Using Multifunctional Standardized Stack as Universal Spintronic Technology for IoT

M. Tahoori[1], S.M. Nair[1], R. Bishnoi[1], S. Senni[2], J. Mohdad[2], F. Mailly[2], L. Torres[2], P. Benoit[2], A. Gamatie[2], P. Nouet[2], F. Ouattara[2], G. Sassatelli[2], K. Jabeur[3], P. Vanhauwaert[3], A. Atitoaie[4], I.Firastrau[4], G. Di Pendina[3] and G. Prenat[3]

[1)]*Karlsruhe Institute of Technology, Germany*

[2)]*LIRMM, UMR CNRS 5506, University of Montpellier, France*

[3)]*Univ. Grenoble Alpes, CNRS, CEA, INAC-SPINTEC, F-38000 Grenoble, France*

[4)] *Transilvania University of Brasov, 29 B-dul Eroilor, 500036 Brasov, Romania*

***Abstract*— For monolithic heterogeneous integration, fast yet low-power processing and storage, and high integration density, the objective of the EU GREAT project is to co-integrate multiple digital and analog functions together within CMOS by adapting the Magnetic Tunneling Junctions (MTJs) into a single baseline technology enabling logic, memory, and analog functions, particularly for Internet of Things (IoT) platforms. This will lead to a unique STT-MTJ cell technology called Multifunctional Standardized Stack (MSS). This paper presents the progress in the project from the technology, compact modeling, process design kit, standard cells, as well as memory and system level design evaluation and exploration. The proposed technology and toolsets are giant leaps towards heterogeneous integrated technology and architectures for IoT.**

I. Introduction

Billions of Smart Connected Devices are sold every year with an increase of both the number and complexity for these systems. The interest for developing smart connected systems (Smart Sensors, Secure Elements, etc.) based on an "*Internet of Things*" (IoT) is growing fast. The number of Internet-connected devices surpassed the number of human beings on the planet in 2011, and by 2020, Internet-connected devices are expected to number between 26 billion and 50 billion. New applications in IoT will require storing and computing ever-increasing amounts of data in battery-powered systems. The main components of IoT devices are autonomous battery-operated smart embedded systems comprising communication circuits, sensors, computing/processing devices as well as integrated memories. These smart connected devices embed RF circuits for communications, digital circuits for data processing, memory for data storage as well as analog circuits such as sensors, filters, and converters. To reduce the fabrication costs and improve system integration, it is desirable to have all these functionalities on the same die and using the same fabrication technology.

A *Non-Volatile Memory* (NVM) based technology that would allow realizing digital, RF and analog functions on the same chip could enhance the integration and reduce the cost for the fabrication of high-end embedded platform for smart connected IoT systems, push forward their miniaturization, decrease their power consumption (by reducing the power consumptions of memory and sensor interfaces blocks by 5× or 10×), enhance their security and improve their reliability while meeting the high performance requirements.

The *Magnetic Tunnel Junction* (MTJ), as the building block of non-volatile magnetic memory (MRAM) such as *Spin Transfer Torque* (STT), is a multilayered nanostructure whose resistance depends on its magnetic state. In its MRAM implementation it behaves as a bistable element that can be used for memory and/or logic functions ("processing/storing"). The MTJ, however, can also be used as a variable resistance for analog applications, including magnetic field or current sensor ("sensing"). So far, these different functions have been achieved separately, using dedicated optimized magnetic tunnel junction stacks. The idea of the GREAT project is to adapt the STT-MRAM to a single baseline technology allowing performing logic and analog functions in the same SoC. This will lead to a unique STT-MRAM MTJ cell which we call *Multifunctional Standardized Stack* (MSS). The basic idea consists in using a standard perpendicular STT-MTJ in memory mode with additional permanent magnets around it, generating an in-plane bias magnetic field to change its behavior for sensors or RF application. This requires only one additional lithography step whose additional cost will be very low compared to the gain offered by the co-integration.

For memory applications, MTJs can have adjustable retention by playing with the diameter of the stack thus allowing to minimize the switching current according to the specified retention. For RF and sensor functions, patterned permanent magnets (for instance made of CoCr alloy or NdFeB) can be added on the two sides of the MTJ pillars, as this is done to bias magnetoresistive heads in hard disk drives. For the spin transfer oscillator, the size and shape of the permanent magnet biasing layer will be adjusted to produce a horizontal field in the order of half of the effective perpendicular anisotropy field (~1kOe) so that the free layer magnetization will be tilted at about 30°. For sensor applications, we will develop a sensor sensitive to the out-of-plane component of the field. First, the diameter of the pillar will be increased compared to the MSS used for memory functions. Besides, the size and shape of the permanent magnet biasing layer will be adjusted to produce a horizontal field slightly larger than the effective perpendicular anisotropy field (~1kOe) so that the free layer magnetization will be pulled in-plane by this biasing field. When submitted to an out-of-plane field to be sensed, the free layer magnetization will rotate upwards or downwards producing a resistance change proportional to the out-of-plane field amplitude.

In this paper we focus on the development in the modeling, process design kit, standard cell, memory evaluation and system level analysis in the scope of this European project. We also show how these pieces of the flow can be combined for cross-layer hybrid design and evaluation.

The rest of this paper is organized as follows. The process design kit is explained in Section II. Section III is devoted to variability analysis. Section IV describes the cross-layer analysis framework for hybrid memory evaluation. Finally, Section V concludes the paper.

978-3-9819263-0-9/DATE18/©2018 EDAA

II. CUSTOM CELL DESIGN

A. Micromagnetic simulations

While operational conditions for MRAM and magnetic sensor functions are well known for the targeted MSS stack based on a fully perpendicular magnetic tunnel junction, the conditions for inducing steady state excitations are less well known. The aim of this task is therefore to provide, via simulations, the conditions of field and current and potentially other material parameters under which the steady state oscillations will be observed. A lot of configurations, varying several parameters of the MSS stack, have been considered. The different scenarios can be separated in two mains sets of simulations. In the first one, the MSS stack is very similar to the one used for memory and sensor functions (with a perpendicular to plane soft layer magnetization orientation). It has been shown that steady states of the magnetization could be obtained but in a range or parameters hardly compatible with a real process implementation (combination of a strong second order anisotropy coefficient, non-zero field-like spin-transfer-torque coefficient and application of a tilted bias magnetic field). So, a second set was considered, with a slightly increased thickness of the storage layer, close to the transition to an in-plane soft layer magnetization. In these conditions, large steady-states out-of-plane precessions, with a frequency between 1 and 14 GHz were obtained (see Fig. 1), with the help of an additional bias magnetic field to tilt the trajectory of the precessions to generate a large output signal.

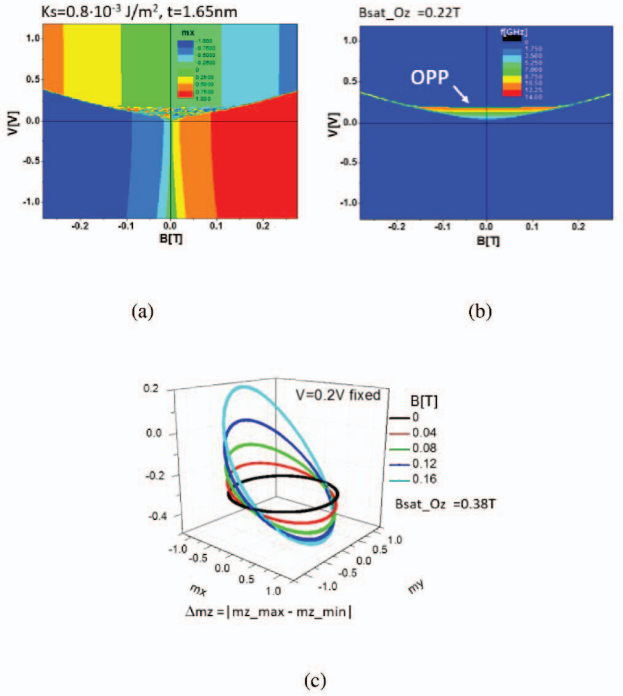


Fig. 1. Static (a) and dynamic (b) voltage-field diagrams of states of free layer magnetization when an out of plane saturation field is also applied (Ks=0.8mJ/m2 and t=1.65nm); (c) Representation of OPP trajectories for a fixed value of voltage and different values of the in plane field.

These results are in good accordance with experimental measurements on the first fabricated devices and are very encouraging for the RF functionality. Although the performance in terms of line-width and power values does not fit the requirements for RF emission applications so far, the device should be relevant to RF detection. However, since the stack slightly differs from the one used for memory or sensor functionalities, the actual co-integrations of the three functionalities still have to be investigated.

B. Compact model of the device

A compact model of the MSS device is required to perform electrical simulations of hybrid circuits. Two approaches are considered for the description of the magnetic behavior in an electrical compact model [1]:

- The first one is based on the Neel-Brown's and Sun's models, giving the typical switching duration as a function of the amplitude of the writing current pulse. This approach offers good performance in terms of simulation time and is particularly adapted for digital circuits where MSS is used in switching mode, but not suitable for use in analog configurations (sensors or RF).
- The second approach is based on the LLG equation which models the dynamics of the magnetization. It is suitable for all the modes of operations and can also represent the stochasticity of the MSS switching thanks to a noise module to be used in a transient noise electrical stimulation.
- Both models have been developed (Fig. 2) in the framework of the project and will be ideally merged to a unified model.

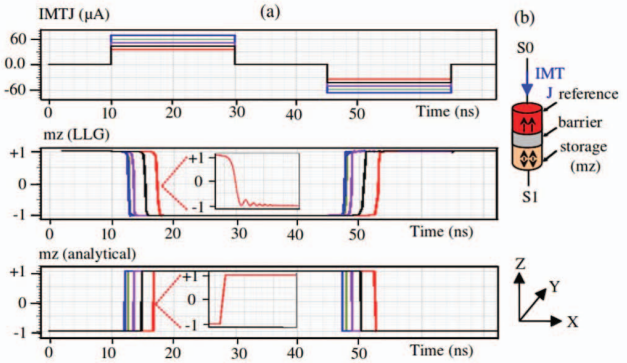


Fig. 2. Comparison of the compact model strategies: the first graph represents the pulse of writing current, whose amplitude is varied. The second one corresponds to the model based on LLG equations, where we can see the expected precessions of the switching and the last one shows the results of the model based on Neel-Brown's and Suns' theories. The models are tuned to give the same results in terms of switching durations

C. Design of standard cells and IP blocks

One of the major objectives of the GREAT project is to integrate a full System on Chip (SoC) embedding memory, logic, sensing and RF functions on the same demonstrator chip. It has been decided to design a processor made non-volatile (NV) by means of the introduction of MSS devices, processing data coming from a MSS-based sensor. This processor will

have sleep and wake-up modes capabilities, thanks to the NV and can be waken-up using a MSS-based wake-up receiver. The schematic of this demonstrator is depicted in Fig. 3.

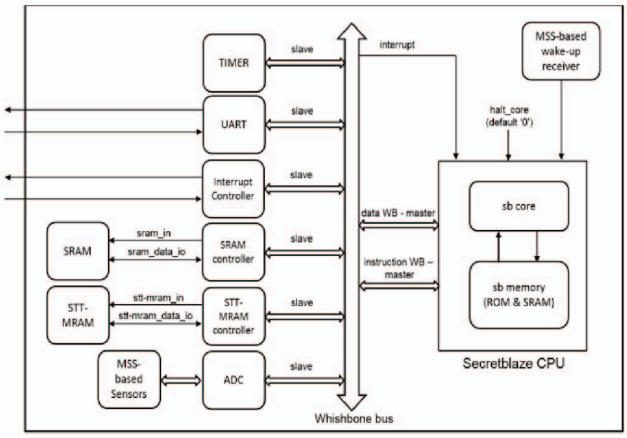
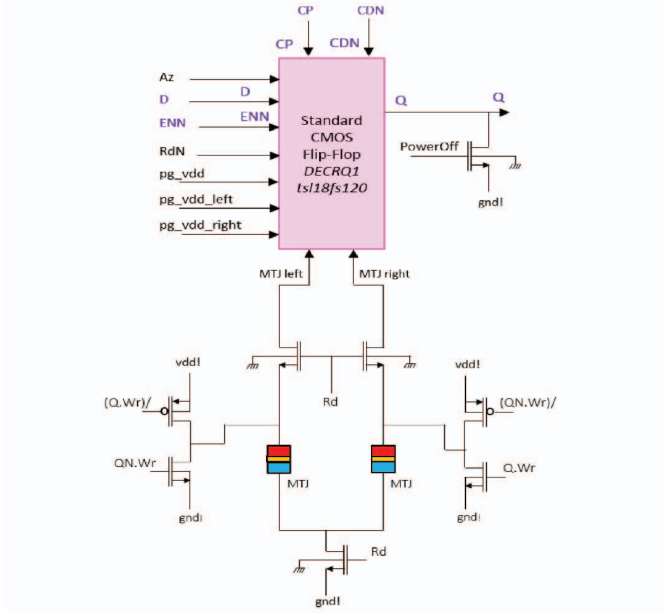
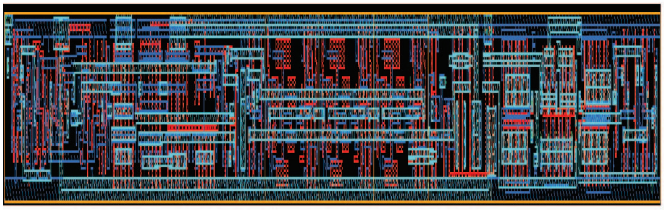


Fig. 3. Schematic of the SoC to be implemented in the final demonstrator



(a)



(b)

Fig. 4. Non-Volatile Flip-Flop based on the MSS device: (a) schematic of the FF, based on a standard FF from the library of TowerJazz (in pink) and (b) layout

The choice has been made to start working with the secretblaze processor [2] developed by the LIRMM. The introduction of the NV is made by replacing some Flip-Flops (FFs) by non-volatile FFs and adding a NV MRAM memory.

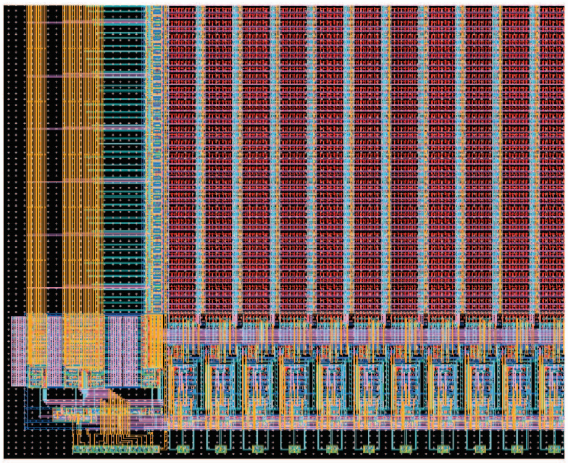
The magnetic FF that has been developed is based on a standard FF from the library of TowerJazz. A circuitry with two MSS devices operating in differential mode has been added to allow back-up and recovery of the output of the FF in and from the MSS devices (Fig. 4(a)). The NV FF has been designed and characterized with Monte-Carlo (MC) and corner simulations. It has been integrated in the digital design flow, with a modified verilog description taking into account the specific operations allowed by the NV, a timing library (.lib) for logic synthesis and an abstract view of the layout for Place and Route operation (Fig. 4(b)).

Concerning the MRAM memory, it has been decided to go for a High Density (HD) implementation of a 128kb memory, with 32 bits IOs and an operating frequency of 10MHz (Fig. 5(a)). The memory has been fully designed with mismatch Monte-Carlo simulations on worst-case corners for reading and writing, on the critical path. The layout has been provided for integration in the final demonstrator (Fig. 5(b)).

A True Random Number Generator (TRNG) circuit, based on the stochasticity of the STT writing and with a controlled

MTJ	200nm	Magnetic techno
TMR, Rp, variability (σ)	100%, Rp=275 Ω	
Retention Endurance	years $>10^{12}$	
Density	*16KB =(0.5KB x 32)*	System Specs
Timing	10 Mhz	
IO width	32	
Bitcell size	W=3µm	Circuit-level Design considerations
Bitcell architecture	1T-1MTJ	
Optimization techniques	SL sharing	
CMOS	*180nm*	CMOS techno
P.Supply (core/IO)	1.8V	

(a)



(b)

Fig. 5. MRAM memory specifications (a) and layout (b)

feedback using an MSS-based programmable current source, has also been proposed and will be integrated in the SoC.

Some of these IPs as well as test structures have been embedded in a first test chip (Fig. 6) whose testing results will be used as a feedback to prepare the final demonstrator.

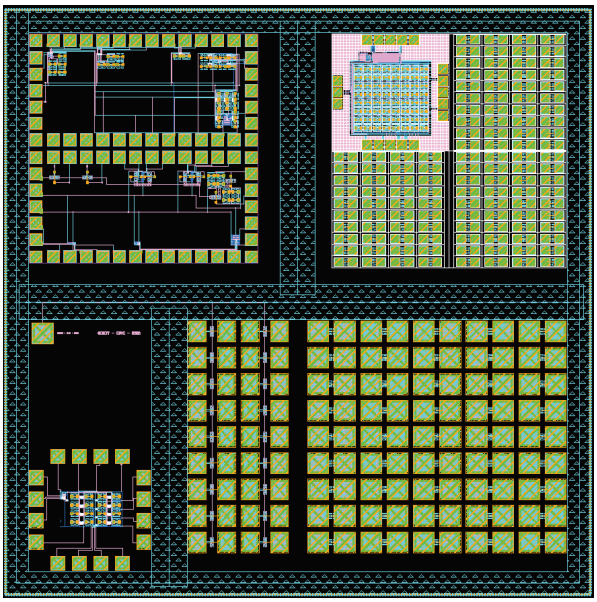


Fig. 6. Layout of the first demonstrator, embedding test structures and circuits from different partners

III. Variability Analysis in STT-MRAM

Like any nano-scale device, STT-MRAM is also affected by manufacturing variations as the technology scales down in the magnetic fabrication process as well as the CMOS process. In addition, the impact of process variation on the magnetic devices exacerbates the stochastic switching behavior of the MTJ. Therefore, quantifying the effect of these variations at the memory architecture level is important for a realistic estimation for the performance, energy and reliability for STT-MRAM. To this end, we have developed a *Variation Aware Estimator Tool for STT-MRAM* (VAET-STT) [6], an early stage design exploration tool for STT-MRAM, which considers process variation, stochastic switching and reliability requirements in its analysis and memory configuration optimization.

The VAET-STT tool is built on the top of NVSim [2] and extends it to account for variability in both the bit-cell and peripheral components. The impact of variability causes the latency and energy of the bit-cell and peripherals to follow distributions instead of being a single (nominal) value as shown in Table 1. The table shows the nominal values (variation-unaware values obtained from NVSim) of the latency and energy along with the mean (μ) and standard deviation (σ) of the distributions at two different technology nodes (45 nm and 65 nm).

From Table 1, we see that μ is much higher than the nominal values. It can also be seen that the effect of variations in write and read latencies is more pronounced in the smaller

	45 nm			65 nm		
	Nominal	μ	σ	Nominal	μ	σ
Write Latency (ns)	4.9	14.7	1.82	4.4	12.1	1.32
Write Energy (pJ)	159.0	425.0	3.73	272.8	512.2	2.79
Read Latency (ns)	1.2	1.7	0.08	1.22	1.5	0.05
Read Energy (pJ)	3.4	4.8	0.002	4.8	5.7	0.001

Table 1: Overall latency and energy values for 45 nm and 65 nm technology nodes for a memory array of 1024×1024

technology node (45 nm) as shown by the higher value of σ/μ in this node. Furthermore, due to the high value of σ for the latencies, a large timing margin is required to keep the error rates within acceptable limits. However, using a smaller technology node helps with both read and write energy reduction. Fig. 7 presents the overall latencies for different values of input Read Error Rate (RER) and Write Error Rate (WER). It can be seen that for lower values of target error rates, high timing margins are required.

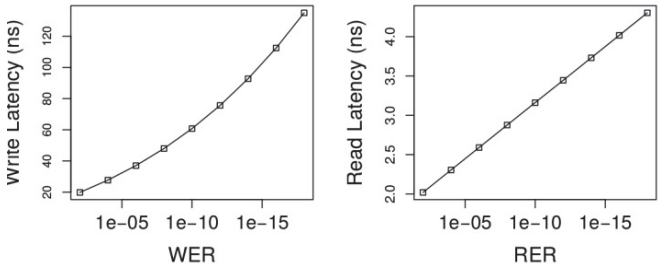


Fig. 7. Overall read and write latencies for various error rates.

Another approach is to reduce the timing margin and employ appropriate Error Correcting Codes (ECCs) to correct errors in the tail of the distribution. The effect of various ECC schemes on the write latency is shown in Fig. 8. The figure shows that compared to the case with no ECC (0-bit correction), there is a drastic improvement in latency by using an ECC with one-bit error correction. However, the improvement in latency for higher bit error correction is comparatively less.

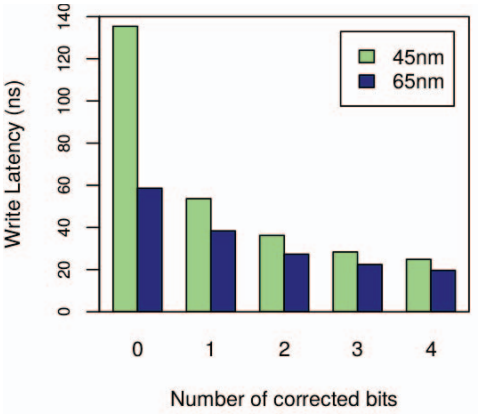


Fig. 8. Effect of ECCs on write latency for WER of 1×10^{-18}

The read operation in STT-MRAM is also affected by read disturb, where the read current accidentally flips the data stored in the MTJ. Fig. 9 shows the read disturb probabilities for different read periods. Even though a higher read latency leads to a lower RER as per Fig. 7, it will lead to increased read disturb probability as shown in Fig. 9. Hence the read period should be fixed considering the conflicting requirements for RER and read disturb.

The VAET-STT tool helps the designer in analyzing the trade-offs between the energy and performance requirements against the target reliability requirements. Moreover, it provides realistic estimations which can better reflect the chip behavior after fabrication. The results show that the variation-aware latency and energy values are significantly higher than those of the nominal case, highlighting the importance of variation-aware analysis.

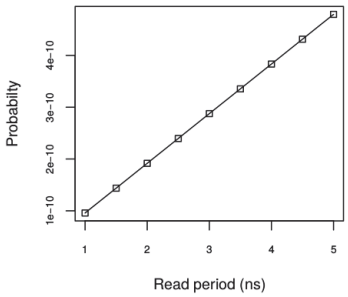


Fig. 9. Read disturb probabilities for different read periods.

IV. Cross-Layer Hybrid Memory Design Exploration Framework

The MAGPIE (Manycore Architecture enerGy and Performance evaluatIon Environment) [8] is a framework with the aim to explore the impact of STT-MRAM on real systems requiring a cross-layer investigation where device, circuit, memory, and system levels are taken into account. Such a simulation platform could be a fast and cost-effective solution to provide essential feedback to enhance the development of STT-MRAM devices. Fig. 10 describes the cross-layer simulation environment for hybrid design exploration.

MAGPIE is built upon three mature and popular tools: the gem5 full-system simulator, the McPAT and VAET-STT power/energy and area estimation tools for CMOS, SOI and emerging non-volatile memory technologies. MAGPIE promotes a script-oriented approach that assists a designer in the design and evaluation tasks.

The MAGPIE evaluation flow considers input parameters comprising information related to the software and hardware components of a target system. The software-related inputs include a gem5 execution script file for each workload/application to be executed, together with the underlying operating system supported by the considered full-system simulator. From the hardware perspective, a number of parameters of the target manycore architectures are required: types and number of cores, memory hierarchy and its technology-specific properties, and the interconnect type.

Provided with these input information, MAGPIE enables a designer to run a seamless evaluation flow that automatically produces results including application outputs, performance numbers, area, power and energy consumption. The user can optionally generate performance and energy numbers in both textual and graphical representations for a convenient design analysis. The ability of MAGPIE to easily evaluate manycore system designs has been shown on the Parsec 3.0 benchmark suite executed on an Exynos 5 Octa SoC model integrating STT-RAM memory at cache level. According to gained insights, MAGPIE improves the productivity by significantly reducing the design exploration.

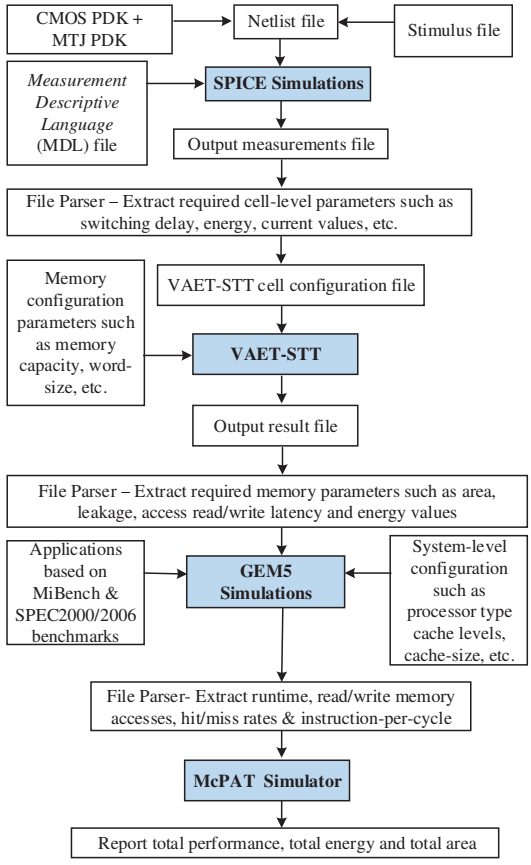


Fig. 10. Hybrid design exploration framework, MAGPIE flow

A. Circuit level

First, a Process Design Kit (PDK) is developed with the device-level parameters, as discussed in Sec. II. This PDK is then used as an input for circuit-level simulation through SPICE. Thus, single bit cells and flip-flops based on MRAM, sense amplifiers, and write circuits can be analyzed. For SPICE simulation, a template file is created for the netlist, stimulus and Measurement Descriptive Language (MDL) after performing several independent simulations. Next, the SPICE simulation generates output measurement file that is then parsed to extract the required cell level parameters such as switching current, delay and energy values. These values are updated into the cell configuration file of the VAET-STT tool.

B. Memory level

Memory-level evaluation is performed thanks to VAET-STT, built on top of NVSIM [3], which is described in Section III. Based on circuit-level data of single bit cell (Section II) and the desired memory architecture information such as capacity, data width, and type of memory (e.g. Cache, RAM, CAM), VAET-STT estimates the access time, the access energy, and the total area of a complete memory chip taking the impact of variations also into account. This tool also includes optimization settings (e.g. buffer design optimization) and various design constraints to facilitate a variation-aware design space exploration before the fabrication of the actual memory chip.

C. System level

Memory-level information is extracted from the previous step to explore the impact of different memory technologies at system level. An accurate performance simulator (gem5 [4]) is used to simulate a single-core or a multi-core architecture with its memory hierarchy. Gem5 generates a detailed report of the system activity including the number of memory transactions (e.g. number of reads/writes, number of hits/misses) and the execution time. This activity information is then used by McPAT [5], a power and area estimator tool at architecture level. Extending the exploration framework with McPAT allows us to analyze not only the energy consumption related to the memory components, but also to evaluate the energy of the complete system including the processor cores, buses, and memory controller.

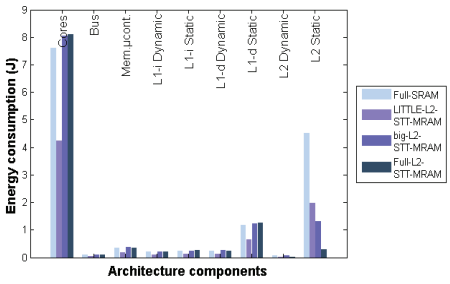


Fig. 11. Energy breakdown by component when executing *bodytrack* kernel on big.LITTLE architecture

D. Results

Fig. 11 provides a summary of typical results obtained with MAGPIE. The following evaluation scenarios are considered: big.LITTLE architecture where all cache memories are in SRAM (i.e., our reference scenario, referred to as Full-SRAM); similar architecture but the L2 cache of the LITTLE cluster is now in STT-MRAM (LITTLE-L2-STT-MRAM), similar architecture but the L2 of the big cluster is in STT-MRAM (big-L2-STT-MRAM), and similar architecture where L2 caches of both clusters are in STT-MRAM (Full-L2-STT-MRAM).

Fig. 12 shows for each kernel its execution time, energy and Energy-Delay-Product (EDP), for the LITTLE-L2-STT-MRAM, big-L2-STT-MRAM and Full-L2-STT-MRAM scenarios compared to the Full-SRAM reference scenario. Here, a 45 nm memory technology is used for illustration.

As expected, putting STT-MRAM in the L2 caches can increase the execution time due to its well-known higher write latency compared to SRAM. Only the scenario with STT-MRAM in the L2 cache of the LITTLE cluster reduces the execution time, up to 50% compared to Full-SRAM. Nevertheless, the overall energy consumption is improved in all scenarios, at least up to 17%. The EDP plotted in Fig. 12 is the product between execution time and energy consumption. It shows that the penalty observed on the execution time when introducing STT-MRAM in L2 caches is compensated by the enabled energy savings.

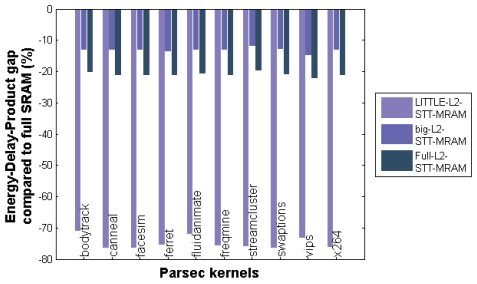


Fig. 12. Energy Delay Product merit

V. Conclusions

In this paper we reviewed the objectives and activities in this European project and presented our preliminary results. This project which spans from technology level all the way to architecture and system, is based on the Multifunctional Standardized Stack (MSS) to enable the use of spintronics for analog and digital sub-systems of IoT platforms. This leads to better integration of embedded & mobile communication systems and a significant decrease of their power consumption.

VI. Acknowledgment

This work was supported by the European Union under Horizon-2020 Program as part of the GREAT project (http://www.great-research.eu/) under grant agreement No 687973. We would like to warmly thank people who worked on the technology in the framework of this project at Spintec Laboratory (Ricardo Sousa, Nathalie Lamard, Laurent Vila, Ursula Ebels, Lucian Prejbeanu), at Singulus (Juergen Langer, Jerzy Wrona) and at TowerJazz (Philippe Azoley, Yakov Roizin).

References

[1] K. Jabeur, et al. "Comparison of Verilog-A compact modelling strategies for spintronic devices," Electronics Letters, pp. 1353–1355, 2014.

[2] L. Barthe, et al. "The secretblaze: A configurable and cost-effective open-source soft-core processor." IPDPSW, pp. 310-313, 2011.

[3] X. Dong, et al. "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," TCAD, vol. 31, no. 7, pp. 994–1007, 2012.

[4] N. Binkert, et al. "The gem5 simulator," ACM SIGARCH Computer Architecture News, vol. 39, no. 2, pp. 1–7, 2011.

[5] S. Li, et al. "Mcpat: an integrated power, area, and timing modeling framework for multicore and manycore architectures," MICRO-42. pp. 469–480, 2009.

[6] S.M. Nair, et al. "VAET-STT: A Variation Aware Estimator Tool for STT-MRAM based Memories," DATE, pp. 1456–1461, 2017.

[7] T. Delobelle, et al. "MAGPIE: System-level Evaluation of Manycore Systems with Emerging Memory Technologies," Workshop on Emerging Memory Solutions - Technology, Manufacturing, Architectures, DATE 2017.