Potential Applications Based on NVM Emerging Technologies

Sophiane Senni, Lionel Torres, Gilles Sassatelli, Abdoulaye Gamatié, Bruno Mussard

Potential Applications Based on NVM Emerging Technologies

Sophiane Senni[1,2], Raphael Martins Brum[1], Lionel Torres[1], Gilles Sassatelli[1] and Abdoulaye Gamatie[1]
LIRMM – UMR CNRS 5506 – University of Montpellier
Montpellier, France
{lastname[1]}@lirmm.fr

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

***Abstract*— Energy efficiency is a critical figure of merit for battery-powered applications. Today's embedded systems suffer from significant increase of power consumption essentially due to a high leakage current in advanced technology node. A significant portion of the total power consumption is spent into memory systems because of an increasing trend of embedded volatile memory area among the building components in System-on-Chips (SoCs). That is why new Non-Volatile Memory (NVM) technologies are considered as a potential solution to solve the energy efficiency issue. Among these NVM technologies, Magnetic RAM (MRAM) is a promising candidate to replace current memories since it combines non-volatility, high scalability, high density, low latency and low leakage. This paper explores use of MRAM into a memory hierarchy (from cache memory to register) of a processor-based system analyzing both performance and energy consumption.**

I. INTRODUCTION

Major issues encountered in ICs for advanced technology node include high leakage current, performance saturation, increased device variability and process complexity. For battery-powered applications, energy consumption is of course the most critical metric. In dynamic mode, fast switching at low power is targeted. In static mode, low leakage power is desired. Current systems embed volatile devices such as Flip-Flops, Static Random Access Memories (SRAM) and Dynamic Random Access memories (DRAM) which lose information when powered off. Circuit design techniques, such as clock and power gating, are currently used to reduce the power consumption during standby mode. A potential solution to overcome these energy challenges is non-volatile SoCs using non-volatile devices. Hence, a complete power-down is possible without losing data and logic states. A promising candidate for non-volatile SoCs is MRAM based on Magnetic Tunnel Junction (MTJ). Both academia and industry regard MRAM as a suitable technology to become a universal memory as it combines good scalability, low leakage, low access time and high density. Although MRAM is presenting a lot of attractive features, there are still two challenges under intensive investigation. First, MTJ switching requires a significant amount of current. Second, even if it is orders of magnitude faster than conventional NVMs, MTJ is slower than conventional SRAM, especially for write operation. 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.

This paper evaluates the performance and energy impacts of integrating MRAM into a memory hierarchy of processor architectures. An exploration on L2 cache, L1 cache and at register level is discussed. Useful information of the memory traffic are extracted to analyze accurately performance and energy consumption for several benchmarks.

II. MRAM BASICS

MRAM bit is a MTJ consisting of two ferromagnetic layers separated by a thin insulator. 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 (RF), provides a fixed reference magnetic orientation required for reading and writing. To switch the orientation of the FL, three methods have been proposed: Toggle [1], Spin Transfer Torque (STT) [2] and Thermally Assisted Switching (TAS) [3]. The Toggle scheme uses a specific current pulse sequence through the conductive lines to generate a magnetic field to switch the magnetic orientation of the FL to its opposite direction. STT-MRAM uses the spin transfer torque effect to switch magnetic orientation of the FL. A highly spin polarized current flowing through the MTJ induces a "torque" applied by the injected electron spins on the magnetization of the FL. TAS-MRAM adds 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.

III. NVM EXPLORATION FLOW

gem5 [4] is a processor architecture simulator widely used by the research community. It currently supports most commercial ISAs like ARM, ALPHA, MIPS, Power, SPARC and x86. gem5 is able to simulate a complete processor-based system with devices and operating system in full system mode. The use of gem5 allows defining the overall processor system architecture, including the memory hierarchy specifications: cache size, cache and main memory latencies etc. Execution time and memory transactions can be extracted for a given application, i.e. cache read/write accesses including cache hits and misses.

In order to calibrate the memory hierarchy defined in gem5 for access time, NVSIM [5] is used, a circuit-level model for NVM performance, energy, and area estimation, which supports various NVM technologies. Using NVSIM, a fast estimation of electrical features of a complete memory chip is possible comprising read/write access time and read/write access energy. However, if more accurate values are needed, SPICE simulation results of a design or electrical features of a real prototype can be easily used.

Combining electrical features of memories with gem5 allows evaluating different memory hierarchy strategies using different memory technologies in order to find a good trade-off between performance and energy consumption. Few studies upon integration of NVMs into the memory hierarchy of processor architectures were made in [6], [7] and [8] also using the gem5 simulator. Contrary to these investigations, we do not restrict the analysis to performance and energy results against a reference memory technology or architecture, but rather observe and analyze memory activity over time so as to better understand performance and energy issues.

IV. EXPERIMENTAL SETUP

As a case study, some applications of SPLASH-2 benchmark suite [9] are used, which are mostly in the area of High Performance Computing (HPC), to explore STT-MRAM and TAS-MRAM based caches for quad-core processor ARM architecture. Table I shows the architecture configuration and Table II gives details on the simulated benchmarks.

TABLE I. ARCHITECTURE CONFIGURATION

Hierarchy Level	Configuration
Processor	4-core, 1 GHz, 32-bit RISC ARMv7 (Linux OS)
L1 I/D cache	Private, 32kB, 4-way associative, 64B cache line
L2 cache	Shared, 512kB, 8-way associative, 64B cache line
Main memory	DRAM, DDR3, 100-cycle latency

TABLE II. SPLASH-2 BENCHMARKS

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

Cache latency parameters in gem5 are calibrated using simulation results of NVSIM for both SRAM and STT-MRAM while for TAS-MRAM, outcomes from a real prototype were used thanks to the support of Crocus Technology. To take into account the state-of-the-art of MRAM technology and to be fair for performance and energy evaluation, 45 nm STT-MRAM results are normalized to a baseline 45 nm SRAM cache, and 130 nm TAS-MRAM results are normalized to a baseline 120 nm SRAM cache.

V. L2 CACHE EXPLORATION

A. Performance Evaluation

Table III shows latencies for L2 caches implemented with the three considered memory technologies. As expected, both MRAM technologies have write latency higher than SRAM. Regarding hit latency, STT-MRAM (45 nm) is faster than SRAM (45 nm). It is not surprising since MRAM is denser than SRAM. As a result, for the same capacity, the total L2 cache area for STT-MRAM is smaller than for SRAM, which results in smaller bit line delay. This difference on hit latency in favor of STT-MRAM is noticeable only for large cache capacity. For TAS-MRAM, write and hit latencies are respectively about 8.5 and 6 times higher than SRAM (120 nm) write and hit latencies.

Fig. 1 shows the execution time of SPLASH-2 benchmarks for both STT-MRAM and TAS-MRAM based L2 caches. Observing Fig. 1, performance of STT-MRAM-based L2 scenario is similar and sometimes better than the baseline for the simulated benchmarks. It could be explain by a smaller hit latency for STT-MRAM compared to SRAM.

For TAS-MRAM-based L2, performance penalties from 3% (lu1) to 38% (ocean2) are observed. To better understand these results, we trace the L2 cache miss rate over time, displayed in Fig. 2. For ocean2 benchmark, a high L2 cache miss rate is observed, which explains the high penalty on execution time using TAS-MRAM. For other benchmarks, such as barnes, the small penalty on execution time is justified by a lower L2 cache miss rate compared to the ocean2 benchmark.

TABLE III. CACHE FEATURES

Field	32 kB L1 cache		512 kB L2 cache			
	SRAM (45 nm)	STT (45 nm)	SRAM (45 nm)	STT (45 nm)	SRAM (120 nm)	TAS (130 nm)
Hit Latency	1.25 ns	1.94 ns	4.28 ns	2.61 ns	5.95 ns	35 ns
Hit Energy	0.024 nJ	0.095 nJ	0.27 nJ	0.28 nJ	1.05 nJ	1.96 nJ
Write Latency	1.05 ns	5.94 ns	2.87 ns	6.25 ns	4.14 ns	35 ns
Write Energy	0.006 nJ	0.04 nJ	0.02 nJ	0.05 nJ	0.08 nJ	4.62 nJ
Static Power	22 mW	3.3 mW	320 mW	23 mW	82.23 mW	10 mW

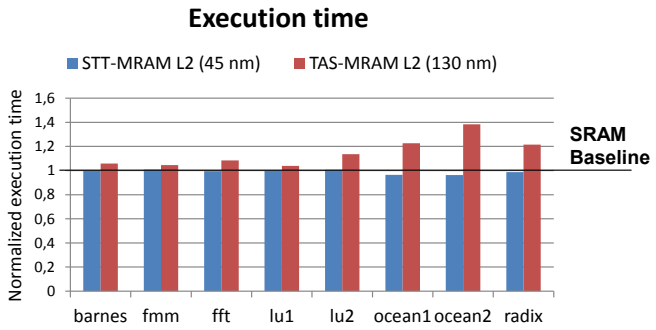

Fig. 1. Execution time of MRAM-based L2 cache

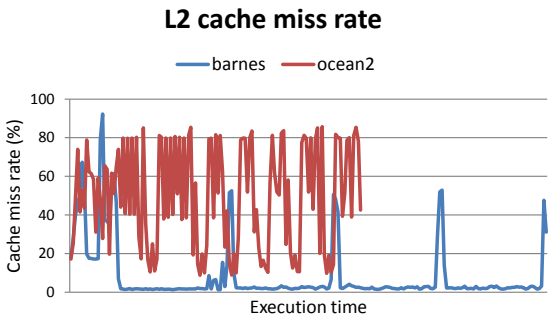

Fig. 2. L2 cache miss rate of barnes and ocean2 benchmarks

B. *Energy evaluation*

Table III gives energy consumption of L2 cache for the three considered technologies. As expected, write energy for both MRAM is higher compared to SRAM. While STT-MRAM hit energy is almost the same as SRAM hit energy. Considerable gain of MRAM over SRAM is however noticeable on the leakage power : 45 nm STT-MRAM-based L2 consumes over one order of magnitude less power than 45 nm SRAM-based L2 while TAS-MRAM is around 8 times less power consuming than 120 nm SRAM. Indeed, most of the static power of memories comes from cell arrays. Because MRAM cell has zero standby power and the CMOS access transistor does not need to be power supplied, all static power of MRAM-based memory is due to peripheral circuitry such as address decoding, drivers and sense amplifiers.

Fig. 3 displays the total L2 energy consumption (including dynamic and static energy). Simulation results show a gain over SRAM of more than 80% for both MRAM technologies in terms of static energy consumption regarding to L2 cache. This large gap in leakage power between MRAM and SRAM makes MRAM-based cache memory an attractive alternative to reduce energy while keeping reasonable performance.

To analyze more accurately energy consumption gain variation between the simulated benchmarks, the cache bandwidth evolution over time is traced in Fig. 4 to have a representation of the dynamic activity of the L2 cache. Analyzing the total L2 energy consumption for TAS-MRAM in Fig. 3, for instance a large gap between lu1 and lu2 benchmarks is observed. This energy difference is explained by the different memory activity in L2 cache between the two benchmarks. The higher the L2 cache bandwidth is, the higher the dynamic energy contribution is. In Fig. 4, since L2 read and write bandwidths for lu1 are significantly lower than those of lu2, total TAS-MRAM-based L2 energy consumption for lu1 is lower than for lu2.

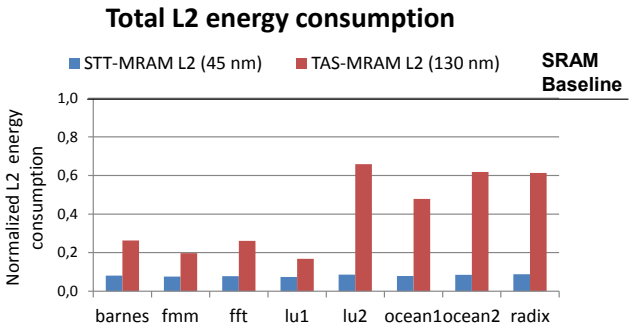

Fig. 3. MRAM-based L2 energy consumption

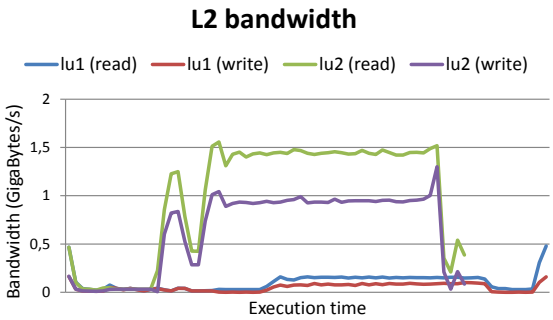

Fig. 4. L2 bandwidth for lu1 and lu2 benchmarks

VI. L1 CACHE EXPLORATION

A. *Performance evaluation*

As shown in the Table III for L1 cache, STT-MRAM read latency is quite similar to that of SRAM. Regarding the write operation, a higher latency is observed for STT-MRAM. In terms of CPU cycle, read latency is the same for both technologies (2 cycles). For write latency, SRAM takes 2 cycles while STT-MRAM needs 6 cycles. As a result, for a STT-MRAM-based L1 cache, the more the number of writes is high, the more the execution time penalty is significant.

Fig. 5 shows the execution time of different hierarchy strategies: a scenario where both I-Cache and D-Cache are based on STT-MRAM, a scenario with STT-MRAM-based I-Cache only and a scenario with STT-MRAM-based D-Cache only. For some benchmarks, using MRAM in L1 D-cache degrades overall performance due to high write latency. While for other benchmarks, such as radix, the execution time penalty is not so high, even with a MRAM-based D-Cache. To understand better these observations, Fig. 7 traces the write bandwidth evolution over time of fft and radix benchmarks. Analyzing the traces, L1 caches are more often accessed in write for fft, resulting in higher execution time penalty, than for radix. Integrating MRAM only into L1 I-Cache improves

the overall performance to be almost the same as the baseline scenario because I-Cache is read only. Globally, according to the simulation results, the execution time penalty does not exceed 21%. The read/write ratio observed for the simulated benchmarks shows the L1 cache is much more accessed by read operations.

B. Energy evaluation

Analyzing energy results of L1 cache in the Table III, STT-MRAM consumes around four and seven times more energy than SRAM respectively for read and write operations. Regarding the static power, a significant gain is observed replacing SRAM with STT-MRAM. Fig. 6 depicts the total L1 energy consumption (including L1 cache of each core). Replacing SRAM with MRAM in L1 cache does not lead to an energy gain as good as the gain observed for the L2 cache. Indeed, L1 cache is much more accessed than L2 cache. As a result, the dynamic energy impact of MRAM is clearly more visible. For the best cases, an energy gain of 62%, 34% and 35% is noted respectively for MRAM in both I-Cache and D-Cache, MRAM in I-Cache only and MRAM in D-Cache only.

Fig. 8 displays the L1 cache bandwidth over time for fmm and lu2 benchmarks. Traces are illustrated just for one core since the same behavior is noticed for the other cores. Regarding the I-Cache, the read bandwidth is higher for fmm than for lu2. This correlates with the energy consumption results in Fig. 6. For STT-MRAM-based I-Cache only, higher energy consumption is observed for fmm compared to lu2. On the other hand, for STT-MRAM-based D-Cache only, more energy is consumed for lu2 compared to fmm. This also correlates with the bandwidth traces in Fig. 8 since D-Cache read/write bandwidth is higher for lu2 compared to fmm.

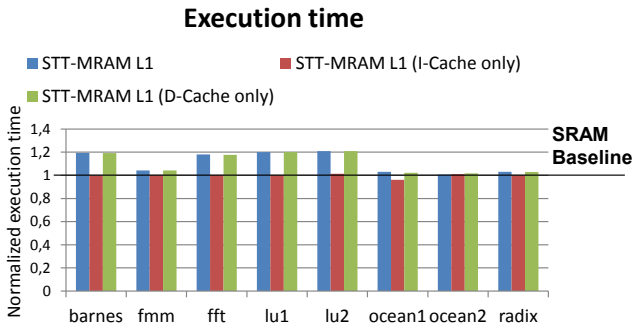

Fig. 5. Execution time of MRAM-based L1 cache

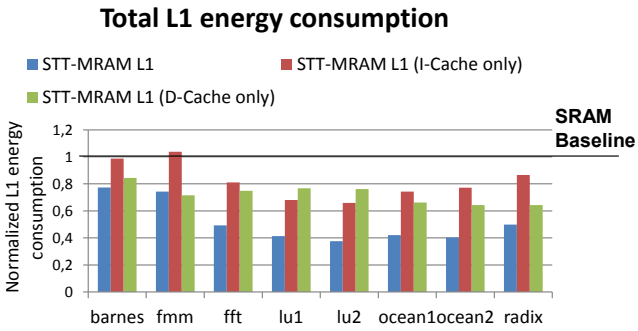

Fig. 6. MRAM-based L1 energy consumption

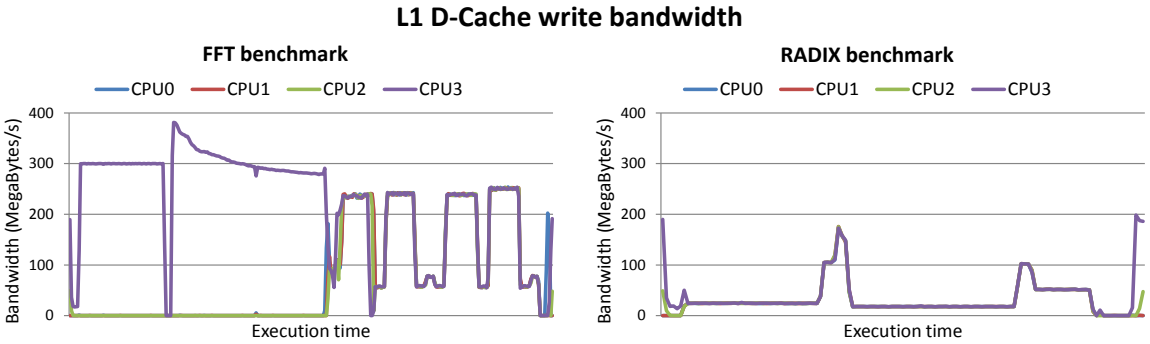

Fig. 7. L1 D-Cache write bandwidth for fft and radix benchmarks

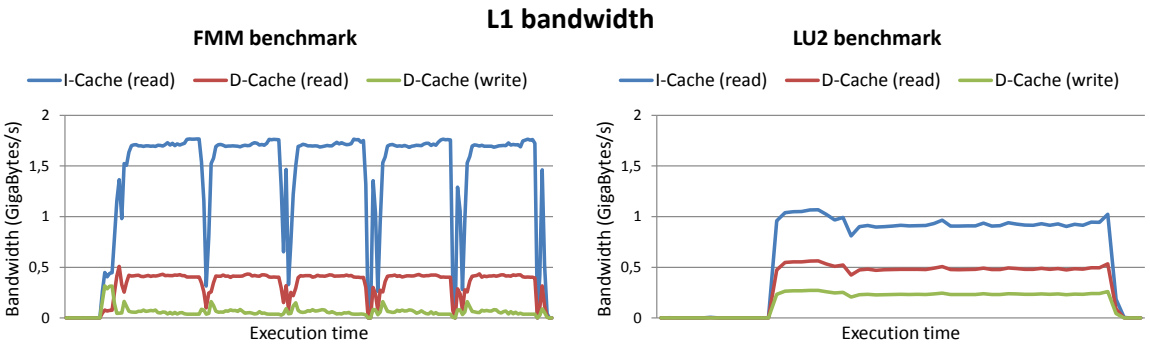

Fig. 8. L1 bandwidth for fmm and lu2 benchmarks (Since I-Cache is read only, the write bandwidth for this cache is not illustrated)

VII. NON-VOLATILE REGISTERS

Whereas most of the research in this field is focused on the development and the applications of non-volatile memory arrays, NVMs can also be used to build non-volatile sequential cells, such as latches and flip-flops. These cells can be later employed in the construction of sequential circuits, including single or multi-bit registers, which are found in any application-specific or general-purpose digital circuit. In this section, our objective is to demonstrate that the use of Non-Volatile Flip-Flop (NVFF) or register bank (part of the memory hierarchy) is realistic for the design of NVM processor architecture.

Na et al. [10] classified NVFFs in two different categories, according to their structure: separated latch/sense flip-flops (SLS) and merged latch/sense flip-flops (MLS). We investigated the first approach by designing and laying-out two different SLS-NVFFs using the ST 28nm FDSOI CMOS process, combined with a perpendicular magnetic anisotropy (PMA) STT-MTJ post-process developed by CEA/Spintec. In this technology, the magnetic junctions are cylindrical nanopillars having a nominal radius of 200 nm. Due to limitations of the post-process, the required spin-polarized current is in the order of 1 mA. State-of-the-art PMA-STT processes require currents as low as 50 µA [11], though.

A. *Non-volatile Flip-Flop Structure*

Both flip-flops follows the architecture depicted in Fig. 9a, composed of standard CMOS master and slave latches, as well as of the MTJ-specific read and write circuits. The master latch input samples data from the external signal *ff_d* or from the read circuit, which translates the MTJ configuration to a valid logic level.

The sense or read circuit is based on the self-referenced sense amplifiers known as Black & Das (Fig. 9b) [12] and Pre-Charged Sense Amplifier, or PCSA (Fig. 9c) [13]. Both implementations require two complementary configured MTJs for each bit. Write circuits (not shown) are based on paired CMOS tri-state buffers, whose sizing is proportional to the write current. Given this and the particular requirements of the magnetic post-process, a hybrid flip-flop is 8 to 9 times larger than a standard flip-flop. In turn, processes requiring lower write currents will present a substantially lower area overhead.

B. *Energy profile*

As the magnetic post-process model does not include process variation data, we devised nine simulation corners (shown in Table IV), allowing us to capture the behavior of the implemented cells under the variation of MTJ dimensions. These corners were combined with ST statistical models using Monte-Carlo simulations to produce the energy profiles shown in Fig. 10.

The Black & Das flip-flop presents peak read currents in the order of 150 µA. Its contender, on the other hand, may consume as much as 1 mA, as depicted in Fig. 10a. Write current, shown in Fig. 10b, is calibrated to meet the technology requirements (1 mA per MTJ). However, variations on the MTJ geometry have a strong impact on the resulting write current, explaining the variation seen in the graph.

Leakage current (Fig. 10c) is in the order of 10 nA for the PCSA-based cell (SLS_NVFF1) a small overhead when compared to the standard CMOS FF. The SLS_NVFF0 cell, however, do not remain disconnected while idle as the PCSA does. Its leakage current is thus in the order of 200 nA, a 20x increase when compared to SLS_NVFF1.

C. *Building non-volatile registers from NVFFs*

NVFFs enable the introduction of new features to processor architectures, such as instant on/off capabilities. This technique consists of replacing standard volatile registers with non-volatile counterparts, made of sets of NVFFs. Upon an external command, these registers can save their volatile state into the non-volatile part, creating a checkpoint. Restoring the state after a system shutdown is a matter of reading the values stored in the MTJs.

Koike et al. [14] implemented a complete MIPS-based processor using this approach. However, their implementation is based on the MLS-NVFF architecture, composed of a volatile master latch and a non-volatile slave latch. Whereas this structure is more compact, SLS-NVFFs, such as the ones presented in this paper, have the advantage of operating normally over the fast, volatile context, while reading and writing from the non-volatile context only when needed. For a more complete performance comparison of read circuits, refer to [15].

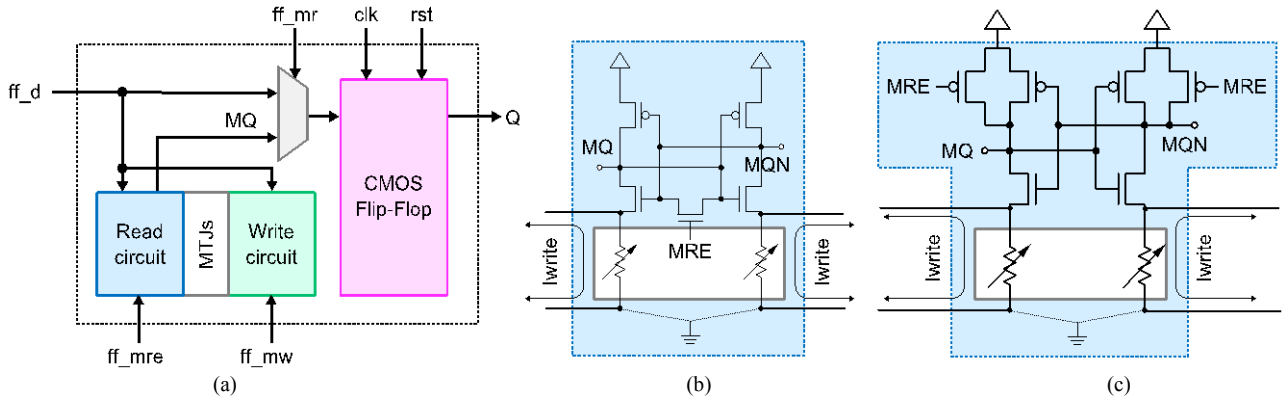

Fig. 9. SLS-NVFF Architecture: (a) block diagram. (b) Detail of SLS-NVFF0 read circuit, based on the Black & Das cell. (c) Detail of SLS-NVFF1 read circuit, based on the Pre-Charged Sense Amplifier (PCSA) cell.

TABLE IV. MTJ CORNERS USED FOR RELIABILITY SIMULATION

Dimension	SL	ST	TS	SS	TT	LL	TL	LT	LS
Left MTJ radius (nm)	140	140	200	140	200	260	200	260	260
Right MTJ radius (nm)	260	200	140	140	200	260	260	200	140

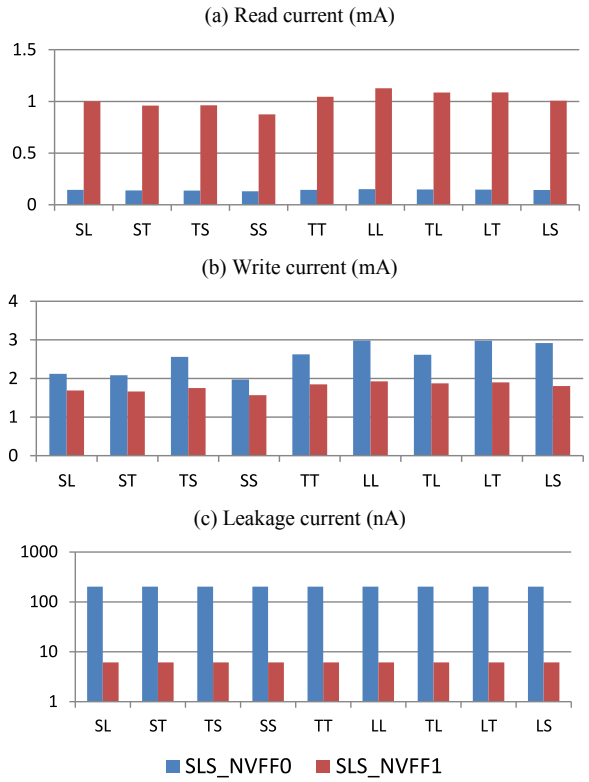

Fig. 10. Energy profile of SLS_NVFF0 and SLS_NVFF1 hybrid flip-flops.

VIII. CONCLUSIONS

This paper explored NVM-based memory hierarchy from cache memory to register. Performance and energy analysis were made capturing useful information of the memory traffic. Simulation results demonstrate it is possible to reduce significantly the total energy consumption of caches thanks to the low leakage power of MRAM. Regarding NVFFs, NVM emerging technologies can lead to a new era of non-volatile SoCs with the ability to retain stored data when powered off. In such systems, power supply can be cut every time there is no need to operate: this is the normally-off computing. Among NVMs, MRAM is a promising candidate. Currently, STT-MRAM shows the best results in terms of speed. However, it still suffers from reliability issues for advanced technology nodes. Although it is slower, TAS-MRAM demonstrates excellent data retention leading to high reliability. Moreover, TAS-MRAM can be designed to become a storage element and a logic Exclusive-OR gate at the same time: this is the Magnetic Logic Unit (MLU) [16]. In addition to non-volatility, high reliability and low leakage power, MLU can lead to highly secured devices, an important issue with the emergence of the Internet of Things.

ACKNOWLEDGMENT

The Authors wish to acknowledge ADAC team at LIRMM, Crocus technology and the French National Research (ANR), through the DIPMEM project, for their support in this work.

REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[5] 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

[6] J. Wang, X. Dong and Y. Xie, “OAP: an obstruction-aware cache management policy for STT-RAM last-level caches,” in Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 847-852, March 2013.

[7] R. Bishnoi, M. Ebrahimi, F. Oboril and M. Tahoori, “Architectural Aspects in Design and Analysis of SOT-Based Memories,” in the 19th Asia and South Pacific Design Automation Conference (ASP-DAC), pp. 700-707, January 2014.

[8] E. Arima et al., “Fine-Grain Power-Gating on STT-MRAM Peripheral Circuits with Locality-aware Access Control,” in The Memory Forum (in conjounction with the 41st International Symposium on Computer Architecture), June 2014, unpublished.

[9] 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.

[10] T. Na, K. Ryu, J. Kim, S.H. Kang, and S. Jung., “A Comparative Study of STT-MTJ based Non-Volatile Flip-Flops,” in Circuits and Systems (ISCAS), IEEE International Symposium on, pages 109–112, May 2013.

[11] D. Suzuki, M. Natsui, A. Mochizuki et al., “Design and fabrication of a perpendicular MTJ based nonvolatile programmable switch achieving 40% less area using shared-control transistor structure,” Journal of Applied Physics, 115, 17B742 (2014).

[12] W. C. Black and B. Das, “Programmable logic using giant-magnetoresistance and spin-dependent tunneling devices,” J. Applied Physics, 87(9):6674–6679, 2000.

[13] W. Zhao, C. Chappert, V. Javerliac, and J-P Noziere, “High speed, high stability and low power sensing amplifier for mtj/cmos hybrid logic circuits,” Magnetics, IEEE Transactions on, 2009.

[14] H. Koike et al., “A power-gated MPU with 3-microsecond entry/exit delay using MTJ-based nonvolatile flip-flop,” Solid-State Circuits Conference (A-SSCC), IEEE Asian , pp.317-320, Nov. 2013

[15] B. Jovanovic, R. M. Brum and L. Torres, “Comparative Analysis of MTJ/CMOS Hybrid Cells based on TAS and In-plane STT Magnetic Tunnel Junctions,” Magnetics, IEEE Transactions on, vol.PP, no.99, pp.1,1

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