A Universal Spintronic Technology based on Multifunctional Standardized Stack

M. Tahoori[*], S.M. Nair[*], R. Bishnoi[*], L. Torres[**], S. Senni[**], G. Patrigeon[**], P. Benoit [**], G. Di Pendina[***] and G. Prenat[***]

[*]Karlsruhe Institute of Technology, Karlsruhe, Germany
[**]LIRMM, UMR CNRS 5506, University of Montpellier, France
[***]Univ. Grenoble Alpes, CNRS, CEA, INAC-SPINTEC, F-38000 Grenoble, France

***Abstract*—The goal of the GREAT RIA project is to co-integrate multiple functions like sensors ("Sensing"), RF emitters or receivers ("Communicating") and logic/memory ("Processing/Storing") together within CMOS technology by adapting the Spin-Transfer Torque Magnetic Tunnel Junction (STT-MTJ), elementary constitutive cell of the MRAM memories, to a single baseline technology. Based on the STT unique set of performances (non-volatility, high speed, infinite endurance and moderate read/write power), GREAT will achieve the same goal as heterogeneous integration of devices but in a much simpler way. This will lead to a unique STT-MTJ cell technology called Multifunctional Standardized Stack (MSS). This paper presents the lessons learned 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

The interest for developing smart systems based on interconnected objects (Smart Sensors, Secure Elements for the Internet of Things...) is growing fast [1]. It is assumed that 50 billion objects will be connected in a few years. The main components of "Internet of Things" (IoT) devices are autonomous battery-operated smart embedded systems comprising communication circuits, sensors, computing/processing devices as well as integrated memories. Consequently, the key requirements for IoT devices are ultra-low power, high processing capabilities, wireless communication, and autonomy. 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, converters, not to mention cameras, GPS systems etc. In battery-operated Machine to Machine (M2M) and Machine to Human (M2H) operations, the key processing cycle includes the actions sleep, wake-up, sense, store, process, and send. Hence, the enabling technology for IoT should provide 1) ultra-low power, 2) high performance processing, 3) fast, dense, and low power storage, and 4) heterogeneous integration based on "More than Moore" to enable different digital and analog functionalities.

Indeed, according to International Roadmap for Devices and Systems (IRDS) [2], IC power consumption will continue to grow with an increasing contribution of leakage power, which is becoming dominant. With smart connected objects or mobile devices now used as terminals, the need to store and to access simultaneously an increasing amount of data requires energy-efficient embedded architectures. However, the continuously decreasing size of devices and increasing operation frequency leads to critical power consumption and heating issues, which are major challenges.

As pointed out by the IRDS [2], one of the best solutions to overcome this trend is the modification of the memory hierarchy by the integration of Non-Volatility (NV) as a new feature of memories, which would immediately minimize static power as well as pave the way towards normally-off/instant-on computing [3] [4]. Among the existing emerging NV technologies, Spin Transfer Torque Magnetic Random Access Memories (STT-MRAM) have been identified by IRDS together with Redox RAM as the two most promising technologies for embedded memories beyond the 16 nm technology node. Moreover, the Magnetic Tunnel Junctions (MTJs), which are the basic elements of MRAMs, can be used for RF and analog applications, which is specific to this technology. However, so far, MTJs are optimized to perform each function independently. Using the same device for all the functions would allow integrating them on the same chip.

The Magnetic Tunnel Junction (MTJ), as the building block of STT-MRAM, is a multilayered nanostructure whose resistance depends on its magnetic state. In its standard 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 MTJ stacks. The idea of the GREAT project is to adapt the STT-MRAM to a single baseline technology allowing performing logic and analog functions on the same System on Chip (SoC) [5], [6]. This will lead to a unique STT-based MTJ cell which we call Multifunctional Standardized Stack (MSS). The goal of the GREAT RIA project is to co-integrate multiple functions like sensors ("Sensing"), RF emitters or receivers ("Communicating") and logic/memory ("Processing/Storing") together within CMOS by adapting the MTJ structure to a single baseline technology. Based on the STT unique set of per-

978-3-9819263-4-7/DATE20/©2020 EDAA

formances (non-volatility, high speed, infinite endurance and moderate read/write power), GREAT has achieved the same goal as heterogeneous integration of devices but in a much simpler way (memory technology convergence paradigm). The overview of the GREAT project is shown in Figure 1.

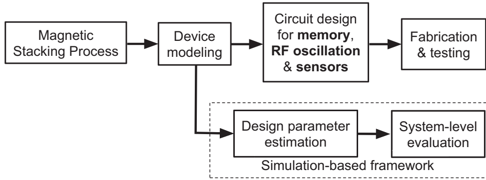


Fig. 1: Overview of EU GREAT project. More details can be found at http://www.great-research.eu/

The basic idea consists of 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 cost is very low compared to the gain offered by the co-integration.

The EU GREAT project had addressed the following three challenges:

- Firstly, in order to realize an MSS technology, each flavor of MTJ device are modified to enable a single core MTJ technology, which support multiple functions (RF, sensors, memory, etc.). As all functions respond similarly to such a universal stack, an intensive R&D effort on the technology (materials, processing) and simulations are conducted.
- Secondly, the design of these circuits also adapted to the specifications of the new technology. This implied a large design effort to accommodate a potentially sub-optimal technology. Moreover, in the digital part, the memory hierarchy also considered the unique performance set of magnetic memories/cells (co-design and co-optimization).
- Thirdly, at the SoC level, it is necessary to ensure that different heterogeneous functions are able to interact efficiently, regardless of their specifications/performance differences to be able to meet high-end IoT hardware architecture platforms for battery-powered M2M and M2H systems (system 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 [7] [8]. 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 was adjusted to produce a horizontal field in the order of half of the effective perpendicular anisotropy field ($\approx$1kOe) so that the free layer magnetization can be tilted at about 30°. For sensor applications, we developed a sensor sensitive to the out-of-plane component of the field. First, the diameter of the pillar is increased compared to the MSS used for memory functions. Besides, the size and shape of the permanent magnet biasing layer is adjusted to produce a horizontal field slightly larger than the effective perpendicular anisotropy field ($\approx$1kOe) so that the free layer magnetization is pulled in-plane by this biasing field. When submitted to an out-of-plane field to be sensed, the free layer magnetization rotates upward or downward producing a resistance change proportional to the out-of-plane field amplitude.

The rest of this paper is organized as follows. The manufacturing details are explained in Section II. Section III is devoted simulation-based design flow explanation. Section IV describes the non-volatile system on chip design in the scope of the project. Finally, Section VI concludes the paper.

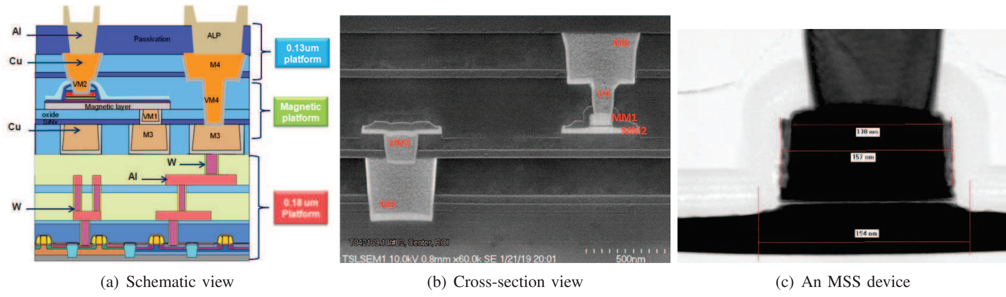
II. MANUFACTURING DETAILS

The schematic and cross section view of the hybrid process developed in the GREAT project is shown in Figure 2 and the cross-section of MSS stack is shown in Figure 2(c). Few experiments were realized to evaluate three different Tunneling Magneto-Resistance (TMR) Hard Mask, finally a thin layer of 80 $\mathring{A}$ of Ta has been chosen and validated on the manufacturing line. The TMR stack has been defined by the consortium and a special Ion Beam Etching (IBE) etch recipe has been developed in the project and provided to the foundry partner. Process development was done to adapt and verify the recipe on IBE tool of the foundry and etch the 45 unit layers that composed the TMR stack. Thanks to TEM x-section, the IBE etch process has been validated (good profile, good stop on expected layers).

The minimum Critical Dimension (CD) was around 200 nm and we succeed to achieve 150 nm dot size with good yield. The engineering work focused on 3 steps: Photo patterning, Hard Mask etch and IBE etch. By modifying photo parameters, we changed negative profile resist dots to positive and solved by this way the falling dot failure. After this first step, a new recess process has been added to the Hard Mask etch operation in order to decrease dot size. Finally, we validated that after IBE etch, CDs and profiles are according to expectation.

Following this activity, engineering work was needed to adapt Tower standard via process (Photo alignment and etch) to small dots size and avoid any misalignment that will create device shorts. To reach the CD's target, a trimming process has been added before the via etch step. Results were checked in SEM x-section and have shown good alignment between TMR dots and contact vias. All these developments were validated during final electrical test at the foundry using standard test chip.

The results of TMR test are summarized in Table I. Figure 3 shows the yield and CD of fabricated devices. As can be seen in the fabrication results for the MSS stack, the obtained yield and TMR values are quite acceptable, confirming the manufacturability of MSS stack.



(a) Schematic view (b) Cross-section view (c) An MSS device

Fig. 2: Demonstration of the hybrid process developed in GREAT project.

TABLE I: Electrical tests were performed on two different TMR stack of CoFeB: 1.3 nm and 1.4 nm

RA	$7.5\ \Omega.\mu m^2$				$10\ \Omega.\mu m^2$					
CoFeB thickness	1.3 nm		1.4 nm		1.3 nm			1.4 nm		
Electric CD	125 nm	125 nm	140 nm	105 nm	140 nm	140 nm	140 nm	100 nm	90 nm	80 nm
Median Die Yield (TMR>120%)	76%	77%	86%	77%	89%	91%	90%	88%	72%	3%
Median TMR	140%	140%	145%	150%	150%	150%	150%	160%	160%	30%

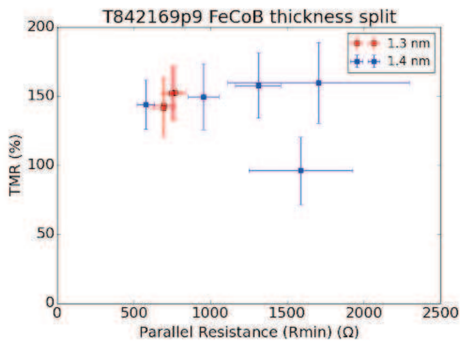


Fig. 3: Device yield and CD.

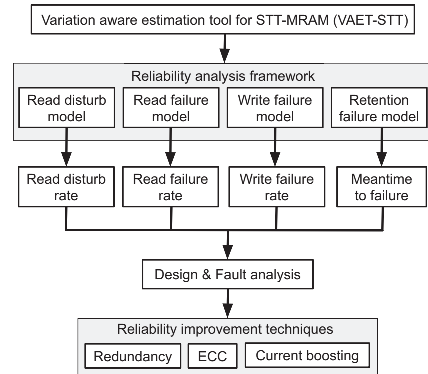
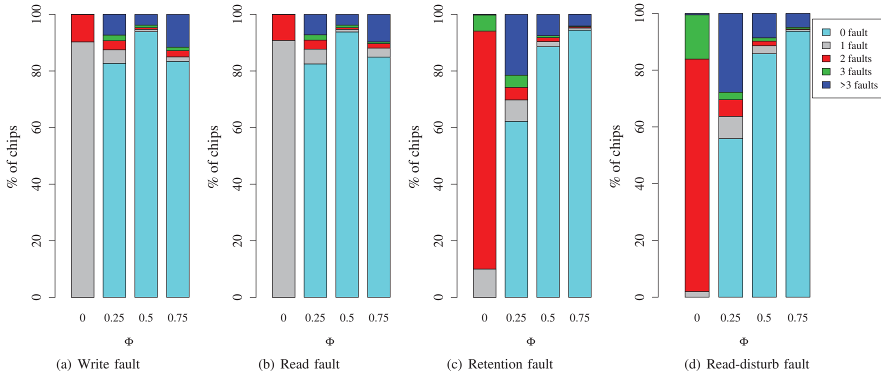


Fig. 4: Design and reliability framework for STT-MRAM

III. Simulation-Based Design and Yield Analysis Framework

Exploring the impact of STT-MRAM on real systems requires 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. Moreover, this exploration framework would also give the possibility to evaluate hybrid designs by considering several memory technologies inside the system. We employed different sets of tools to build an accurate exploration framework for performance, energy and area analysis of a full system based on STT-MRAM. More details about this simulation framework can found in [6] [9] [3].

These memory design parameters of STT-MRAM can significantly vary under the influence of variations. The functionality of this memory is influenced by variations in CMOS components as well as by the magnetic devices due to their fabrication process steps. The manufacturing imperfections in the magnetic components disturb the device characteristics of the cell such as switching characteristics, resistance differences, etc. Therefore, quantifying the effect of these variations at the memory architecture level is important for a realistic estimation for the energy, performance and reliability for STT-MRAM. To tackle this, we have developed a Variation Aware Estimator Tool for STT-MRAM (VAET-STT) [11]. This is 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. Overall design and reliability analysis framework used in this project is depicted in Figure 4.



(a) Write fault (b) Read fault (c) Retention fault (d) Read-disturb fault

Fig. 5: Percentage of chips with their fault types for a 512×512 memory at 25°C for various correlation coefficients (ϕ) [10].

In this project, we have developed model for both reliability failures such as read/write failures, read-disturb and retention failures as well as permanent faults due to parametric variations of the STT-MRAM. The yield analysis flow is presented in detail in [10]. The parameters considered in our analysis are the radius (r) of the MTJ and the threshold voltage (V_{th}) of all CMOS components. The correlation maps for these parameters are then obtained from the VARIUS tool [12]. For each of these correlation maps, we get the parametric failures by performing Monte-Carlo simulations for the entire memory system. The yield is then obtained by performing Monte-Carlo over multiple maps (corresponding to different chip instances). We then explore the right combination and efficiency of different defect tolerance techniques such as Error Correcting Code (ECC) [13] [14], Redundancy (RR) [15] and current boosting [16] [17] to obtain a target yield.

Yield exploration was done from the failure maps of different Monte-Carlo runs corresponding to different chip instances and by analyzing the number of faults in a row or column. If there are large number of faults per row or column, i.e., the faults are clustered, then Redundancy (RR) is a good technique to mitigate these faults. On the other hand, for a small number of faults per row or column, i.e., when faults are more uniformly distributed, ECC might be a good option. In case of more number of single isolated faults, advanced techniques such as those proposed in [18] could be optimal for yield improvement. Besides the conventional yield improvement techniques, we also explore some of the techniques specific to STT-MRAM. Since the switching probability and the latency of STT-MRAM is highly sensitive to the write current, increase in current (using current boosting technique) can significantly decrease the write latency resulting in reduced write failures. A 10% increase in write current can decrease the write latency of a bit-cell to around one-third. This will result in a significant reduction in the number of write failures.

Figure 5 shows the percentage of chips and the corresponding fault types for a 512×512 memory array. Here the percentage of chips with 0 faults indicates the yield. As shown in the figure, when ϕ = 0, the yield is very low. In such case, most of the chips fail due to 2 or 3 line faults due to permanent read and write faults. However, for retention and read-disturb case, most of the lines fail due to greater than 3 faults. The reason is that the number of faults is much higher for retention and read-disturb, which increase the probability of having more faults per line. As ϕ increases, the faults become clustered, which increases the probability of having a large number of faults on some chips, and less number of faults on some other chips. This means that as ϕ increases, the probability of having chips with 0 faults also increases, thus improving the yield. The amount of yield improvement depends on the nature and actual distribution of faults.

On the other hand, the system-level simulation were performed using an accurate performance simulator (GEM5) [19] 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 cache hits/misses) and the execution time [20]. This activity information is then used by McPAT [21], a power and area estimator tool at architecture level. Extension of the exploration framework with McPAT is done to allow 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. This highlights which memory components are power hotspots and should be evaluated to be replaced with other memory technologies. Thus, this extension is key in order to perform evaluation and combination of different memory technologies (memory convergence) to achieve the best compromise between performance, energy demand, area and cost [22].

IV. Non-volatile SoC design

One of the major objectives of the GREAT project is to integrate a full SoC embedding memory, logic, sensing and RF functions on the same demonstrator chip. It has been decided to design an MCU (Microcontroller Unit) made non-volatile (NV) by means of the introduction of MSS devices, processing data coming from a MSS-based sensor [23]. This processor will have sleep and wake-up modes capabilities

(normally-off computing), thanks to the NV and can be woken-up using an MSS-based wake-up receiver. The schematic of this demonstrator is depicted in Figure 6. A full layout of the MCU is shown in Figure 7 and some features are re given in Table II. This non-volatile MCU can run at 20 MHz and has a backup/recovery time of about 4 μs. The energy costs are respectively 437 nJ (backup) and 98.7 nJ (recovery). It is worth noting that the current implementation of the MCU aims at giving a proof of concept instead of an optimized design. For instance, this work has made all the flip-flops of the processor non-volatile to preserve the system state after a power supply shutdown. Thus, optimizations are still possible for the backup/recovery time and energy. Moreover, considering a more advanced technology node would further improve the overall performances.

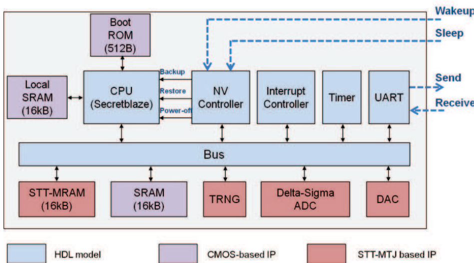


Fig. 6: Schematic of the SoC to be implemented in the final demonstrator.

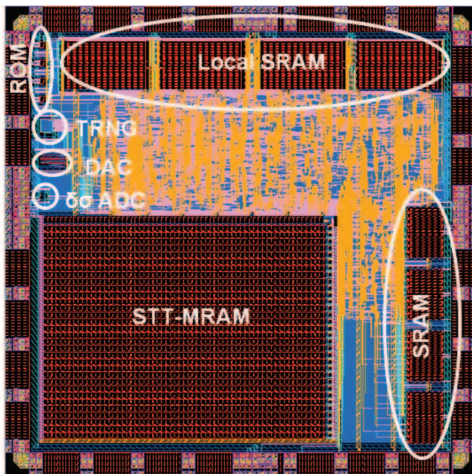


Fig. 7: SoC Layout of the project demonstrator.

Based on the real application behavior obtained from smart agriculture monitoring use case, illustrated in power profile of an industrial IoT (see Figure 8), the normally-off computing has been compared more precisely with the traditional MCU solution. Several case studies were considered regarding the number of times the IoT device has to wake up during the execution of the application. The objective is to determine the ratio of time spent in run/sleep modes and analyze in which cases the non-volatile MCU could offer better energy efficiency. The case studies taken into account consider a periodic wake-up of 15 minutes to accomplish the main task of the application. The number of additional wake-ups (minor task) along each period depends on the sensor. Three case studies were considered with 15 (e.g. rain gauge), 200 (e.g. water meter) and 9000 (e.g. anemometer) additional wake-ups.

TABLE II: SoC features

Die Area	23 mm^2
Process	180 nm CMOS/200 nm STT-MTJ
Supply voltage	1.8V Core/ 3.3V IOs
Frequency	20 MHz
# of NV Flip-flops	2126
Backup time	4.15 μs
Recovery time	4.15 μs
Backup energy	437 nJ
Recovery energy	98.7 nJ

In run mode, the MCU power consumption with the Flash+SRAM configuration (traditional MCU) remains lower than that of the MCU based on the STT-MRAM (non-volatile MCU) by a factor of 1.8 (normal due to the size of the MTJ). However, this overhead is compensated by the energy saved during the sleep phase, making the NV MCU more energy efficient after a few hundreds of ms (it takes around 65 ms to compensate the backup energy). Considering the ratio of time spent in run and sleep modes between two periodic wake-ups, the SoC energy consumption for each case study and for both MCU solutions has been summarized in Table III. If no sensor event wakes up the device in addition to the periodic wake-up, the non-volatile MCU clearly offers better energy efficiency than the traditional MCU by a large factor of 663. In the cases of 15 and 200 additional wake-ups, the non-volatile MCU still remains the best solution by a factor of 50 and 4.5, respectively. However, the last case, which considers the large number of 9000 sensor events, shows a better energy efficiency for the traditional MCU due to its low power consumption in run mode compared to the non-volatile MCU.

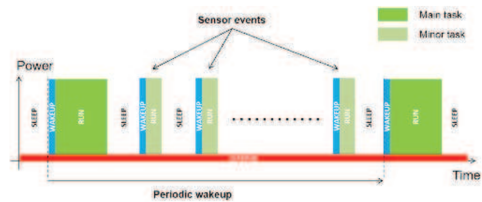


Fig. 8: Power profile of an industrial IoT.

The MCU power analysis presented above definitely demonstrates the interest of the non-volatility inside the processor of an MCU. Even when considering large MTJs of 200 nm with high read/write currents during run mode, the energy benefit in sleep mode is quickly noticeable with a low minimum time T_{sleep}. Although the MCU design of this work is based on a process node not comparable with industry, the

TABLE III: Energy comparison between NV-MCU STT-MRAM and SRAM based MCU for different sensor events scenarios.

Sensor events	Energy Consumption (E_{MCU})	(E_{NV-MCU})	$\frac{E_{MCU}}{E_{NV-MCU}}$
0	6.03 mJ	9.09 μJ	663
15 (rain gauge)	6.08 mJ	121 μJ	50
200 (water gauge)	6.81 mJ	1.5 mJ	4.5
9000 (anemometer)	41.2 mJ	67.2 mJ	0.6

study remains particularly interesting. Today, most MCUs uses old processes, some as high as 350 nm [24]. This is because there is enough performance in existing products to satisfy most IoT applications, and most embedded applications do not require leading-edge performance but the lowest power consumption and the best power efficiency [25]. However, the need for more functionalities induces manufacturers to move to lower technology nodes with more memory, processing, and connectivity into the same space.

Based on these experiments, and MCU use, we compare MCU performances between Flash based MCU and STT-MRAM based MCU. A first extrapolation on a 28 nm Flash vs 40 nm STT-MRAM (28 nm CMOS) has been realized and we demonstrated that STT-MRAM for this node technology is really interesting for MCU. The overall gain is about 70 % - Same CPU, same memory size (128 kB) and same application (DES algorithm). All the comparison methodology flow has been realized and will serve has a foundation to MCU comparison with various memory technologies.

V. Conclusions

In this paper, we reviewed the objectives and activities in this EU GREAT project and presented our project 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 and mobile communication systems and a significant decrease of their power consumption.

VI. Acknowledgement

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] Internet of Things - number of connected devices worldwide 2015-2025, https://www.statista.com/, 2019.

[2] International Roadmap for Devices and Systems, https://irds.ieee.org/, 2017.

[3] S. Senni, L. Torres, G. Sassatelli, and A. Gamatie, "Non-volatile processor based on MRAM for ultra-low-power IoT devices," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 13, no. 2, p. 17, 2017.

[4] S. Senni, L. Torres, P. Benoit, A. Gamatie, and G. Sassatelli, "Normally-Off Computing and Checkpoint/Rollback for Fast, Low-Power, and Reliable Devices," *IEEE Magnetics Letters*, vol. 8, pp. 1–5, 2017.

[5] M. Tahoori, S. Nair *et al.*, "Using multifunctional standardized stack as universal spintronic technology for IoT," in *DATE*, 2018, pp. 931–936.

[6] M. Tahoori *et al.*, "GREAT: heteroGeneous integRated magnetic tEchnology using multifunctional standardized sTack," in *ISVLSI*, 2017, pp. 344–349.

[7] N. Sayed, L. Mao *et al.*, "Compiler-Assisted and Profiling-Based Analysis for Fast and Efficient STT-MRAM On-Chip Cache Design," *TODAES*, vol. 24, no. 4, p. 41, 2019.

[8] N. Sayed *et al.*, "Process variation and temperature aware adaptive scrubbing for retention failures in STT-MRAM," in *ASPDAC*, 2018, pp. 203–208.

[9] F. Oboril *et al.*, "Evaluation of hybrid memory technologies using sot-mram for on-chip cache hierarchy," *TCAD*, vol. 34, no. 3, pp. 367–380, 2015.

[10] S. M. Nair *et al.*, "A Comprehensive Framework for Parametric Failure Modeling and Yield Analysis of STT-MRAM," *TVLSI*, 2019.

[11] S. M. Nair, R. Bishnoi *et al.*, "VAET-STT: Variation aware STT-MRAM analysis and design space exploration tool," *TCAD*, vol. 37, no. 7, pp. 1396–1407, 2017.

[12] S. R. Sarangi *et al.*, "VARIUS: A model of process variation and resulting timing errors for microarchitects," *Transactions on Semiconductor Manufacturing*, vol. 21, no. 1, pp. 3–13, 2008.

[13] N. Sayed *et al.*, "Fast and Reliable STT-MRAM Using Nonuniform and Adaptive Error Detecting and Correcting Scheme," *TVLSI*, vol. 27, no. 6, pp. 1329–1342, 2019.

[14] N. Sayed, F. Oboril *et al.*, "Leveraging systematic unidirectional error-detecting codes for fast STT-MRAM cache," in *VLSI Test Symposium (VTS)*, 2017, pp. 1–6.

[15] C. Münch *et al.*, "Reliable in-memory neuromorphic computing using spintronics," in *Asia and South Pacific Design Automation Conference*, 2019, pp. 230–236.

[16] N. Sayed *et al.*, "A cross-layer adaptive approach for performance and power optimization in STT-MRAM," in *DATE*, 2018, pp. 791–796.

[17] R. Bishnoi *et al.*, "Improving write performance for STT-MRAM," *Transactions on Magnetics*, vol. 52, no. 8, pp. 1–11, 2016.

[18] W. Kang *et al.*, "Yield and reliability improvement techniques for emerging nonvolatile STT-MRAM," *J. Emerg. Sel. Topics Circuits Syst.*, vol. 5, no. 1, pp. 28–39, 2014.

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

[20] G. Prenat *et al.*, "Ultra-fast and high-reliability SOT-MRAM: From cache replacement to normally-off computing," *TMSCS*, vol. 2, no. 1, pp. 49–60, 2015.

[21] S. Li *et al.*, "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures," in *International Symposium on Microarchitecture*, 2009, pp. 469–480.

[22] T. Delobelle *et al.*, "Magpie: System-level evaluation of manycore systems with emerging memory technologies," 2017.

[23] S. Senni, F. Ouattara, J. Modad, K. Sevin, G. Patrigeon, P. Benoit, P. Nouet, L. Torres, F. Duhem, G. Di Pendina *et al.*, "From Spintronic Devices to Hybrid CMOS/Magnetic System On Chip," in *VLSI-SoC*, 2018, pp. 188–191.

[24] Evolution of the mcu, https://semiengineering.com/evolution-of-the-mcu/, 2017.

[25] 2017 embedded processor report: At the edge of moore's law and iot, http://www.embedded-computing.com/, 2017.