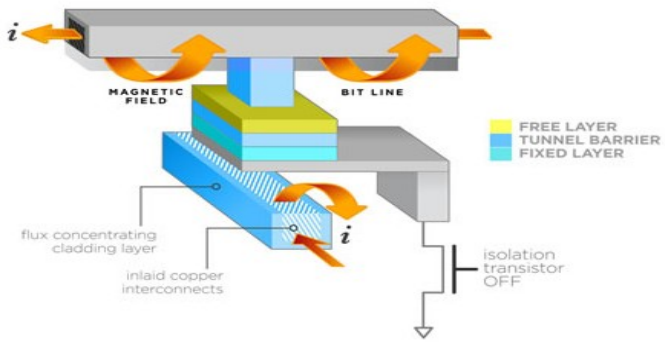

Fig. 1. Typical Toggle MRAM 1T-MTJ bit cell

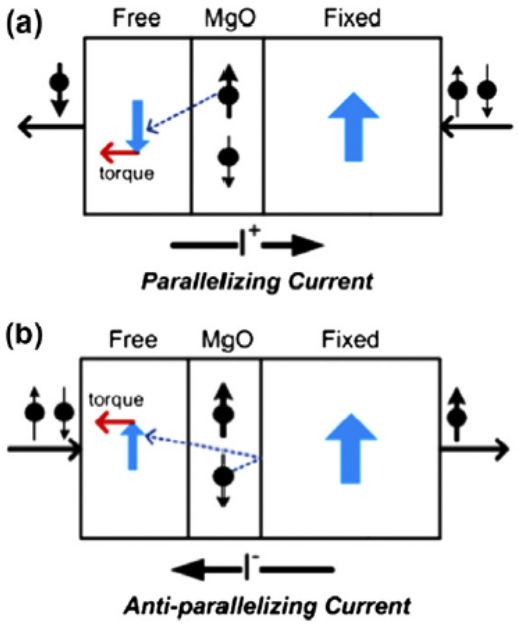

Fig. 2. STT effect of STT-MRAM

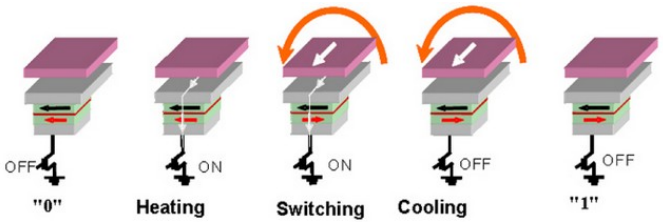

Fig. 3. TAS-MRAM principle

III. EXPLORATION FLOW

This section will present the high level simulation tools we are using to explore integration of MRAM technology into the memory hierarchy of processor architecture.

A. NVSIM Simulator

NVSIM [8], a modified environment of CACTI [9], is a circuit-level model for NVM performance, energy, and area estimation, which supports various NVM technologies, including STT-MRAM, PCRAM, RRAM, and legacy NAND Flash. It also includes the volatile SRAM memory. NVSIM is successfully validated against industrial NVM prototypes in [8], and it is expected to help boost architecture-level NVM-related studies. With NVSIM, we can estimate electrical features of a complete memory chip such as read/write access time, power consumption and so on, which can be used to calibrate a memory hierarchy of, for instance, a processor architecture simulator.

B. GEM5 Simulator

GEM5 [10] is a cycle accurate processor architecture simulator whose accuracy was validated against real hardware platform in [11]. It currently supports most commercial ISAs like ARM, ALPHA, MIPS, Power, SPARC and x86. The simulator's modularity allows these different ISAs to plug into the generic CPU models and the memory system without having to specialize one for the other. GEM5 can simulate a complete processor-based system with devices and operating system in full system mode and it supports also simulation of multi-core systems. The use of GEM5 allows us to define the overall processor system architecture, including the memory hierarchy specifications: cache size, L1/L2 cache and main memory latencies. Hence, we are able to extract execution time and all the memory transactions for a given application: number of L1/L2 read/write accesses, cache hits and misses, among other parameters.

C. Evaluation flow

Combining NVSIM with GEM5 allows us to evaluate different memory hierarchy strategies using SRAM and MRAM in order to find the best trade-off in terms of performance and power consumption. Memory hierarchy defined in GEM5 can be calibrated in access latency using simulation results of NVSIM. Fig. 4 depicts the exploration flow used for our study.

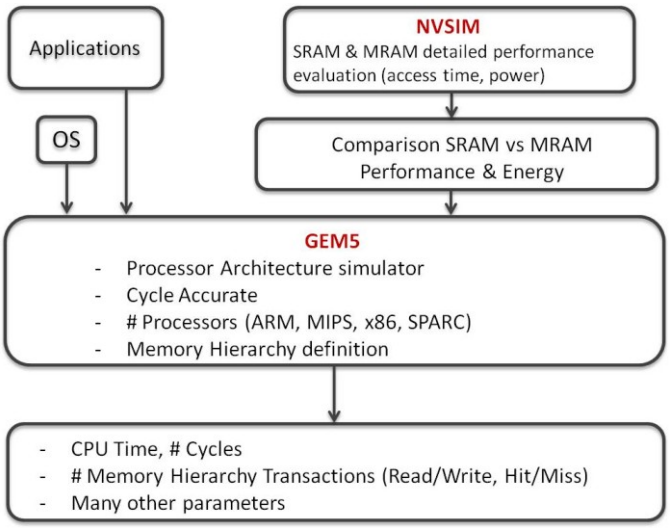

Fig. 4. MRAM-based memory hierarchy exploration flow

TABLE I. SPLASH-2 BENCHMARKS INPUT SETS

Benchmark	Input set
barnes	16K Particles, Timestep = 0.25, Tolerance 1.0
fmm	16K Particles, Timestep = 5
fft	2^{20} total complex data points
lu1	Contiguous blocks, 512x512 Matrix, Block = 16
lu2	Non-Contiguous blocks, 512x512 Matrix, Block = 16
ocean1	Contiguous partitions, 514x514 Grid
ocean2	Non-Contiguous partitions, 258x258 Grid
radix	4M Keys, Radix = 4K
water	15^3 Molecules, Timestep = 3

IV. EXPERIMENTAL SETUP

For our study, we propose to use some applications of SPLASH-2 benchmark suite [12], which are mostly in the area of High Performance Computing (HPC), to evaluate the impact of TAS-MRAM for shared L2 cache on two-core processor architecture. Table I gives details on input sets used for the benchmarks. We considered a 1GHz 32-bit RISC ARMv7 processor, with a complete Linux operating system running on top of it. We assume a two-level cache configuration: private 32kB L1 Instruction-cache (I-cache) 4-way associative, private 32kB L1 Data-cache (D-cache) 4-way associative, shared 512kB L2 cache 8-way associative. The main memory is a DDR3 type whose latency is fixed to 100 cycles. For our study, we integrate a model into NVSIM for the TAS-MRAM technology which was validated against industrial prototype in terms of performance, but not yet in terms of energy. Hence, we will use targeted performances of the industrial chip to evaluate the TAS-MRAM technology against SRAM.

V. PERFORMANCE EVALUATION

Performance comparison is made at node 120nm for SRAM and 130nm for TAS-MRAM, the main idea is to have a fair comparison for same technology node. First of all, we characterize each level of the SRAM-based memory hierarchy by simulation using NVSIM in order to calibrate latency parameters in GEM5. Table II describes performance of SRAM-based L2 cache, and targeted performance of TAS-MRAM-based L2 cache. TAS-MRAM write latency and hit latency are respectively about 8.5 times and 6 times higher than SRAM write and read latencies.

Fig. 5 shows the total execution time of several benchmarks of SPLASH-2 for two memory hierarchy schemes: a baseline scenario where all the cache memory hierarchy is implemented in SRAM ('SRAM') and a second scenario where L1 I-cache and L1 D-cache are in SRAM while L2 cache is based on TAS-MRAM ('TAS_MRAM'). Results are normalized to the execution time spent with the 'SRAM' scenario. Observing Fig. 5, we can notice performances of the two scenarios are quite similar for the simulated benchmarks. We can notice a performance penalty from 3%, for the lu1 application, to 28%, for the ocean1 and ocean2 applications.

Fig. 6 and Fig. 7 depicts respectively the average L2 cache miss rate evolution and the average L2 miss latency evolution at runtime for the barnes benchmark. Analyzing Fig. 6, we can notice that changing latencies does not affect the L2 cache miss rate behavior. It is not surprising since cache miss rate depends more on parameters related to size, such as cache size

and cache line size. In our study, these parameters are identical for the both scenarios. The small time gap between the curves comes from the difference on the execution time. Observing Fig. 7, we can examine the large L2 cache miss latency gap replacing SRAM with TAS-MRAM, which is due to the high latencies of the MRAM technology. At the end of the execution time, we can examine a ratio of two between the both scenarios for the average L2 cache miss latency. The small gap observed on the execution time concerning the barnes benchmark is explained by the L2 cache miss rate evolution. As the L2 cache miss reaches quickly enough a rate under 10%, the impact of large cache miss latency for TAS-MRAM-based L2 cache is not significant in terms of performance. For the other benchmarks, such as ocean and radix, the larger difference on execution time is justified by a higher average L2 cache miss rate compared to the barnes benchmark. However, since L2 cache is not the upper level of the memory hierarchy, performance of 'TAS_MRAM' scenario is not so far from the baseline scenario, e.g. 'SRAM'.

TABLE II. CACHE FEATURES

Field	512 kB L2 cache	
	SRAM	TAS-MRAM
Hit latency	5.95 ns	35 ns
Hit energy	1.05 nJ	1.96 nJ
Write latency	4.14 ns	35 ns
Write energy	0.08 nJ	4.62 nJ
Static power	82.23 mW	10 mW

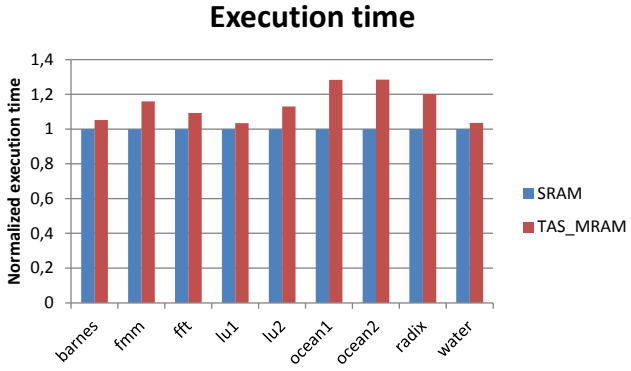

Fig. 5. Execution time (Normalized to execution time of 'SRAM' scenario)

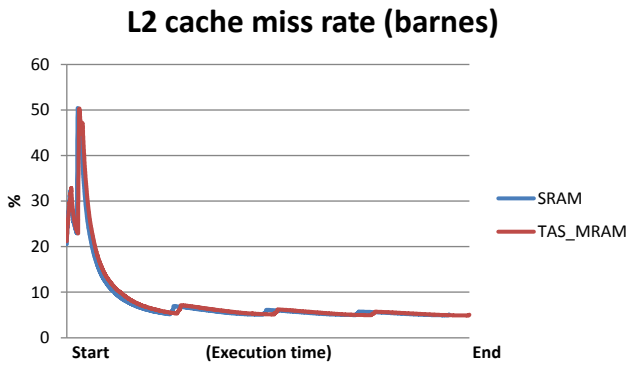

Fig. 6. Average L2 cache miss rate for barnes benchmark

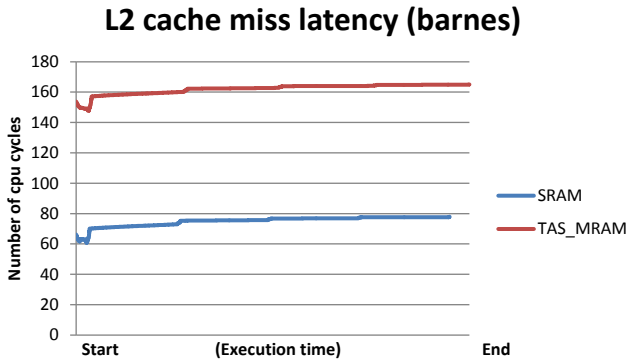

Fig. 7. Average L2 cache miss latency for barnes benchmark

VI. ENERGY EVALUATION

Table II describes energy consumption of SRAM and TAS-MRAM-based L2 cache. As expected, TAS-MRAM write access energy is much higher than SRAM whereas we have approximately a ratio of two for hit energy in favor of SRAM. But the considerable gain of TAS-MRAM over SRAM is on the leakage power: TAS-MRAM is more than 8 times less power consuming than SRAM. Indeed, most of the static power of memory systems comes from cell arrays. Because intrinsically non-volatile, TAS-MRAM cell has zero standby power, and the CMOS access transistor does not need to be power supplied. All static power for TAS-MRAM memory is due to peripheral circuitry such as address decoding, drivers and sense amplifiers.

Fig. 8 displays the total L2 dynamic energy. While total TAS-MRAM-based L2 read energy does not exceed twice of the total SRAM-based L2 read energy, total write energy is much higher using the MRAM technology due to its high write energy per access compared to SRAM. However, because L2 cache is much more accessed by read operations, the total TAS-MRAM-based L2 dynamic energy is about 2 to 4.5 times higher than the total L2 dynamic energy of our baseline. A maximum L2 read ratio of 94% of overall L2 accesses was noticed for the barnes application while the minimum L2 read ratio, observed for the lu2 benchmark, is

60%. Considering all the simulated benchmarks, we have an average L2 read ratio of around 77%.

Examining Fig. 9, we note the major benefit of using TAS-MRAM technology into memory systems. Simulation results show a gain over SRAM of more than 80% in terms of static power consumption for L2 cache. Total L2 cache power consumption, including dynamic and static energy, is shown in Fig. 10. Results demonstrate L2 cache memory can save 51% to 84% of power consumption using TAS-MRAM memory technology instead of SRAM. The large gap in leakage power between the two memories makes TAS-MRAM-based cache memory a very attractive alternative to save energy keeping overall application performance.

VII. Related Work

Several studies were made upon integration of MRAM into the memory hierarchy of processor architectures. Also, integration of MRAM into reconfigurable architectures such as FPGAs was evaluated. Evaluation of the benefit of 3D stacking ability of MRAM for 3D microprocessor was made in [13]. NUCA study with intra hybrid cache partitioned in regions of different memory technologies including MRAM was explored in [14] and [15]. Optimizations techniques such as early write termination which prevent unnecessary writes, or write buffers, to deal with high write latency and high write dynamic energy of MRAM were proposed in [16] and [17]. Trade-off between data retention and write latency of STT-MRAM were analyzed in [18]. In the field of programmable logic, investigations were made to evaluate performance and power consumption of a non-volatile FPGA using TAS-MRAM technology in [19], [20] and [21]. In our case, we explore the use of TAS-MRAM into the memory hierarchy of processor architectures while previous studies in this field have used the STT-MRAM technology. So far, TAS-MRAM was mostly examined for integration into reconfigurable architectures.

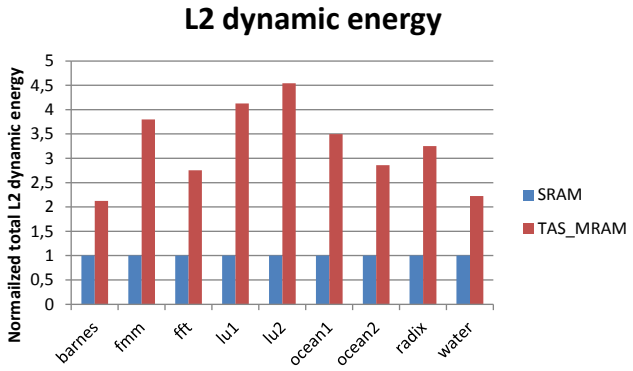

Fig. 8. Total L2 dynamic energy (Normalized to L2 dynamic energy of 'SRAM' scenario)

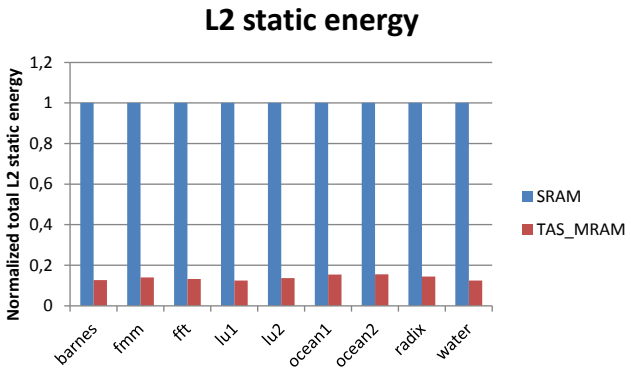

Fig. 9. Total L2 static energy (Normalized to L2 static energy of 'SRAM' scenario)

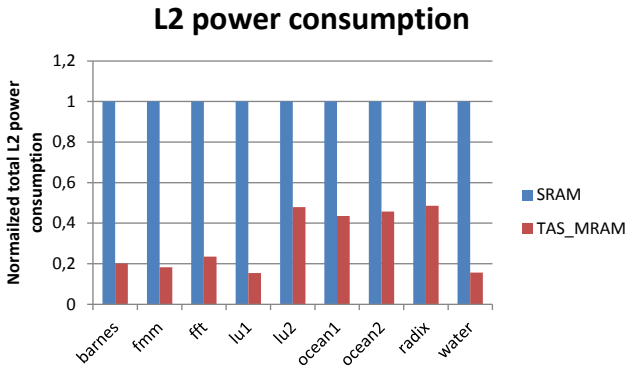

Fig. 10. Total L2 power consumption (Normalized to L2 power consumption of 'SRAM' scenario)

VIII. Future Work

In order to explore all benefits the TAS-MRAM can bring into future memory systems, we plan to develop a complete model of this memory technology into the NVSIM simulator. Hence, accurate performance, energy and area investigations will be possible. For instance, TAS-MRAM write scheme has ability to minimize power consumption using only two pulses of magnetic field to write an entire word [22]. Fig. 11 describes how power consumption can be optimized for TAS-MRAM write process.

IX. Conclusions

Among the emerging memory technologies, MRAM is a very promising candidate to help resolve one of the major challenges faced in continuing CMOS scaling: Power dissipation. Among MRAM technologies, TAS-MRAM is the most promising since it is the only memory combining simultaneously high scalability, excellent data retention and low power programming. Moreover, an innovative concept based on another implementation of TAS-MRAMs, called Magnetic Logic Unit (MLU) [23], allows one to introduce new functionalities such as the Match In Place (MIP) [23]. Fields of use of this new architecture are quite wide including

secure microcontrollers, SIM cards, banking cards, biometric authentication chips, near field communication and magnetic sensors. Integrating these attractive features, we believe that TAS-MRAM technology can lead to a new era of secure, low cost and low power non-volatile nanoelectronic systems.

ACKNOWLEDGMENT

The Authors wish to acknowledge all people from ADAC team at LIRMM and people from Crocus technology for their support in this work.

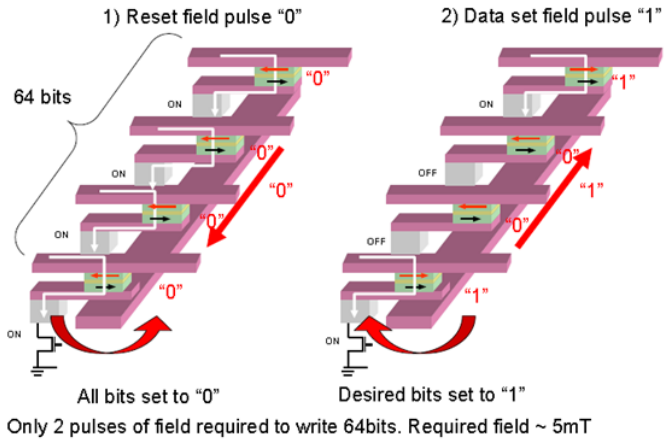

Fig. 11. Power consumption optimization by using only two pulses of magnetic field to write all the words. Energy of a heating pulse ~1pJ. Energy of a field pulse ~35pJ. In the first step, all bits are heated and reset to '0' with a '0' field pulse. In the second step, only the bit to be written '1' is heated together with the application of a '1' field pulse.

REFERENCES

[1] C. Hafer, M.V. Thun, M. Mundie, D. Bass and F. Sievert, "SEU, SET and SEFI Test Results of a Hardened 16Mbit MRAM Device," in Radiation Effects Data Workshop, 2012.

[2] J. Moodera, L. Kinder, T. Wong and R. Meservey, "Large magnetoresistance at room temperature in ferromagnetic thin film tunnel junctions," in Physical Review Letters, vol. 74, 1995.

[3] B.N. Engel et al., "A 4-Mb Toggle MRAM Based on a Novel Bit and Switching Method," in IEEE Transactions on Magnetics, vol. 41, no. 1, January 2005.

[4] A.V. Khvalkovskiy et al., « Basic principles of STT-MRAM cell operation in memory arrays," in Journal of Physics D: Applied Physics, vol. 46, no. 7, 2013.

[5] I.L. Prejbeanu et al., "Thermally assisted MRAM," in Journal of Physics: Condensed Matter, vol. 19, no. 16, 2007.

[6] K. Lewotsky, "Tech trends: Details on Everspin's ST-MRAM," in eetimes.com. Available online at http://www.eetimes.com/document.asp?doc_id=1280267.

[7] H. Naeimi, C. Augustine, A. Raychowdhury, S.L. Liu and J. Tschanz, "STTRAM Scaling and Retention Failure," in Intel Technology Journal, vol. 17, no. 1, 2013.

[8] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 31, no. 7, pp. 994-1007, Jul. 2012 .

[9] N. Muralimanohar, R. Balasubramonian, and N. P. Jouppi, "Cacti 6.0: A tool to model large caches," Published in International Symposium on Microarchitecture, Chicago, Dec 2007, Tech. Rep., Apr. 2009.

[10] N. Binkert, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish,M. D. Hill, D. A. Wood, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, and T. Krishna, "The gem5 simulator," ACM SIGARCH Computer Architecture News, vol. 39, no. 2, pp. 1–7, Aug. 2011.

[11] A. Butko, R. Garibotti, L. Ost, and G. Sassatelli, "Accuracy Evaluation of GEM5 Simulator System," in the proceedings of the 7th International Workshop on Reconfigurable Communication-Centric Systems-on-Chip (ReCoSoC). 2012, pp. 1–7.

[12] S.C. Woo, M. Ohara, E. Torrie, J.P. Singh and A. Gupta, "The SPLASH-2 Programs: Characterization and Methodological Considerations," in Proceedings of the 22nd Annual International Symposium on Computer Architecture, pp. 24–36, June 1995.

[13] 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'08: Proceesings of the 45th annual Design Automation Conference, pp. 554-559, New York, NY, USA, 2008, ACM.

[14] X. Wu, J. Li, L. Zhang, E. Speight, R. Rajamony, and Y. Xie, "Hybrid cache architecture with disparate memory technologies," in ACM SIGARCH Computer Architecture News, vol. 37, no. 3, 2009, pp. 34–45.

[15] X. Wu, J. Li, L. Zhang, E. Speight, R. Rajamony, and Y. Xie, "Power and Performance of Read-Write Aware Hybrid Caches with Non-volatile Memories," in Design Automation and Test in Europe (DATE), 2009.

[16] P. Zhou, B. Zhao, J. Yang, and Y. Zhang, "Energy reduction for stt-ram using early write termination," in International Conference on Computer-Aided Design, 2009, pp. 264-268.

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

[18] A. Jog el al., "Cache revive: architecting volatile STT-RAM caches for enhanced performance in CMPs," in 49th Annual Design Automation Conference, 2012, pp. 243–252.

[19] W. Zhao, E. Belhaire, B. Dieny, G. Prenat and C. Chappert, "TAS-MRAM based Non-volatile FPGA logic circuit," in International Conference on Field-Programmable Technology, 2007, pp. 153-160.

[20] Y. Guillemenet, L. Torres, G. Sassatelli, N. Bruchon and I. Hassoune, "A non-volatile run-time FPGA using thermally assisted switching MRAMs," in International Conference on Field Programmable Logic and Applications, 2008, pp. 421-426.

[21] W. Zhao, E. Belhaire, C. Chappert, B. Dieny and G. Prenat, "TAS-MRAM-Based Low-Power High-Speed Runtime Reconfiguration (RTR) FPGA," in ACM Transactions on Reconfigurable Technology and Systems (TRETS), vol. 2, no. 2, June 2009.

[22] I.L. Prejbeanu et al., "Thermally assisted MRAMs: ultimate scalability and logic functionnalities," in Journal of Physics D: Applied Physics, vol. 46, no. 7, 2013.

[23] B. Cambou, "Match In Place. A novel way to perform secure and fast user's authentication," available online at www.crocus-technology.com