UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Exploiting In-Memory Computing for High-Security SOT-MRAM Based Physical Unclonable Functions

Proyecto de Investigación y Desarrollo

Eduardo Javier Holguín Weber

Raffaele de Rose, Ph.D.
Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito
para la obtención del título de Master en Nanoelectrónica

Quito, 11 de diciembre del 2020

Universidad San Francisco de Quito USFQ

Colegio de posgrados

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Exploiting In-Memory Computing for High-Security SOT-MRAM Based Physical Unclonable Functions

Eduardo Javier Holguín Weber

Nombre del Director del Programa: Omar Aguirre
Título académico: Doctor of Philosophy
Director del programa de: Maestría en Nanoelectrónica

Nombre del Decano del colegio Académico: César Zambrano
Título académico: Doctor of Philosophy
Decano del Colegio: Colegio de Ciencias e Ingenierías

Nombre del Decano del Colegio de Posgrados: Hugo Burgos
Título académico: Doctor of Philosophy

Quito, 11 de diciembre del 2020

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombre del estudiante:	Eduardo Javier Holguín Weber
Código de estudiante:	00208079
C.I.:	0919663609
Lugar y fecha:	Quito, 11 de diciembre del 2020

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following graduation project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

AGRADECIMIENTOS

A la Universidad San Francisco de Quito en especial a todos los miembros que componen la Maestría en Nanoelectrónica la misma que me permitió ganar los conocimientos necesarios para desarrollar este trabajo de titulación y por medio de su programa de doble titulación me permitió realizar mi segundo año de maestría en la Universitta della Calabria en Italia, también a los profesores del DIMES quienes me guiaron durante proceso de Tesis y mis compañeros quienes me apoyaron a lo largo de este camino.

Resumen

Esta tesis presenta el diseño de una Physical Unclonable Function (PUF) basada en Spin-Orbit-Torque Magnetic Random Access Memory (SOT-MRAM), que incorpora Computing in memory (CiM) para mejorar la calidad y seguridad de la respuesta PUF. El secreto de la PUF se almacena como un estado aleatorio en una matriz de dispositivos SOT-MRAM de tres terminales que consta de una Magnetic Tunnel Junction (MTJ) con un Free Layer (FL) perpendicular colocada en la parte superior de un Heavy Metal (HM). Dicho estado aleatorio se obtiene aplicando una corriente suficientemente grande en el HM, que impulsa la magnetización de FL a lo largo de la dirección en el plano. Una vez que se apaga la corriente, la magnetización de FL evoluciona a configuraciones perpendiculares ascendentes o descendentes con una probabilidad bastante similar cercana al 50%, dando lugar a escribir un bit aleatorio “0” o “1” en el MTJ. Luego, siguiendo el enfoque de In-Memory Computing, se explota un circuito de detección mejorado en el diseño PUF basado en SOT-MRAM para realizar operaciones lógicas XOR bit a bit durante la operación de lectura.

La arquitectura PUF se diseñó en (Cadence Virtuoso) mediante el uso de un enfoque de modelado híbrido CMOS/Spintronic. Para los dispositivos CMOS, se han considerado modelos de transistores proporcionados por una tecnología FinFET de 0,8 V / 1,8 V de 18 nm. Para dispositivos SOT-MRAM de tres terminales, se ha empleado un modelo compacto basado en Verilog-A. Se han realizado simulaciones eléctricas y estadísticas para evaluar características energéticas y métricas de seguridad del circuito PUF. Los resultados obtenidos demuestran que la implementación de la estrategia CiM en el circuito diseñado permite mejorar considerablemente tanto la aleatoriedad como la unicidad de la respuesta PUF, aumentando así la calidad y la seguridad del dispositivo PUF.

Palabras clave: Magnetic Tunnel Junction (MTJ), Spin-Orbit torque magnetic RAM (SOT-MRAM), Physical Unclonable Function (PUF), Computing in Memory (CiM), Free Layer (FL).

ABSTRACT

This thesis presents the design of a Spin-Orbit-Torque Magnetic Random Access Memory (SOT-MRAM) based Physical Unclonable Function (PUF), which incorporates Computing-in-Memory (CiM) to enhance the quality and security of the PUF response. The secret of the PUF is stored into a random state of an array of three-terminal SOT-MRAM devices consisting of a Magnetic Tunnel Junction (MTJ) with a perpendicular free layer (FL) placed on the top of a heavy metal (HM). Such random state is obtained by applying a large enough current into the HM, which drives the FL magnetization along the in-plane direction. Once the current is switched off, the FL magnetization evolves to either up or down perpendicular configurations with quite similar probability close to 50%, thus giving rise to write a random bit “0” or “1” in the MTJ. Then, by following In-Memory Computing approach, an enhanced sensing circuitry is exploited in the SOT-MRAM based PUF design to perform bitwise XOR logic operations during the reading operation.

The PUF architecture has been designed into a commercial circuit design tool (Cadence Virtuoso) by using a hybrid CMOS/spintronics modeling approach. For CMOS devices, transistor models provided by a 0.8V/1.8V 18-nm FinFET technology have been considered. For three-terminal SOT-MRAM devices, a Verilog-A based compact model has been employed. Electrical and statistical simulations have been performed to evaluate energy characteristics and security metrics of the PUF circuit. Obtained results demonstrate that the implementation of the CiM strategy in the designed circuit allows considerably improving both randomness and uniqueness of the PUF response, thus increasing the quality and the security of the PUF device.

Contents

I. Introduction **14**

II. STT-MRAM and SOT-MRAM Devices **20**

2.1 Magnetic Tunnel Junction (MTJ) 20

2.2 STT and SOT switching 22

2.3 Reading operation 24

2.4 STT-MRAM and SOT-MRAM array 25

III. Computing-in-Memory (CiM) with STT- and SOT-MRAM **27**

3.1 CiM strategy 27

3.2 Circuitry for voltage sensing scheme 33

IV. SOT-MRAM Based PUF with Computing-in-Memory **36**

4.1 Evaluation metrics of PUF devices 36

4.2 Three-terminal SOT-MRAM device 39

4.3 SOT-MRAM based PUF design with CiM operation 31

4.3.1 Electrical simulation results 45

4.3.2 Statistical simulation results 46

V. Summary and Conclusions **52**

References **54**

List of Figures

Figure 1.1: (a) CPU and memory connected through a bandwidth-limited bus for data transfer. (b) Performance improvement for CPU and memory in time [3] 15

Figure 1.2: Four approaches to deal with the Von Neumann Bottleneck [4] 16

Figure 1.3: Operating principle of the SOT-MRAM based PUF proposed in [7] 18

Figure 2.1: Basic structure of an MTJ with perpendicular magnetic anisotropy (PMA) 20

Figure 2.2: Two different resistance states in an MTJ with PMA 21

Figure 2.3: STT switching in a two-terminal MTJ-based device 22

Figure 2.4: SOT switching in a three-terminal MTJ-based device [14] 23

Figure 2.5: Different sensing schemes: (a) current sensing (CS) scheme and (b) voltage sensing (VS) scheme 24

Figure 2.6: STT-MRAM array with the detail of the 1T-1MTJ bitcell 25

Figure 2.7: SOT-MRAM array with the detail of the 2T-bitcell 26

Figure 3.1: Operating principle of the considered CiM solution referred to a current sensing (CS) scheme [8]: (a) bitcell resistive equivalent, (b) bitcell sensing current depending on the MTJ state, (c) bitwise CiM operation between two bitcells, and (d) source-line (SL) sensing current resulting from bitwise CiM operation 28

Figure 3.2: Current sensing schemes for bitwise (a) OR and (b) AND logic operations, along with the detail of the required reference currents [8] 29

Figure 3.3: Current sensing scheme for bitwise XOR operation 30

Figure 3.4: Voltage sensing schemes for bitwise (a) OR and (b) AND logic operations, along with the detail of the required reference voltages 31

Figure 3.5: Latch-type voltage sensing amplifier [17] 33

Figure 3.6: Circuits to generate reference voltages (i.e. $V_{ref\text{-}or}$ and $V_{ref\text{-}and}$) according to the CiM scheme of Fig. 3.4 34

Figure 3.7: Comparison between reference voltages generated by the circuit of Fig. 3.6 and the BL voltage depending on the states of the MTJs belonging to the two activated bitcells: (a) $V_{P\text{-}P}$, (b) $V_{AP\text{-}P}$, and (c) $V_{AP\text{-}AP}$ 35

Figure 3.8: Timing diagram of the voltage sensing CiM solution. Here, the resistance states of the MTJs belonging to the two activated bitcells is changed every 20 ns to evaluate the output of the whole sensing circuitry for three different cases: R_P-R_P, R_P-R_{AP} (or R_{AP}-R_P), and R_{AP}-R_{AP} 35

Figure 4.1: Example of PUF uniqueness evaluation 38

Figure 4.2: Example of PUF reliability evaluation 39

Figure 4.3: (a) Sketch of the SOT-MRAM device and operating principle to write a random bit thanks to the application of the current J_{SHE}. (b) Time description of the J_{SHE} current pulse. (c) Example of the time evolution of the FL magnetization 40

Figure 4.4: Block diagram of the PUF architecture 42

Figure 4.5: Scheme of the bitcell array in the circuit with conventional voltage sensing scheme, along with the detail of the required reference voltages 43

Figure 4.6: Scheme of the bitcell array in the circuit with enhanced voltage sensing scheme to implement a 2-bit XOR operation between two BCs within the same column 43

Figure 4.7: Color map of the average HW (over 10,000 challenges and 1,000 PUF instances) for the circuit with conventional sensing scheme (i.e.

w/o 2-bit XOR operations) in the case where all the SOT-MRAM devices feature the same switching probabilities ($P_{up\text{-}to\text{-}down}$ and $P_{down\text{-}to\text{-}up}$ ranging from 39% up to 61%)..47

Figure 4.8: Color map of the average HW (over 10,000 challenges and 1,000 PUF instances) for the circuit with enhanced sensing scheme (i.e. w/ 2-bit XOR operations) in the case where all the SOT-MRAM devices feature the same switching probabilities ($P_{up\text{-}to\text{-}down}$ and $P_{down\text{-}to\text{-}up}$ ranging from 39% up to 61%) 47

Figure 4.9: Color map of the average HD_{INTER} (over 10,000 challenges and 1,000 PUF instances) for the circuit with conventional sensing scheme (i.e. w/o 2-bit XOR operations) in the case where all the SOT-MRAM devices feature the same switching probabilities ($P_{up\text{-}to\text{-}down}$ and $P_{down\text{-}to\text{-}up}$ ranging from 39% up to 61%)..48

Figure 4.10: Color map of the average HDINTER (over 10,000 challenges and 1,000 PUF instances) for the circuit with enhanced sensing scheme (i.e. w/ 2-bit XOR operations) in the case where all the SOT-MRAM devices feature the same switching probabilities ($P_{up\text{-}to\text{-}down}$ and $P_{down\text{-}to\text{-}up}$ ranging from 39% up to 61%)..48

Figure 4.11: Top-view of the FL geometry in the cases of (a) nominal geometry and in presence of defects: (b) same area, (c) smaller area, and (d) larger area 49

List of Tables

Table 1.1: Comparison between conventional CMOS-based and emerging non-volatile memories [2] 15

Table 3.1: Possible values of bit-line (BL) sensing voltage resulting from bitwise CiM operation 31

Table 4.1: Device parameters of the SOT-MRAM device 41

Table 4.2: Summary results of electrical simulations 46

Table 4.3: Switching probabilities in the case of defects in the planar geometry of SOT-MRAM devices 50

Table 4.4: Summary results of statistical simulations 51

I. INTRODUCTION

The idea of having many of our day-to-day devices interconnected is no longer a fantasy thanks to the Internet of Things (IoT), i.e. a system of interrelated computing devices with unique ID and the capability to transfer a large amount of data over a network without requiring human-to-human or human-to-computer interaction. In the IoT era, the growth in the data processed and the increase in the number of cores have placed high demands on memory of modern computing systems. Accordingly, a growing fraction of power consumption and area is related to memories. Conventional semiconductor-based memories (e.g. SRAM and DRAM) have been the mainstays of memory in the past decades. However, fundamental challenges of CMOS scaling along with the increased demand for memory capacity and performance have led the research towards alternative memory technologies. In particular, spintronic memories have recently emerged as a very promising technology to overcome the limitation of CMOS scaling towards the end of Moore’s law [1]. Among spintronic solutions, Spin-Transfer-Torque (STT) and Spin-Orbit-Torque (SOT) Magnetic Random Access Memories (MRAMs) are widely considered as premiere candidates for post-CMOS on-chip non-volatile storage thanks to their potential for low-power and high-speed operation, near-zero leakage, long endurance, and technological scalability. A comparison between conventional CMOS-based and emerging non-volatile memories is reported in Table 1.1.

Table 1.1: Comparison between conventional CMOS-based and emerging non-volatile memories [2].

	Typical Memories			Emerging Memories				
Parameters	SRAM	DRAM	Flash	FeRAM	ReRAM	PCRAM	STT-MRAM	SOT-MRAM
Non-Volatility	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Cell Size F^2	$50-120$	$6-10$	5	$15-34$	$6-10$	$4-19$	$6-20$	$6-20$
Read Time (ns)	≥ 2	30	10^3	≥ 5	$1-20$	≈ 2	$1-20$	≥ 10
Write Time (ns)	≥ 2	50	10^6	≈ 10	50	10^2	≈ 10	≥ 10
Write Power	*Low*	*Low*	*High*	*Low*	*Medium*	*Low*	*Low*	*Low*
Endurance (cycles)	10^{16}	10^{16}	10^5	10^{12}	10^6	10^{10}	10^{15}	10^{15}
Future Scalability (ns)	*Good*	*Limited*	*Limited*	*Limited*	*Medium*	*Limited*	*Good*	*Good*

Conventional computing systems are based on the Von Neumann architecture that mainly consists in two separate block for processing units (CPUs) and memories, respectively. Accordingly, the data transfer between CPU and memory units through a bus causes a massive overhead in terms of both performance and energy consumption (Fig. 1.1a). This phenomenon is called Von Neumann Bottleneck (VNB), which is emphasized with technology scaling due to the fact that the performance of CPUs and memories differently scales [3]. More specifically, CPU doubles its performance every two years, while memory performance doubles every ten years (Fig. 1.1b).

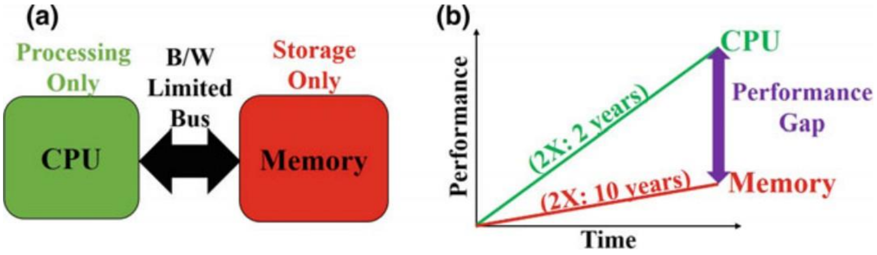


Figure 1.1: (a) CPU and memory connected through a bandwidth-limited bus for data transfer. (b) Performance improvement for CPU and memory in time [3].

Different approaches have been recently introduced to address the processor-memory data transfer bottleneck in computing systems (Fig. 1.2) [4]:

- <u>Computation-near-Memory</u> (Fig. 1.2a): thanks to the 3D structure of integrated circuit technologies, there is the possibility to stack CPU and memory circuits closer together in order to reduce the length of the connections, thus broadening the bandwidth; however, computation modules and memories are still built on two different blocks;

- Computation-in-Memory (Fig. 1.2b): the memory structure takes advantage of its intrinsic analog changes, e.g. the resistivity changes in emerging memory devices, to perform computation during the reading operation within the memory block; this approach is typically implemented by exploiting emerging resistive non-volatile memories like STT-MRAMs and SOT-MRAMs;
- Computation-with-Memory (Fig. 1.2c): in this case, the memory is used as a Content-Addressable-Memory (CAM) where obtained results are retrieved by a Look Up Table (LUT); the operating principle of this approach is that, by storing their truth tables, logical operations that involve more than one input can be directly encoded in the memory; the results are kept in the CAM, while the inputs are kept in the LUT, which can be accessed through an input combination giving a specific address;
- Logic-in-Memory (Fig. 1.2d): in this case, the logic is directly integrated in the memory cell, without requiring to extract stored data from the memory array.

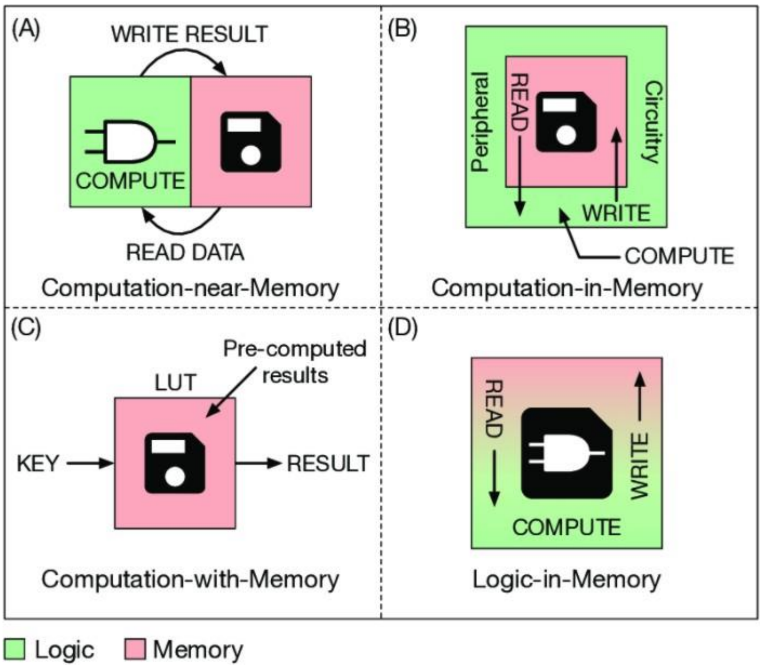


Figure 1.2: Four approaches to deal with the Von Neumann Bottleneck [4].

Computing-in-Memory (CiM) is regarded as a promising approach to deal with the VNB by reducing the number of memory accesses and the amount of data transferred between processor and memory. As stated before, it is particularly suitable for emerging resistive non-volatile memories where the simultaneous enabling of multiple bitcells in the memory array allows directly computing logic functions of the stored bits.

In the IoT era, security is also becoming a very crucial aspect. In this regard, hardware authentication represents a promising solution to enhance the security of physical devices. In order to provide a secure authentication, hardware cryptographic operations having a secret key stored in non-volatile memory devices can be used. In particular, physical unclonable functions (PUFs) are recently going to play a key role to enhance the security of electronic devices with minimal additional hardware cost [5].

PUFs are innovative primitives that implement a chip-unique challenge (C)–response (R) mechanism by typically exploiting the intrinsic randomness of electronic devices related to manufacturing process variability [5]. Such randomness ensures non-replicable code outputs. Therefore, by stimulating the PUF circuit with a challenge, a corresponding response is produced, and this challenge-response pairing (CRP) behavior is device-specific and difficult to predict.

There are two main types of PUFs [6]: “weak PUFs” and “strong PUFs”. The former store the secret key in a potentially vulnerable hardware, while the latter implement more complex challenge–response mechanisms from the physical disorder characterizing the PUF device. The most common implementation of weak PUFs is the static random access memory (SRAM) based PUF. Conversely, typical implementations of strong PUF are based on optical scattering. Recently, the major semiconductor foundries have integrated spintronic technology within the standard CMOS process. This opens a route to exploit spintronic technology in designing possible implementations of high-security PUF. Indeed, CMOS-based commercial PUFs exhibit several problems, which typically affect their performance and reliability, related to environmental (e.g. temperature) and operative (e.g. supply voltage) variations. Spintronic technology can be then used to deal with some of these problems.

Recently, a memory-based PUF implementation exploiting three-terminal SOT-MRAM devices has been proposed in [7] with possible advantages over the STT-MRAM based PUF in terms of robustness against temperature and device-to-device variations. In the proposed solution, the secret key of the PUF is stored into a random state of an array of SOT-MRAM devices consisting of a Magnetic Tunnel Junction (MTJ) with a perpendicular free layer (FL) placed on the top of a heavy metal (HM) strip. Such random state is achieved by applying a large enough current into the HM, which drives the FL magnetization along the in-plane direction, as shown in Fig. 1.3. Once the current is removed, due to the effect of the stochastic thermal field, the FL magnetization evolves to either up or down perpendicular direction with quite similar probability close to 50%, thus giving rise to write randomly a bit “0” or “1” in the MTJ.

In the above context, this thesis focuses on the design of a PUF exploiting the SOT-MRAM based solution proposed in [7]. The designed circuit incorporates a CiM strategy [8] to implement bitwise XOR logic operations within the memory array to enhance the quality and security of the PUF response.

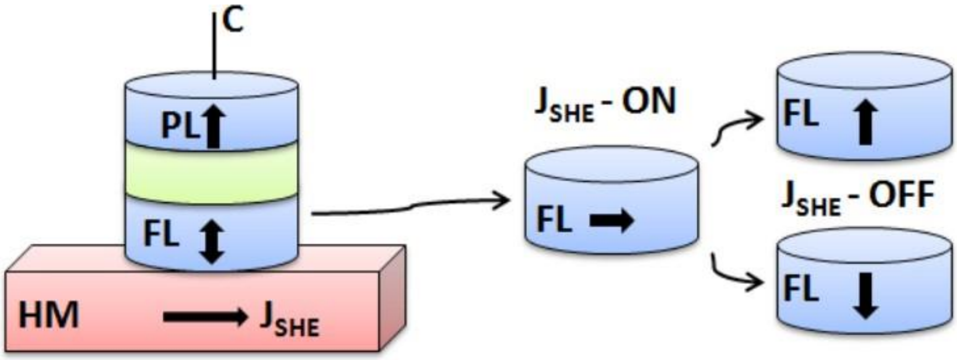


Figure 1.3: Operating principle of the SOT-MRAM based PUF proposed in [7].

The PUF circuit has been designed into a commercial circuit design tool (Cadence Virtuoso) by using a hybrid CMOS/spintronics modeling approach [9-12]. For CMOS devices, transistor models provided by a 0.8V/1.8V 18-nm FinFET technology [13] have been considered. For three-terminal SOT-MRAM devices, a Verilog-A based compact model [14-15] has been employed. Electrical and statistical simulations have been then performed to evaluate energy characteristics and security metrics of the designed PUF circuit.

The rest of this thesis is organized as follows. Chapter II introduces STT- and SOT-MRAM spintronic devices. Chapter III discusses and investigates a CiM solution for STT- and SOT-MRAMs. Then, Chapter IV details the implemented SOT-MRAM based PUF circuit along with the discussion of obtained simulation results. Finally, Chapter V summarizes the main conclusions of this work.

II. STT-MRAM AND SOT-MRAM DEVICES

This chapter briefly introduces the fundamental principles of Magnetic Tunnel Junctions (MTJs), which are basic key devices for spintronic Spin-Transfer-Torque (STT)- and Spin-Orbit-Torque (SOT)-Magnetic Random Access Memories (MRAMs). STT and SOT switching mechanisms used for the writing operation are discussed, along with the implementation of the reading operation. Finally, the architectural organization of STT- and SOT-MRAM array is described.

2.1 Magnetic Tunnel Junction (MTJ)

An MTJ is basically composed of three fundamental layers (Fig. 2.1): two ferromagnetic (FM) layers separated by an extremely thin oxide barrier [16]. One FM layer, namely Pinned or Reference Layer (PL or RL), has a fixed magnetization orientation through the use of an antiferromagnetic layer (AFM), while the other FM layer, namely Free Layer (FL), has a variable magnetization orientation. The relative orientation, i.e. parallel (P) or antiparallel (AP), of the magnetization of the two FM layers leads to two different states corresponding to two different resistance values (low resistance R_P in P state and high resistance R_{AP} in AP state) owing to tunnel

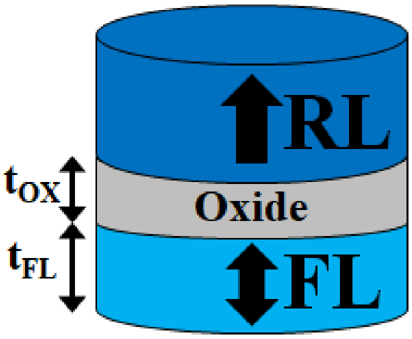


Figure 2.1: Basic structure of an MTJ with perpendicular magnetic anisotropy (PMA).

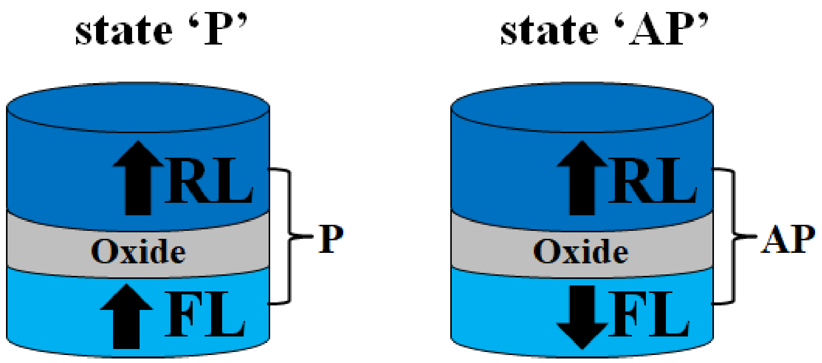


Figure 2.2: Two different resistance states in an MTJ with PMA.

magnetoresistance (TMR) effect (Fig. 2.2), where the TMR ratio is given by:

$$TMR = \frac{R_{AP} - R_P}{R_P} \tag{2.1}$$

One of the main features of the MTJs is the non-volatility, i.e. the capability of retaining stored data for a long time. The parameter quantifying such capability is the thermal stability (Δ), defined as

$$\Delta = \frac{E_b}{k_B T} \tag{2.1}$$

where E_b is the energy barrier between P and AP states, k_B is the Boltzmann constant, and T is the FL temperature [10]. Higher Δ, better the capability to retain the stored bit. Typically, a value of Δ larger than $60k_BT$ at 300 K is the commercial requirement to guarantee a retention time of 10 years.

There are two main types of MTJs: (*i*) MTJs with magnetization orientation in the in-plane (IP) direction, i.e. parallel to the film plane, and (*ii*) MTJ with out-of-plane magnetization orientation, i.e. perpendicular to the film plane (as in Figs. 2.1 and 2.2). In this thesis, only perpendicular MTJs (p-MTJs) will be considered in view of their well-known performance superiority compared to the in-plane counterparts [11].

Different mechanism can be exploited to switch the FL magnetization from one state to the opposite, thus enabling the writing of a bit "0" or "1" in the MTJ. In the following, STT- and SOT-based switching mechanisms will be introduced and discussed.

2.2 STT and SOT switching

The current-induced STT switching process consists of applying a current with enough large pulse amplitude (i.e. above a critical value) and duration through the MTJ stack [9]. The direction of the applied current ensures both switching events, i.e. from P to AP state (P→AP) and from AP to P state (AP→P). In particular, P→AP switching occurs when the current is applied to the RL toward the FL. Conversely, AP→P switching occurs when the current is applied to the FL toward the RL. Nowadays, two-terminal MTJs based on STT switching are regarded as one of the most promising candidates for the next generation of Systems on Chip with on-chip non-volatile memory implemented at nanoscaled technological nodes. STT-based writing operation has been proven to be enough fast, robust and reliable. Nevertheless, two-terminal STT-MTJ devices exhibit some drawbacks. The stochastic nature of the STT switching resulting from the effect of the unavoidable thermal fluctuations on the FL magnetization represents one of the most important challenges. This phenomenon is responsible for large fluctuations in the switching time of STT-MTJs, which can deeply affect the reliability of the writing operation [9]. The STT switching process across the two switching transitions (i.e. P→AP and AP→P) is also asymmetric. In addition, two-terminal STT-MTJs also suffer from the shared writing and reading paths, which can

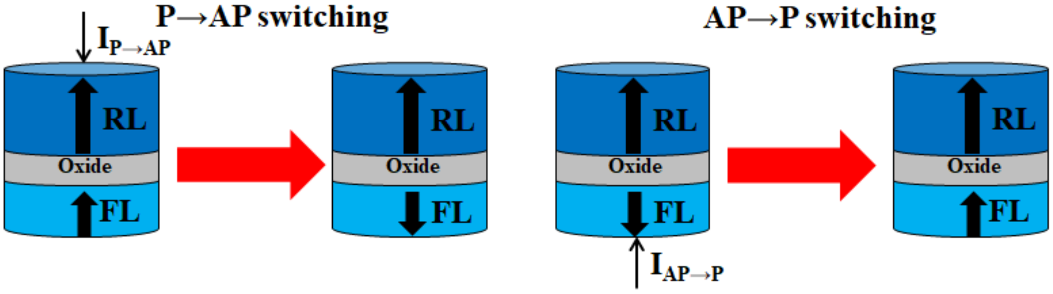


Figure 2.3: STT switching in a two-terminal MTJ-based device.

cause unwanted switching during the reading operation [16].

To deal with the above challenges, alternative switching methods have been explored. Recently, some experiments have demonstrated the SOT switching mechanism mainly related to the Spin-Hall Effect (SHE) in an MTJ-based device (Fig. 2.4) [14]. This consists of a three-terminal structure, where an MTJ is a placed on the top of a heavy metal (HM) strip. In particular, the HM is attached to the FL of the MTJ. Owing to the SHE, an in-plane write current flowing through the HM strip can produce an enough large spin torque to enable the switching of the FL magnetization. There are some advantages of the SOT-based three-terminal devices with respect to conventional STT-based two-terminal counterparts. One of these is related to the fact that the writing current flows through the HM instead of passing through the MTJ, thus giving rise to separate writing and reading paths. This prevents barrier breakdown (i.e. long endurance) and it also enables separate optimization of reading and writing operations. In addition, SOT switching also allows higher speed and lower energy writing than conventional STT switching [14]. However, there are also some challenges and

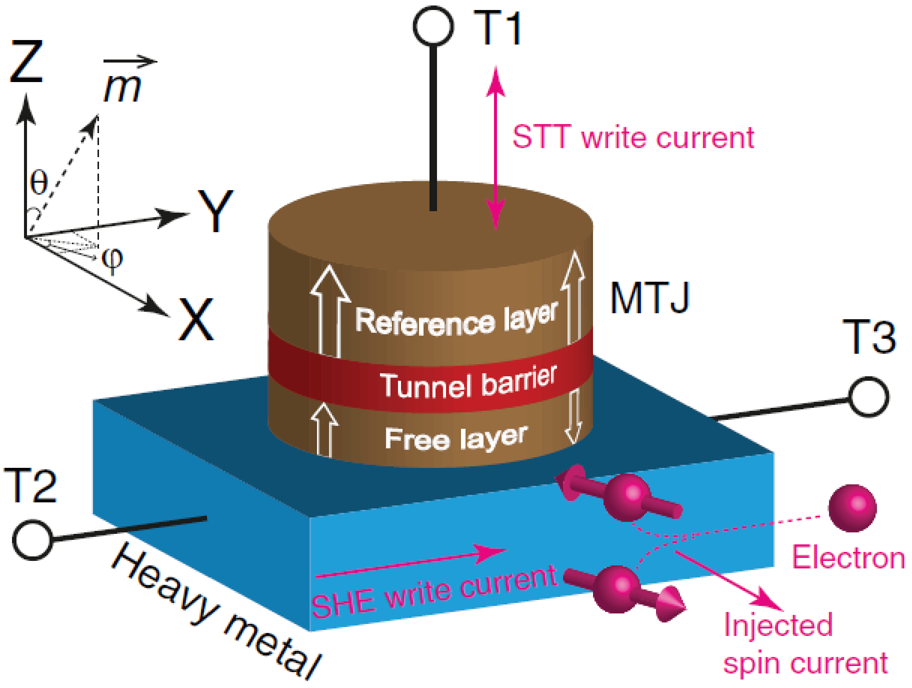


Figure 2.4: SOT switching in a three-terminal MTJ-based device [14].

limitations in SOT-based three-terminal devices, especially for structures with p-MTJs. Indeed, in these devices an additional magnetic field is typically required to achieve deterministic switching since the spin torque originated by the applied in-plane current cannot allow for a stable perpendicular magnetization [14]. The applying of such an external magnetic field significantly limits the technological effectiveness of SOT-based devices. Therefore, additional efforts to find ways to eliminate the need for the external field have to be made for perpendicular SOT-based devices.

2.3 Reading operation

The reading operation is implemented in the same way for STT- and SOT-based devices. Indeed, in both cases the stored bit can be read through the resistance of the MTJ. In this regard, two different sensing scheme can be adopted (Fig. 2.5): (a) a current sensing (CS) scheme consisting of applying a read voltage (V_{read}) across the MTJ and then comparing the generated sensing current (I_{sense}) with a reference current (I_{ref}) by a current-mode sensing amplifier; (b) a voltage sensing (VS) scheme consisting of applying a read current (I_{read}) through the MTJ and then comparing the generated sensing voltage (V_{sense}) with a reference voltage (V_{ref}) by a voltage-mode sensing

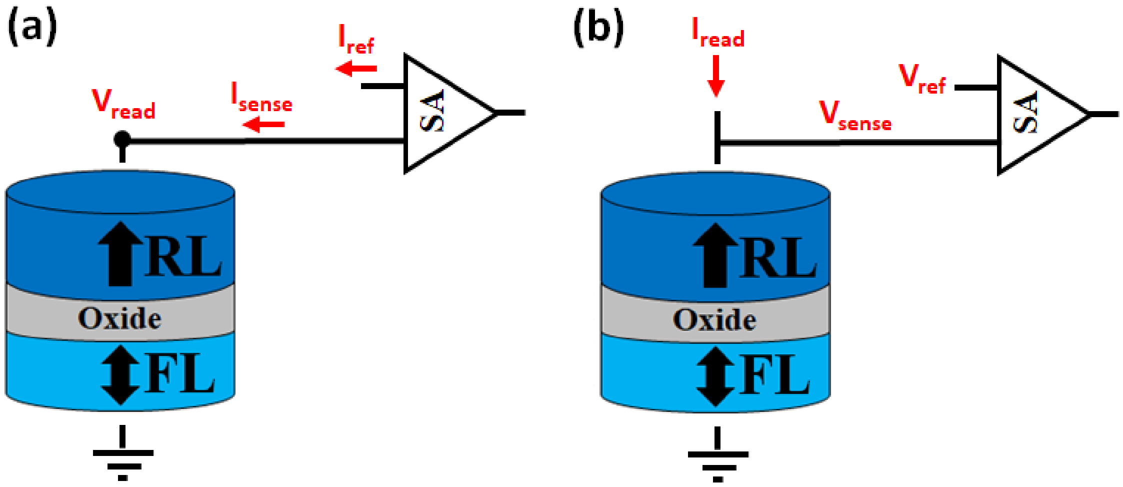


Figure 2.5: Different sensing schemes: (a) current sensing (CS) scheme and (b) voltage sensing (VS) scheme.

amplifier. In both cases, the current flowing through the MTJ has to be sufficiently low to avoid any disturbing of the stored data during the reading operation.

2.4 STT-MRAM and SOT-MRAM array

Fig. 2.6 shows the typical organization of an STT-MRAM array along with the detail of the memory bitcell. The latter consists of a conventional one access transistor (1T)- one MTJ structure [10]. Bitcells within the same row share the word-line (WL), which drives the access transistors. Conversely, bitcells within the same column share the bit-line (BL) and the source-line (SL). The writing operation is performed per row by enabling the corresponding WL to switch on the access transistors, while rising the BL (SL) to the supply voltage (V_{DD}) and grounding the SL (BL). The reading operation is also performed per row by enabling the corresponding WL and implementing a CS or VS scheme via the BL and the SL.

Fig. 2.7 shows the typical organization of an SOT-MRAM array along with the detail of the memory bitcell. Here, each bitcell is composed by an SOT-MRAM device along with two access transistors. The write access transistor connecting the HM to the write bit-line (WBL) is driven by the write word-line (WWL) shared by BCs within the same row. The read access transistor connecting the MTJ to the read bit-line (RBL) is

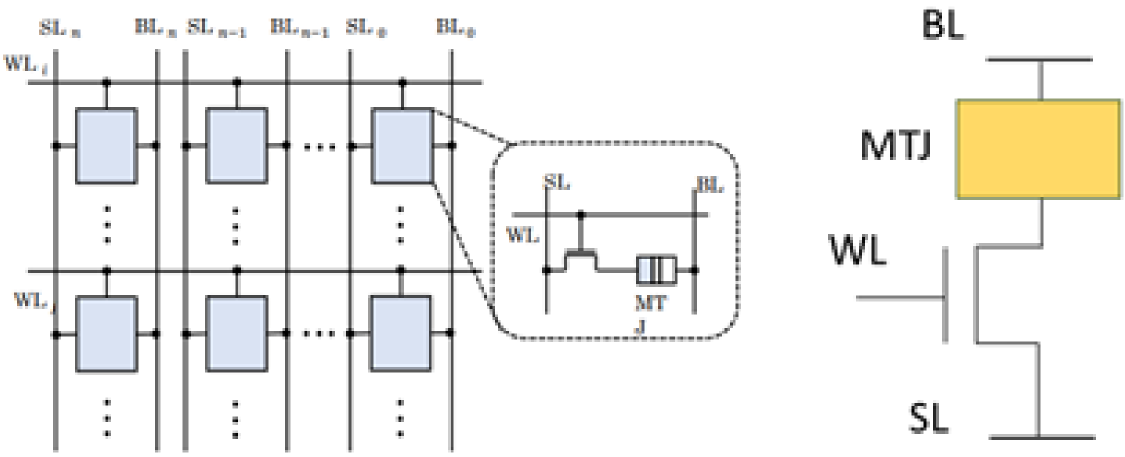


Figure 2.6: STT-MRAM array with the detail of the 1T-1MTJ bitcell.

driven by the read word-line (RWL) shared by BCs within the same row. BCs within the same column share the WBL, the RBL and the SL. Again, both writing and reading operation are performed per row. In particular, during the writing operation, the corresponding WWL is raised high, while disabling the RWL. Then, a current passing from the WBL to the SL flows through the HM strip to enable the SOT-based switching mechanism. On the contrary, for the reading operation the corresponding RWL is raised high, while disabling the WWL. Then, a CS or VS scheme is implemented via the RBL and the SL through the read access transistors.

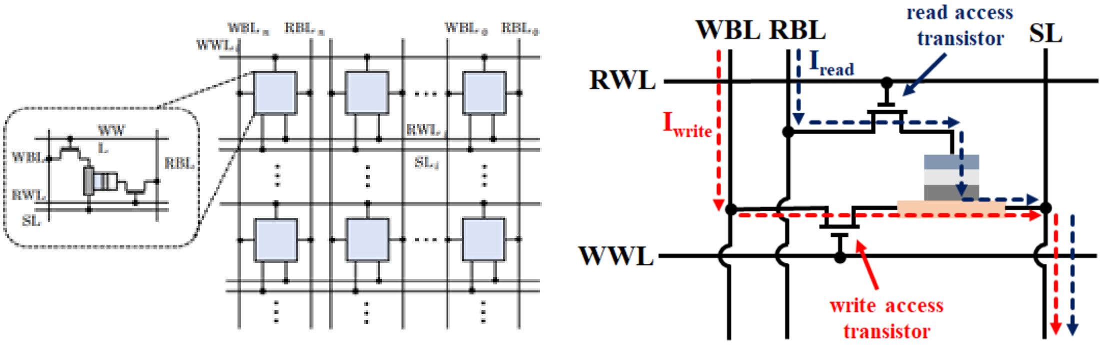


Figure 2.7: SOT-MRAM array with the detail of the 2T-bitcell.

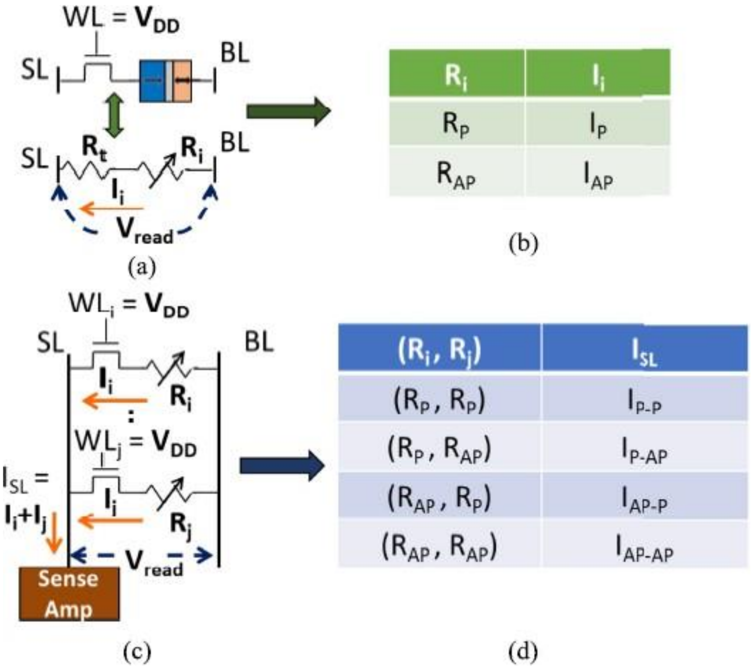
III. COMPUTING-IN-MEMORY (CiM) WITH STT- AND SOT-MRAM

In this chapter, a Computing-in-Memory (CiM) solution that can be exploited in STT-MRAM and SOT-MRAM based circuits is described. In particular, the design of enhanced sensing circuits to support bitwise logic operations within the memory array is discussed,

3.1 CiM strategy

As stated in Chapter I, CiM is motivated by the fact that the movement of data from memory to the CPU and back is a major bottleneck in terms of both performance and energy consumption in modern computing systems. An interesting CiM solution for STT-MRAMs has been recently presented in [8]. The basic idea behind the proposed approach is to activate multiple word-lines (WLs) at the same time in an STT-MRAM array and sensing the effective resistance of each bitline (BL) that is connected to the multiple activated bitcells. In this way, basic logic operations of the bits stored in the memory bitcells can be directly computed within the memory array, without the need to transfer data in a dedicated computing block. It is worth pointing out that, although the considered CiM scheme has been originally proposed for STT-MRAMs [8], such strategy can be also efficiently applied in SOT-MRAMs considering that the reading operation is basically the same in the two different memories.

Fig. 3.1 describes the operating principle of the considered CiM solution referring to a CS scheme. More specifically, Fig. 3.1a the corresponding resistive equivalent



R_i	I_i
R_P	I_P
R_{AP}	I_{AP}

(R_i, R_j)	I_{SL}
(R_P, R_P)	$I_{P\text{-}P}$
(R_P, R_{AP})	$I_{P\text{-}AP}$
(R_{AP}, R_P)	$I_{AP\text{-}P}$
(R_{AP}, R_{AP})	$I_{AP\text{-}AP}$

Figure 3.1: Operating principle of the considered CiM solution referred to a current sensing (CS) scheme [8]: (a) bitcell resistive equivalent, (b) bitcell sensing current depending on the MTJ state, (c) bitwise CiM operation between two bitcells, and (d) source-line (SL) sensing current resulting from bitwise CiM operation.

circuit of a single activated STT-MRAM 1T-1MTJ bitcell, where R_t is the On resistance of the access transistor and R_i represents the resistance offered by the MTJ. Considering a CS scheme, a read voltage (V_{read}) is applied across the MTJ (i.e. between the BL and the SL). Accordingly, the sense current I_i flowing through the bitcell can exhibit two possible values depending on the MTJ state, i.e. I_P if the MTJ is in P state or I_{AP} if the MTJ is in AP state (Fig. 3.1b). Therefore, the use of a current-mode sensing amplifier (SA) and a reference current (I_{ref}) allows distinguishing these two different current values, thus detecting the bit stored in the activated bitcell. Note that, in the following, the bit ‘0’ will be associated while the AP state, while the bit “1” will be associated with the P state. On the other hand, Fig. 3c describes the implementation of a CiM operation by enabling simultaneously two WLs (i.e. WL_i and Wl_j), while applying a V_{read} between the BL and the SL. The resulting current flowing towards the SL (I_{SL}) thus depends on both logic states stored in the two activated bitcells. In particular, I_{SL} is given by the sum of the currents flowing through each of the two bitcells (i.e. I_i and

I_j). The possible values of I_{SL} depending on the states of the MTJs belonging to the two activated bitcells (i.e. $I_{P\text{-}P}$, $I_{P\text{-}AP}$, $I_{AP\text{-}P}$, and $I_{AP\text{-}AP}$) are reported in Fig. 3.1d. Therefore, by exploiting an enhanced sensing circuitry (Fig. 3.2), it is possible to distinguish these different values of I_{SL}, thus performing basic logic operations between the bits stored in the enabled bitcells as follows:

- bitwise OR (NOR): the implementation of logic OR and NOR operations requires the sensing scheme shown in Fig. 3.2a, where I_{SL} is connected to the positive input of the SA and a reference current ($I_{ref\text{-}or}$) is connected to the negative input. To ensure the correct operation, $I_{ref\text{-}or}$ has to be chosen as an intermediate value between $I_{AP\text{-}AP}$ and $I_{AP\text{-}P}$ (equal to $I_{P\text{-}AP}$) as shown in Fig. 3.2c. In this way, among the four possible values of I_{SL} (Fig. 3.1d), only the case with $I_{SL} = I_{AP\text{-}AP}$ leads to a sensing current lower than $I_{ref\text{-}or}$. Accordingly, only the case where both the activated bitcell are in AP state (i.e. both store a bit “0”), corresponds to an output equal to “0” (“1”) for the OR (NOR) operation. All the other cases leads to an output equal to “1” (“0”) for the OR (NOR) operation. Thus, the positive and negative outputs of the SA compute the logic OR and NOR, respectively, of the bits stored in the enabled bitcells [8].

- bitwise AND (NAND): the implementation of logic AND and NAND operations requires the sensing scheme shown in Fig. 3.2b, where I_{SL} is connected to the positive input of the SA and a reference current ($I_{ref\text{-}and}$) is connected to the negative input. To ensure the correct operation, $I_{ref\text{-}and}$ has to be chosen as an intermediate value between

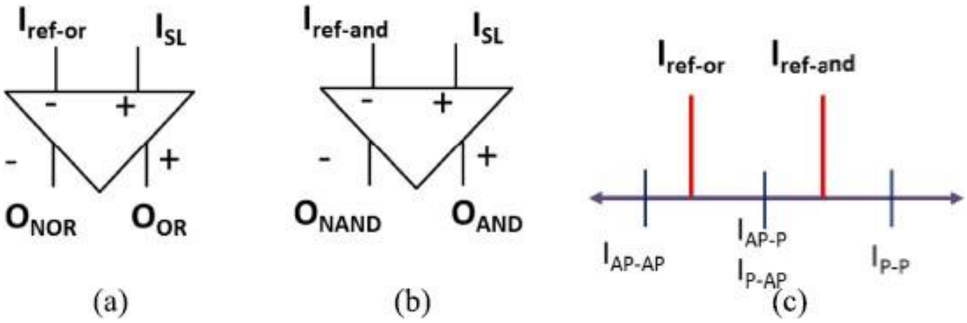


Figure 3.2: Current sensing schemes for bitwise (a) OR and (b) AND logic operations, along with the detail of the required reference currents [8].

$I_{AP\text{-}P}$ (equal to $I_{P\text{-}AP}$) and $I_{P\text{-}P}$ as shown in Fig. 3.2c. In this way, among the four possible values of I_{SL} (Fig. 3.1d), only the case with $I_{SL} = I_{P\text{-}P}$ leads to a sensing current higher than $I_{ref\text{-}and}$. Accordingly, only the case where both the activated bitcell are in P state (i.e. both store a bit "1"), corresponds to an output equal to "1" ("0") for the AND (NAND) operation. All the other cases leads to an output equal to "0" ("1") for the AND (NAND) operation. Thus, the positive and negative outputs of the SA compute the logic AND and NAND, respectively, of the bits stored in the enabled bitcells [8].

- bitwise XOR (NAND): using the two sensing schemes of Fig. 3.2, the bitwise XOR operation can be implemented by exploiting a 2-bit CMOS NOR gate whose inputs corresponds to O_{AND} and O_{NOR} (Fig. 3.3) [8]. This because $O_{XOR} = O_{AND}$ NOR O_{NOR}.

The considered CiM solution can be also implemented with reference to a VS

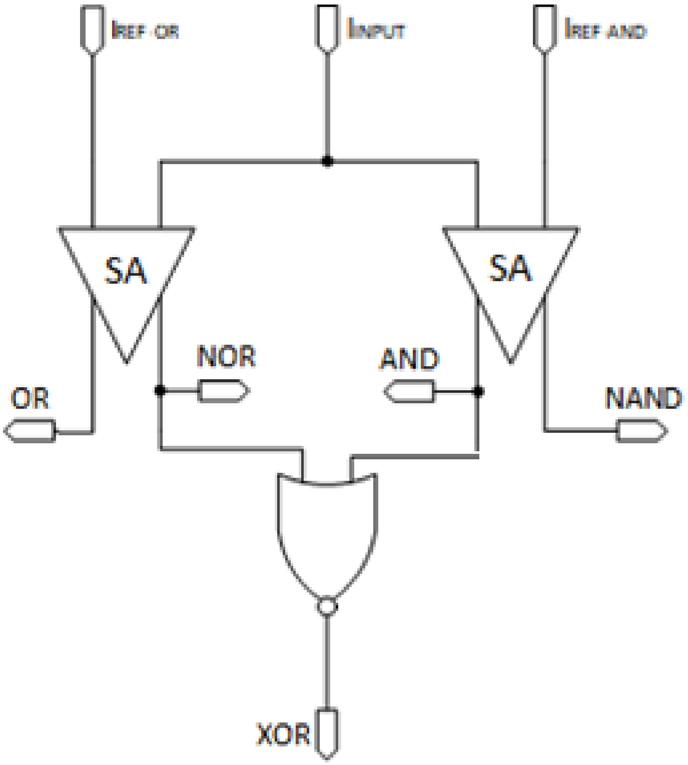


Figure 3.3: Current sensing scheme for bitwise XOR operation.

scheme. This occurs again by enabling simultaneously two WLs, while applying an I_{read}

to the BL (the SL is grounded). The resulting BL voltage (V_{BL}) thus depends on both logic states stored in the two activated bitcells. In particular, V_{BL} is given by the I_{read} multiplied by the equivalent resistance corresponding to the parallel between the resistances associated with the two activated bitcells (i.e. R_i and R_j). The possible values of V_{BL} depending on the states of the MTJs belonging to the two activated bitcells (i.e. V_{P-P}, V_{P-AP}, V_{AP-P}, and V_{AP-AP}) are reported in Table 3.1.

Table 3.1: Possible values of bit-line (BL) sensing voltage resulting from bitwise CiM operation.

(R_i, R_j)	**V_{BL}**
(R_P, R_P)	V_{P-P}
(R_P, R_{AP})	V_{P-AP}
(R_{AP}, R_P)	V_{AP-P}
(R_{AP}, R_{AP})	V_{AP-AP}

Therefore, by exploiting an enhanced sensing circuitry (Fig. 3.4), it is possible to distinguish these different values of V_{BL}, thus performing basic logic operations between the bits stored in the enabled bitcells as follows:

- bitwise OR (NOR): the implementation of logic OR and NOR operations requires the sensing scheme shown in Fig. 3.4a, where V_{BL} is connected to the negative input of the SA and a reference voltage (V_{ref-or}) is connected to the positive input. To ensure the correct operation, V_{ref-or} has to be chosen as an intermediate value between V_{AP-P} (equal

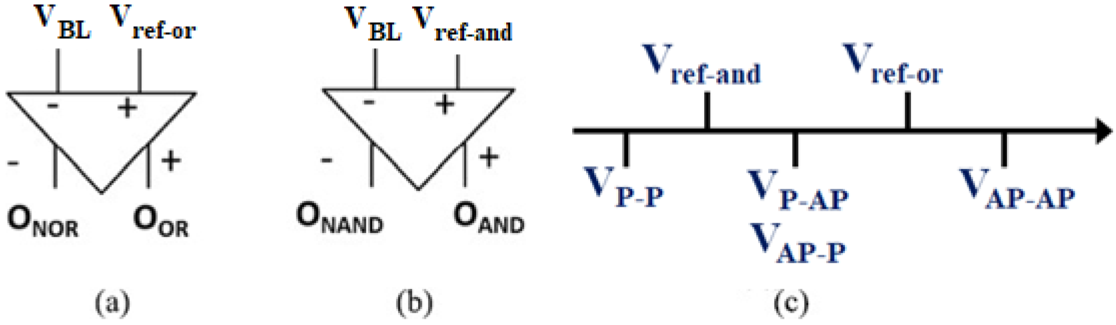


Figure 3.4: Voltage sensing schemes for bitwise (a) OR and (b) AND logic operations, along with the detail of the required reference voltages.

to $V_{P\text{-}AP}$) and $V_{AP\text{-}AP}$ as shown in Fig. 3.4c. In this way, among the four possible values of V_{BL} (Table 3.1), only the case with $V_{BL} = V_{AP\text{-}AP}$ leads to a sensing voltage higher than $V_{ref\text{-}or}$. Accordingly, only the case where both the activated bitcell are in AP state (i.e. both store a bit “0”), corresponds to an output equal to “0” (“1”) for the OR (NOR) operation. All the other cases leads to an output equal to “1” (“0”) for the OR (NOR) operation. Thus, the positive and negative outputs of the SA compute the logic OR and NOR, respectively, of the bits stored in the enabled bitcells.

- bitwise AND (NAND): the implementation of logic AND and NAND operations requires the sensing scheme shown in Fig. 3.4b, where V_{BL} is connected to the negative input of the SA and a reference voltage ($V_{ref\text{-}and}$) is connected to the positive input. To ensure the correct operation, $V_{ref\text{-}and}$ has to be chosen as an intermediate value between $V_{P\text{-}P}$ and $V_{AP\text{-}P}$ (equal to $V_{P\text{-}AP}$) as shown in Fig. 3.4c. In this way, among the four possible values of V_{BL} (Table 3.1), only the case with $V_{BL} = V_{P\text{-}P}$ leads to a sensing voltage lower than $V_{ref\text{-}and}$. Accordingly, only the case where both the activated bitcell are in P state (i.e. both store a bit “1”), corresponds to an output equal to “1” (“0”) for the AND (NAND) operation. All the other cases leads to an output equal to “0” (“1”) for the AND (NAND) operation. Thus, the positive and negative outputs of the SA compute the logic AND and NAND, respectively, of the bits stored in the enabled bitcells.

- bitwise XOR (NAND): following the same scheme used for the CS scheme (see Fig. 3.2), again the bitwise XOR operation can be implemented by exploiting a 2-bit CMOS NOR gate whose inputs corresponds to O_{AND} and O_{NOR} (Fig. 3.4).

3.2 Circuitry for voltage sensing scheme

To implement the above-described CiM solution referred to the VS scheme (Fig. 3.4), a latch-type voltage SA design has been employed [17]. Latch-type SAs are typically exploited to read the contents of several types of memory (e.g. SRAM) since they ensures fast operations thanks to the strong positive feedback [17]. Fig. 3.5 shows the considered conventional latch-type voltage SA, where two cross-coupled inverters (M1-M4) provide the positive feedback, while the enable signal EN 5 starts the sensing operation. This SA combines a high input impedance and positive feedback. The current flow depending on the differential inputs that are applied on M5 and M6 controls the serially-connected latch circuit. A small difference between the current flowing through M5 and M6 converts to a large output voltage.

According to Fig. 3.4, the implementation of the CiM solution with a VS scheme also requires a proper circuit to generate the reference voltages and hence to enable CiM operations. In this regard, Fig. 3.6 shows the schematic of the designed reference generation circuit. It employs a PMOS-based cascode current mirror along with 4 MTJs and 4 access transistors (i.e. one for each MTJ). Two MTJs (both in P state) and two access transistors (ME1 and ME2) are exploited to generate the required voltage reference for the AND/NAND operation ($V_{ref\text{-}and}$). The other two MTJs (one in P state and the other in AP state) and two access transistors (ME3 and ME4) are then exploited

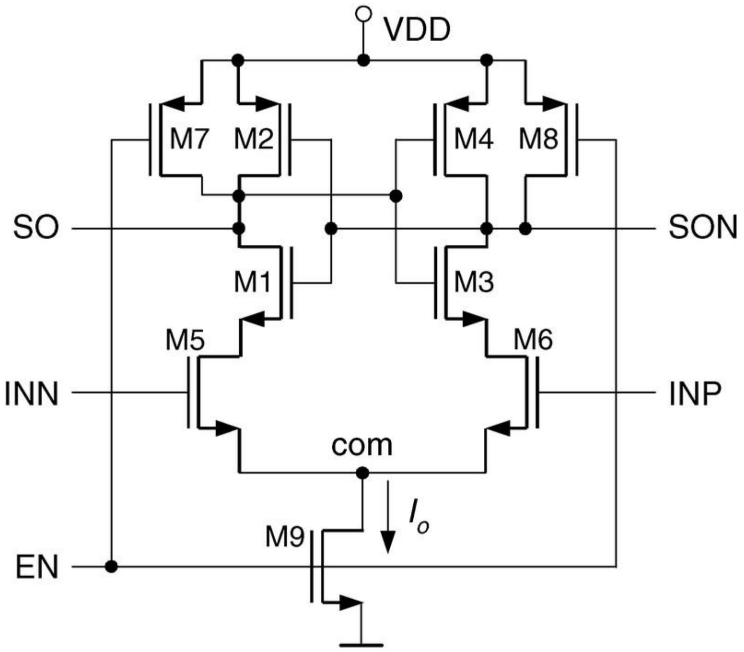


Figure 3.5: Latch-type voltage sensing amplifier [17].

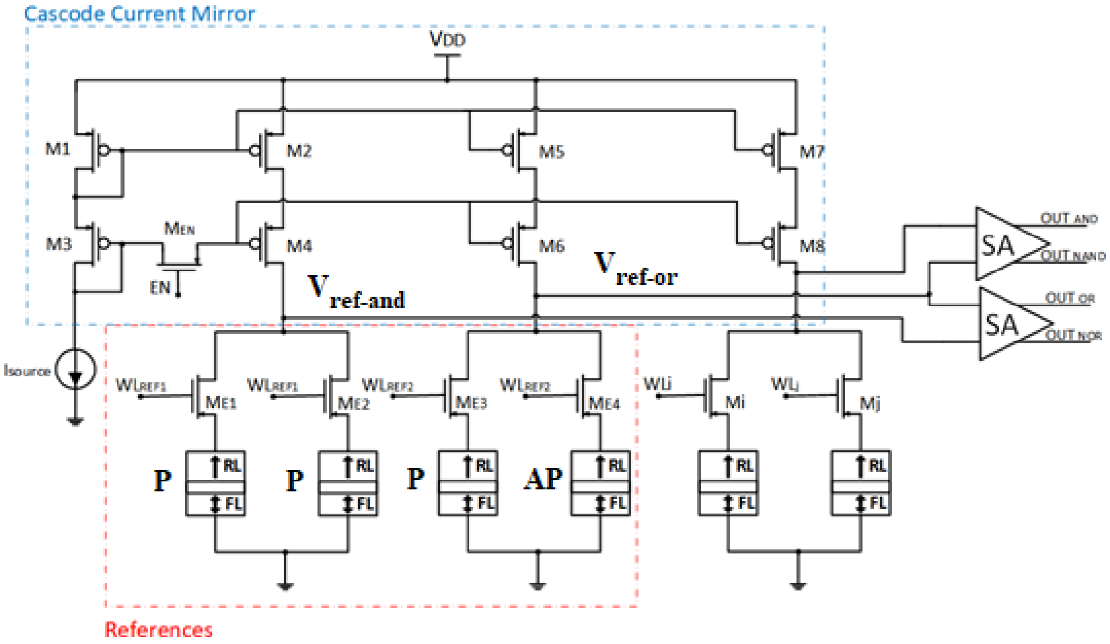


Figure 3.6: Circuits to generate reference voltages (i.e. $V_{ref\text{-}or}$ and $V_{ref\text{-}and}$) according to the CiM scheme of Fig. 3.4.

to generate the required voltage reference for the OR/NOR operation ($V_{ref\text{-}or}$). Fig. 3.7 shows an example of the two generated reference voltages in comparison with the BL voltage depending on the states of the MTJs belonging to the two activated bitcells (i.e. $V_{P\text{-}P}$, $V_{AP\text{-}P}$, and $V_{AP\text{-}AP}$). We can observe that the criteria to ensure correct operations are satisfied according to Fig. 3.4c.

Finally, Fig. 3.8 shows the timing diagram referred to the simulation of the whole voltage sensing circuitry (including the two SAs, the reference generation circuit, and the additional NOR gate) required to implement CiM logic operations.

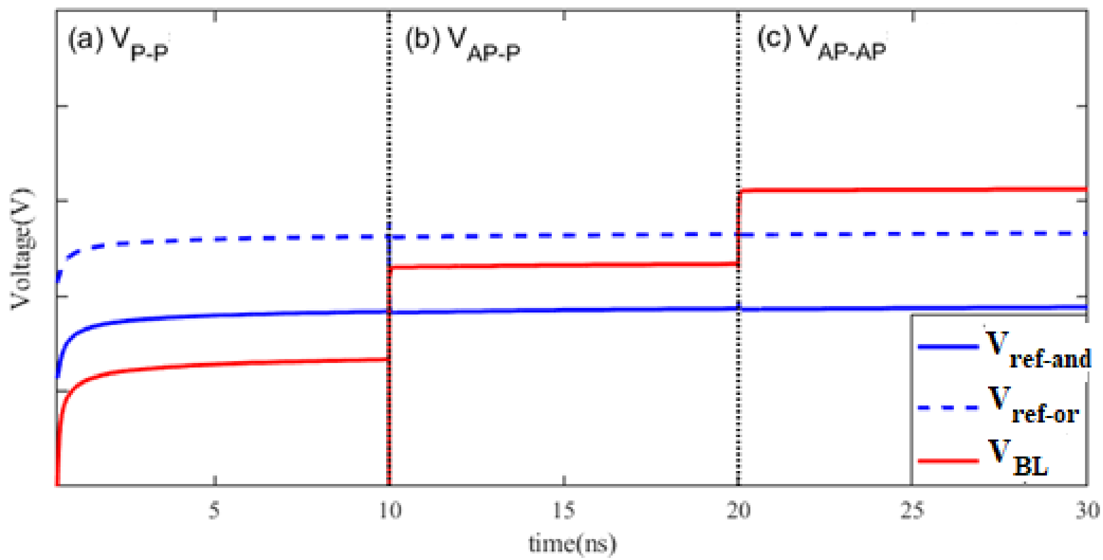


Figure 3.7: Comparison between reference voltages generated by the circuit of Fig. 3.6 and the BL voltage depending on the states of the MTJs belonging to the two activated bitcells: (a) $V_{P\text{-}P}$, (b) $V_{AP\text{-}P}$, and (c) $V_{AP\text{-}AP}$.

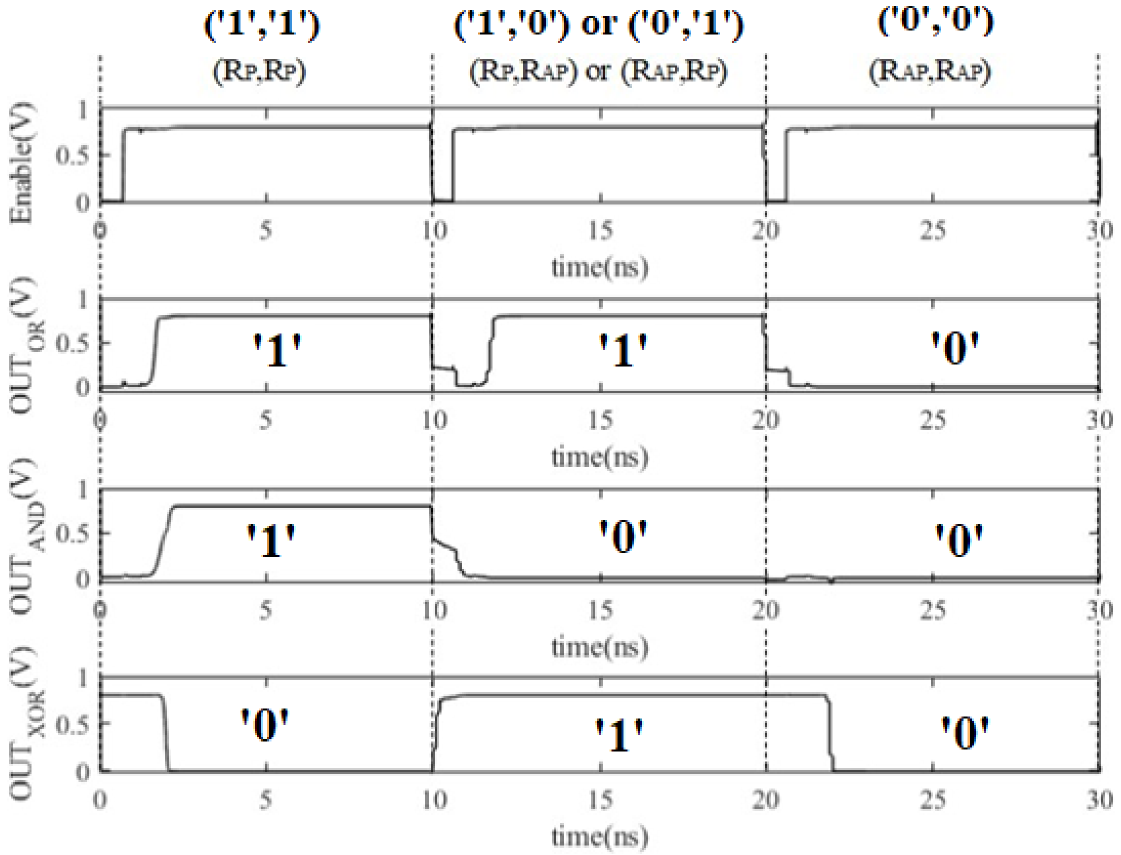


Figure 3.8: Timing diagram of the voltage sensing CiM solution. Here, the resistance states of the MTJs belonging to the two activated bitcells is changed every 20 ns to evaluate the output of the whole sensing circuitry for three different cases: R_P-R_P, R_P-R_{AP} (or R_{AP}-R_P), and R_{AP}-R_{AP}.

IV. SOT-MRAM BASED PUF WITH COMPUTING-IN-MEMORY

This chapter first outlines the typical metrics used to evaluate the quality of a PUF response. Then, the three-terminal SOT-MRAM device acting as a building block of the designed PUF implementation is presented. Finally, the SOT-MRAM based PUF circuit implemented by using a hybrid CMOS/spintronics modeling is described in detail, along with the presentation and the discussion of the results obtained from electrical and statistical simulations of the designed PUF circuit.

4.1 Evaluation metrics of PUF devices

Physically Unclonable Functions (PUFs) have recently emerged as an attractive technology for designing electronic systems with high security [18]. PUFs are innovative primitives that implement a chip-unique challenge (C)–response (R) mechanism to extract instance-specific secrets. In other words, the implemented C–R mechanism converts the unique physical state of the PUF into digital input-output data [7]. Specifically, by stimulating the PUF device with a challenge, a corresponding response is generated by the device. This challenge-response pairing (CRP) behavior is device-specific and difficult to predict or duplicate [5]. The term PUF has been firstly proposed in [19] where the authors introduced silicon-based PUFs, i.e. PUF devices realized using conventional integrated circuits (ICs). Silicon-based PUFs typically exploit the inherent randomness resulting from non-deterministic variations in the manufacturing process of ICs with identical masks to uniquely characterize each IC. However, it is worth pointing out that, in principle, a PUF could be built from any physical entity that ensures inherent randomness [18].

The most important metrics used to evaluate the quality of the response of a PUF implementation are uniformity (or randomness), uniqueness, and reliability [18]. The computation of these metrics is based on the concept of Hamming Distance (HD) and Hamming Weight (HW) [18]:

- Hamming Distance (HD): the Hamming Distance HD(a, b) between two words $a = (a_i)$ and $b = (b_i)$ of length n is defined as the number of positions where they differ, i.e. the number of (i)s such that $a_i \neq b_i$;

- Hamming Weight (HW): considering 0 as the zero vectors, the Hamming Weight HW(a) of a word $a = (a_i)$ is defined as HD(a, 0), i.e. the number of symbols $a_i \neq 0$.

According to the definition of HD and HW, we can define the three main evaluation metrics of PUF devices:

- Uniqueness: it is a measure of the ability of a device to generate unique identification, i.e. the ability of one PUF instance to have a uniquely distinguishable behavior compared with other PUFs with the same structure implemented on different chips [18]. This metric is evaluated using the Inter-chip Hamming Distance (HD_{INTER}). Let consider two chips, i and j ($i \neq j$), with n-bit responses, i.e. $R_i(n)$ and $R_j(n)$, respectively, for a specific challenge C. The average (or normalized) inter-chip HD among k chips is given by:

$$HD_{INTER} = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{HD\left(R_i(n), R_j(n)\right)}{n} \tag{4.1}$$

Fig. 4.1 shows an example of uniqueness evaluation, considering two PUF instances implemented on two different chips. When a challenge (011101) is applied on both instances, each PUF produces a different response. In the case of Fig. 4.1, the HD between the two PUF instances is 1 (i.e. only 14% of the total response bits are different. Ideally, the uniqueness, i.e. the normalized $HD_{INTER,}$ should be close to 0.5 or 50% if expressed as percentage.

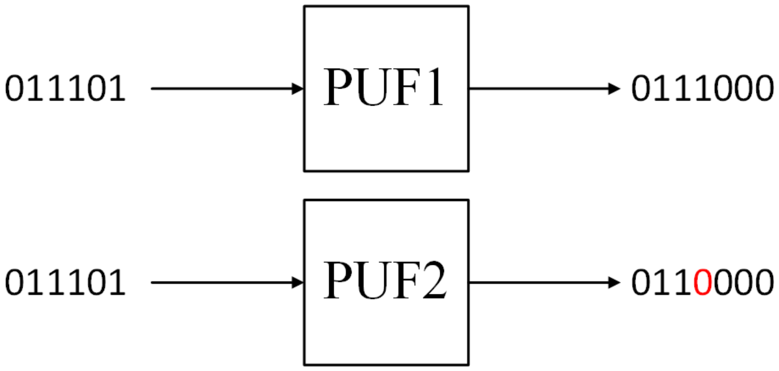


Figure 4.1: Example of PUF uniqueness evaluation.

- Uniformity or Randomness: it is a measure of the "unpredictability" of the PUF responses, which is related to the proportion of 0's and 1's in the PUF response. This metric is evaluated using the average (or normalized) Hamming Weight (HW) of the PUF responses as given by:

$$Uniformity = \frac{1}{k}\sum_{i=1}^{k} HW_i \tag{4.2}$$

where k is the total number of PUF responses and HW_i is the Hamming Weight of the i-th response. Ideally, the uniformity should be close to 0.5 or 50% if expressed as percentage for a truly random response [18].

- Reliability: it is a measure of the ability of the PUF to generate a consistent response *R* for a specific challenge *C*, regardless of any variations in the operating conditions such as the ambient temperature and voltage supply. This metric is evaluated using the Intra-chip Hamming Distance (HD_{INTRA}). Let consider a single chip i with a n-bit response $R_i(n)$ at nominal operating conditions and a n-bit response $R_i^{'}(n)$ at different conditions for the same challenge C. The average (or normalized) HD_{INTRA} for k chips is given by:

$$HD_{INTRA} = \frac{1}{k}\sum_{i=1}^{k} \frac{HD\left(R_i(n), R_i^{'}(n)\right)}{n} \tag{4.3}$$

Fig. 4.2 shows an example of reliability evaluation, considering a PUF instance

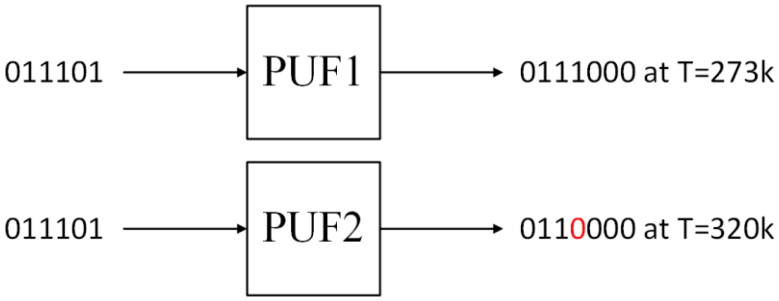


Figure 4.2: Example of PUF reliability evaluation.

operating at two different temperatures. When a challenge (011101) is applied at the two temperatures, ideally the PUF should produce the same response, thus expecting a zero HD between the two responses. In the case of Fig. 4.2., the HD_{INTRA} 1 (i.e. 14% difference from the room-temperature response). Ideally, the HD_{INTRA} should be close to zero, thus corresponding to a reliability close to 100%.

4.2 Three-terminal SOT-MRAM device

To build the PUF circuit, three-terminal SOT-MRAM devices consisting of a p-MTJ with the FL placed on a heavy metal (HM) strip have been considered [7]. Fig. 4.3a shows the sketch of the considered device along with the operating principle to write a random bit in the MTJ. Indeed, when a current (J_{SHE}) flows thorugh the HM, the SOT switching mechanism due to the Spin-Hall Effect (SHE) drives the FL magnetization along the in-plane direction [20]. Once the current pulse is removed, owing to the stochastic thermal field, the FL magnetization moves randomly to either positive (i.e. UP state corresponding to parallel P state) or negative (i.e. DOWN state corresponding to antiparallel AP state) perpendicular direction with quite similar probability close to 50%. This mechanism thus gives the possibility to randomly write a bit “0” (here corresponding to DOWN state) or “1” (here corresponding to UP state) into the MTJ. Then, the stored bit can be read by sensing the resistance of the MTJ using the current or voltage sensing schemes discussed in Chapter III. Figs. 4.3b and c

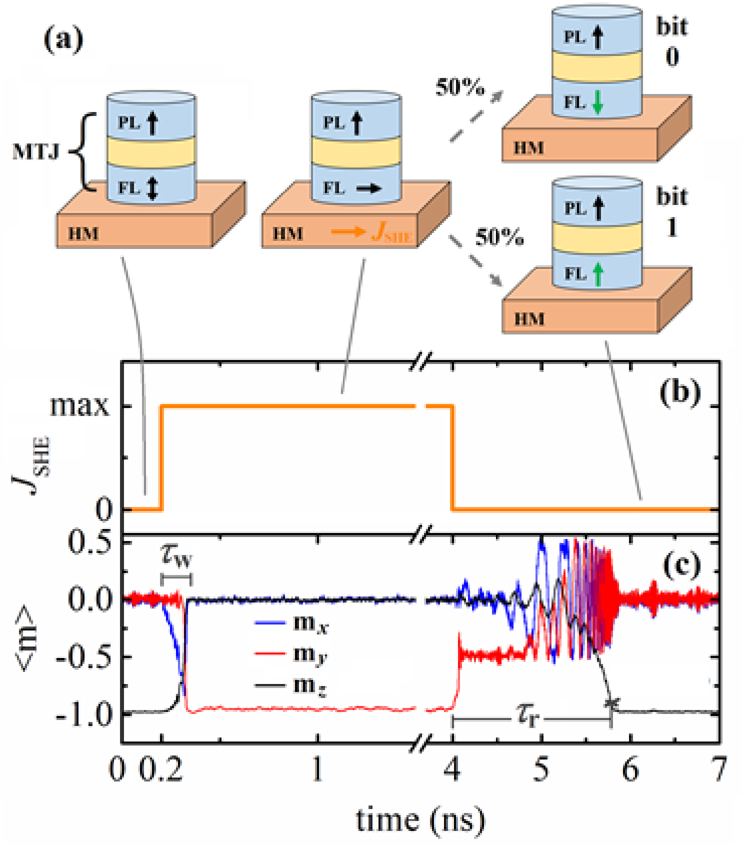


Figure 4.3: (a) Sketch of the SOT-MRAM device and operating principle to write a random bit thanks to the application of the current J_{SHE}. (b) Time description of the J_{SHE} current pulse. (c) Example of the time evolution of the FL magnetization.

show an example of the write current pulse and the time evolution of the FL magnetization, respectively. In particular, in Fig. 4.3c the initial state of the FL magnetization is DOWN (indeed, the z-component of the FL magnetization, i.e. m_z, is equal to -1 at 0 ns). When applying the current pulse, the FL magnetization reaches the in-plane direction (i.e. $m_z = 0$) within a time interval τ_w corresponding to hundreds of ps. Once the current pulse is switched off (at 4 ns in Figs. 4.3b and c), due to the effect of thermal fluctuations, the FL magnetization tends to randomly relax towards one of the two possible perpendicular configurations (i.e. UP or DOWN) [7] within a time interval τ_r corresponding to few ns. Therefore, depending on the initial and final state of the FL magnetization, we have four possible transitions, i.e. up-to-up, up-to-down, down-to-down, and down-to-up, each characterized by a specific switching probability (i.e. $P_{\text{up-to-up}}$, $P_{\text{up-to-down}}$, $P_{\text{down-to-down}}$, and $P_{\text{down-to-up}}$, respectively). Obviously, the two sums ($P_{\text{up-to-up}}$ + $P_{\text{up-to-down}}$) and ($P_{\text{down-to-down}}$ + $P_{\text{down-to-up}}$) are equal to 100%.

In this work, a three-terminal SOT-MRAM device with a circular MTJ featuring a diameter of 25 nm and a CoFeB FL with a thermal stability $\Delta = 76.18$ at 300 K has been considered. Table 4.1 reports the main device parameters. In particular, a tungsten strip featuring a large spin-Hall angle θ_H =-0.33 has been considered [7].

Table 4.1: Device parameters of the SOT-MRAM device.

Parameter	**Description**	**Value**	**Units**
d	MTJ diameter	25	nm
t_{FL}	Free layer thickness	1	nm
α	Magnetic damping	0.03	--
η	Spin polarization factor	0.66	--
Θ_H	Spin Hall angle	-0.33	--
M_S	Saturation magnetization	800×10^3	A/m
K_u	Uniaxial anisotropy constant	1.05×10^6	J/m^3
RA	Resistance-area product	10	$\Omega\cdot\mu m^2$
TMR	Tunnel magnetoresistance	150	%
t_{HM}	HM thickness	6	nm
L_{HM}	HM length	50	nm
W_{HM}	HM width	50	nm
ρ_{HM}	HM resistivity	200×10^{-6}	$\Omega\cdot cm$

Statistical properties of the considered device have been evaluated by a state-of-the-art multi-domain micromagnetic solver, which numerically integrates the Landau–Lifshitz–Gilbert (LLG) equation describing the FL magnetization dynamics [7]. According to the results reported in [7], micromagnetic outcomes show switching probabilities quite close to 50 %, which represents the ideal condition for a PUF implementation.

4.3 SOT-MRAM based PUF design with CiM operation

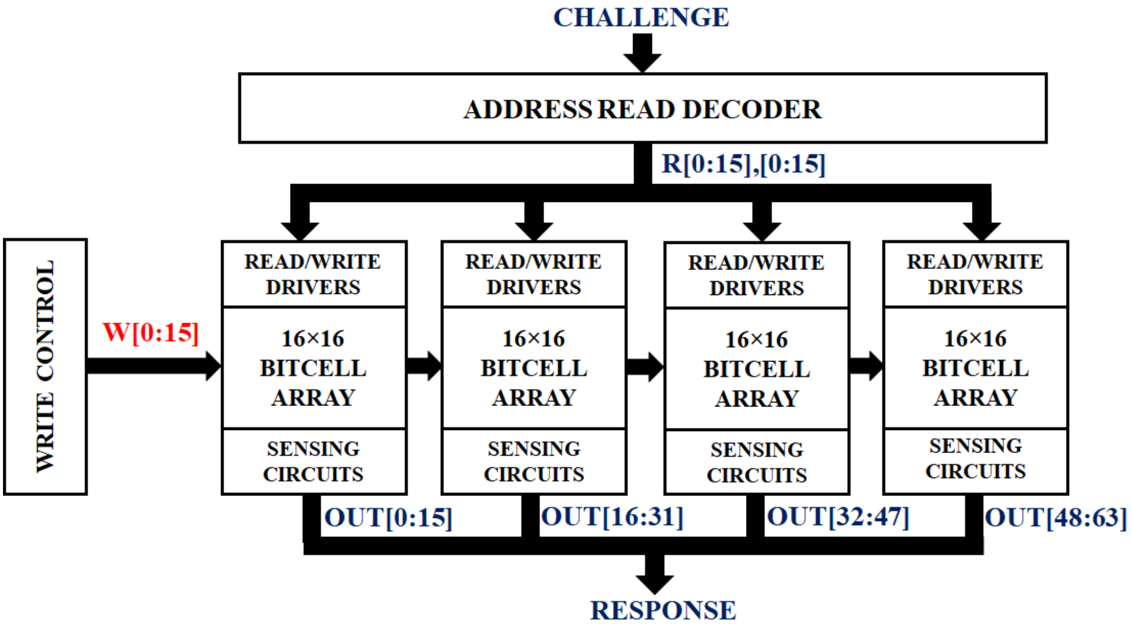


Figure 4.4: Block diagram of the PUF architecture.

By exploiting the above-described operation of the three-terminal SOT-MRAM device, a PUF circuit has been designed into Cadence Virtuoso tool using a hybrid CMOS/spintronics modeling approach [9-12]. To this aim, transistor models provided by a 0.8V/1.8V 18-nm FinFET technology [13] have been used. In addition, a macrospin-based compact model written in Verilog-A language [14-15] has been employed to integrate the behavior of SOT-MRAM devices into the circuit simulator.

Fig. 4.4 shows the general architecture of the designed PUF circuit, featuring four bitcell (BC) blocks, each including a 16×16 BC array along with read/write drivers and sensing circuits (including circuitry to implement a voltage sensing scheme, i.e. latch-type voltage SAs and reference generation circuits as described in Section 3.2) The circuit also includes a block for the generation of 16 write (W) signals, i.e. one for each row of the BC array, and an address read decoder to generate 16×16 read (R) signals

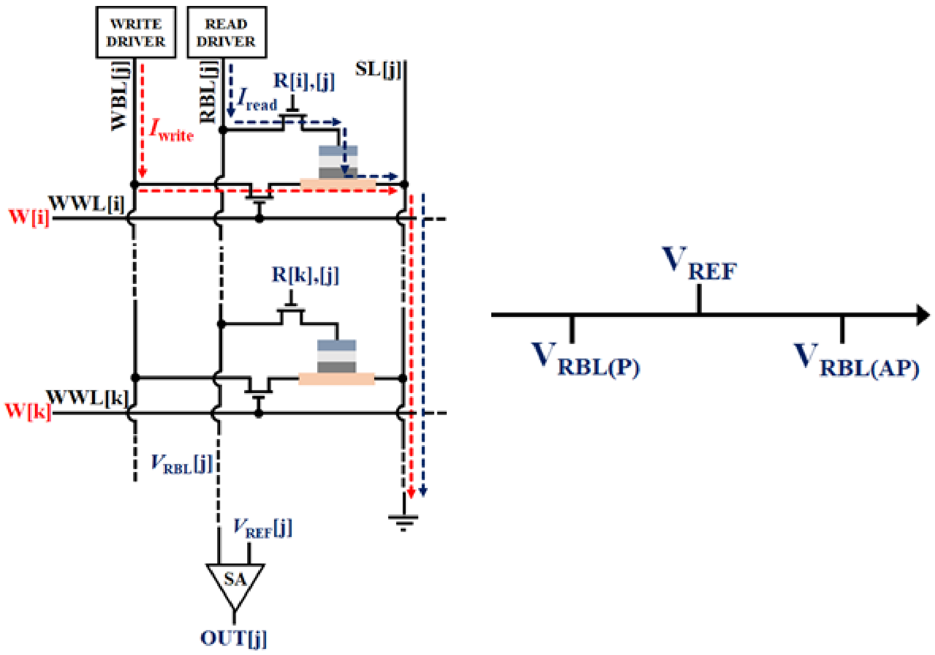


Figure 4.5: Scheme of the bitcell array in the circuit with conventional voltage sensing scheme, along with the detail of the required reference voltages.

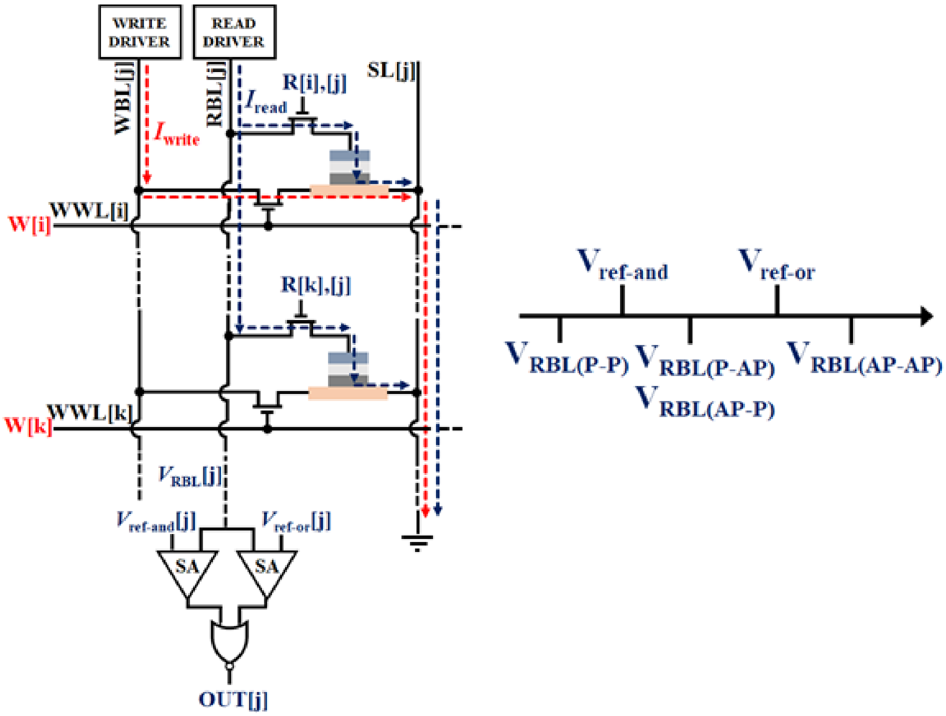


Figure 4.6: Scheme of the bitcell array in the circuit with enhanced voltage sensing scheme to implement a 2-bit XOR operation between two BCs within the same column.

on the basis of the specific input challenge. The R signals are used as inputs for the four BC block, each one providing 16 output (OUT) bits, i.e. one for each column, to obtain a 64-bit output word as response to a given challenge.

Figs. 4.5 and 4.6 then show the structure of the BC block related to two different implementations, respectively. In both circuits, each BC consists of a three-terminal SOT-MRAM device along with two access transistors. The write access transistor is driven by a W signal through the write word-line (WWL), which is shared by the bitcells belonging to the same row. On the other hand, the read access transistor is driven by the a signal. Bitcells within the same column share the write bit-line (WBL) that connects the write access transistors to the write driver, the read bit-line (RBL) that connects the read access transistors to the read driver and sensing circuitry, and the source-line (SL) that is grounded. Figs. 4.5 and 4.6 also show how writing and reading operations are implemented in the two designed circuits. We can note that the writing operation is realized in the same way in both circuits. In particular, the writing of the SOT-MRAM devices is implemented per row by applying a voltage to the WBL, while enabling write access transistors (i.e. the corresponding W signal = “1”) and disabling all the read access transistors (i.e. all R signals = “0”). Accordingly, a write current (I_{write}) flows through the HM strip of SOT-MRAM devices, thus enabling the SOT-based switching process described in Section 4.2. Conversely, the reading operation is implemented differently in the two designed circuits. The circuit of Fig. 4.5 employs a conventional voltage sensing scheme to detect the bit stored in the SOT-MRAM device of one BC per each column. This is done by enabling one read access transistor per column (i.e., only one R signal equal to “1” per each column on the basis of the specific challenge), while all write access transistors are disabled (i.e. all W signals = “0”). Then, by applying a read current (I_{read}) to the RBL, the developed RBL voltage (V_{RBL}) is compared with a reference voltage (V_{REF}) by a voltage SA, thus producing an OUT bit depending on the magnetization state of the MTJ (i.e. P or AP) in the activated bitcell. Instead, the circuit of Fig. 4.6 uses the enhanced voltage sensing scheme described in Chapter 3 to implement a 2-bit XOR operation between two BCs belonging to the same column. This is done by enabling two read access transistors per column at the same time during the sensing operation. As described before, an additional sensing circuitry along with two different reference voltages ($V_{ref\text{-}and}$ and $V_{ref\text{-}or}$) are needed to support the bitwise XOR logic operation. In particular, this requires two voltage SAs, which realize 2-bit AND and NOR operations by comparing the developed V_{RBL} (depending on the magnetization state of the MTJs in the two activated BCs) with $V_{ref\text{-}and}$ and $V_{ref\text{-}or}$, respectively. Then, the AND/NOR outputs of the two SAs are fed to a NOR gate, thus obtaining an OUT bit corresponding to the XOR between the two bits stored in the SOT-MRAM devices of the two activated bitcells.

4.3.1 Electrical simulation results

Circuit-level electrical simulations have been performed into Cadence Virtuoso simulator to evaluate the energy characteristics of the two designed circuits for both writing and reading operation, while considering the effect of the whole 16×16 BC array and the peripheral circuitry. In this regard, it is worth pointing out that, in the PUF implementation, the writing (or program) operation is typically performed only one time (or whenever a reprogramming of the stored secret is needed). Therefore, it does not significantly affect the power dissipation of the circuit during normal PUF operation. Conversely, the reading operation occurs as often as an output has to be generated as a response for a specific challenge [7].

Table 4.2 summarizes energy results extracted from performed circuit-level simulations. As expected, the two PUF circuits exhibit similar write energy per bit (E_{write}), i.e. about 520 fJ. Note that, in both the designed circuits, 1.8V I/O FinFET devices have been used for designing the circuitry involved in the writing operation (i.e. the write control block, the write drivers, and write access transistors). Conversely, standard 0.8V transistors have been used for designing the rest of the circuit. This has allowed achieving an adequate I_{write} (i.e. corresponding to $J_{SHE} = 3\times10^8$ A/cm^2) to guarantee the FL magnetization reaching the in-plane direction within few hundreds of ps (according to Fig. 4.3c). Obtained results also show that the PUF circuit based on the enhanced sensing scheme to support CiM bitwise XOR operations exhibits an about doubled read energy per bit (E_{read}) as compared to that with the conventional sensing scheme, i.e. 102 fJ vs. 49 fJ. This is due to the higher I_{read} required to ensure enough large sensing margins such that to easily distinguish the different values of V_{RBL} (depending on the states of the two MTJs as reported in Table 3.1) when implementing a 2-bit XOR operation, along with the contribution of additional sensing circuitry.

Table 4.2: Summary results of electrical simulations.

PUF circuit	**E_{write} [fJ]**	**E_{read} [fJ]**
Conventional sensing scheme (w/o 2-bit XOR)	517.8	49
Enhanced sensing scheme (w/ 2-bit XOR)	519.6	102

4.3.2 Statistical simulation results

Statistical simulations have been also performed to evaluate the response quality of the two PUF designs in terms of typical evaluation metrics, such as randomness and uniqueness, computed using the HW and the HD_{INTER} [18], respectively, both averaged over 1,000 challenges and 100 PUF instances.

As a first step of the statistical analysis, we considered the case where all the SOT-MRAM devices feature the same switching probabilities (i.e. $P_{up\text{-}to\text{-}down}$ and $P_{down\text{-}to\text{-}up}$ are the same for all the magnetic devices), ranging from 39% up to 61%. In this regard, Figs. 4.7 and Fig. 4.8 show the color map of the average HW obtained for the PUF circuits with conventional (i.e. w/o 2-bit XOR operations) and enhanced (i.e. w/ 2-bit XOR operations) sensing scheme, respectively. From Fig. 4.7, we can observe that the circuit with conventional sensing scheme allows obtaining randomness close to the ideality (i.e. 50%) only when $P_{\text{up-to-down}} \approx P_{\text{down-to-up}}$. Indeed, when $P_{\text{up-to-down}}$ and $P_{\text{down-to-up}}$ are quite different, the randomness moves away from 50% (Fig. 4.7). Conversely, in

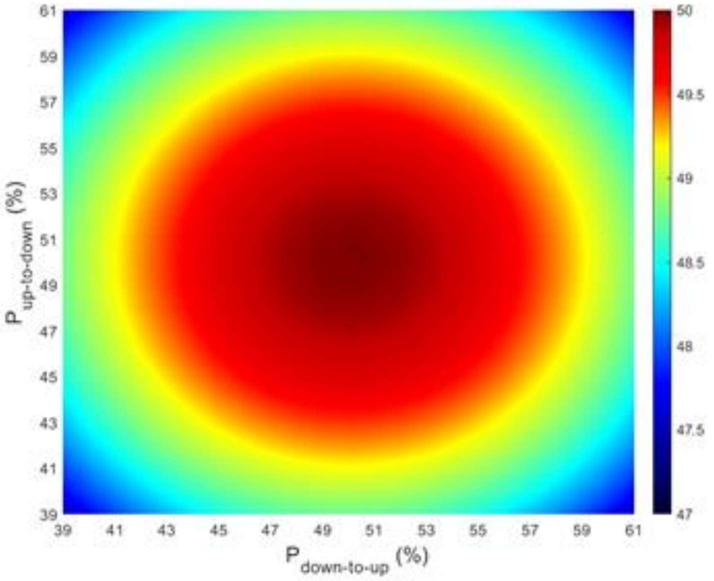


Figure 4.9: Color map of the average HD_{INTER} (over 10,000 challenges and 1,000 PUF instances) for the circuit with conventional sensing scheme (i.e. w/o 2-bit XOR operations) in the case where all the SOT-MRAM devices feature the same switching probabilities ($P_{\text{up-to-down}}$ and $P_{\text{down-to-up}}$ ranging from 39% up to 61%).

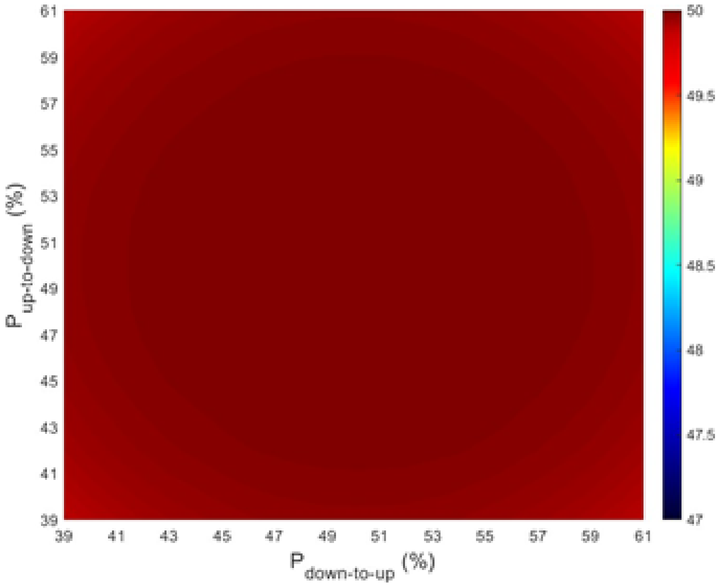


Figure 4.10: Color map of the average HD_{INTER} (over 10,000 challenges and 1,000 PUF instances) for the circuit with enhanced sensing scheme (i.e. w/ 2-bit XOR operations) in the case where all the SOT-MRAM devices feature the same switching probabilities ($P_{\text{up-to-down}}$ and $P_{\text{down-to-up}}$ ranging from 39% up to 61%).

the circuit with enhanced sensing scheme, the randomness is still close to 50% even in the case where $P_{\text{up-to-down}}$ and $P_{\text{down-to-up}}$ are quite different (Fig. 4.8).

Figs. 4.9 and Fig. 4.10 thus show the color map of the average HD_{INTER} obtained for the PUF circuits with conventional and enhanced sensing scheme, respectively. From Fig. 4.9, we can observe that the circuit with conventional sensing scheme allows obtaining uniqueness close to the ideality (i.e. 50%) only when both $P_{\text{up-to-down}}$ and $P_{\text{down-to-up}}$ are close to 50%. Conversely, in the circuit with enhanced sensing scheme, the uniqueness is still close to 50% even in the cases where both $P_{\text{up-to-down}}$ and $P_{\text{down-to-up}}$ move away from 50%. Therefore, statistical results reported in Figs. 4.7-4.10 prove that the implementation of 2-bit XOR operations is significantly beneficial to improve the quality of the PUF response.

In the second step of the performed statistical analysis, we considered the case where, due to the presence of defects in the planar geometry of the MTJ coming from the manufacturing process, the SOT-MRAM devices can exhibit different switching probabilities. This has been confirmed by means of micromagnetic multi-domain simulations where the effect of defects in the MTJ geometry has been accounted for by adding and/or removing some magnetic elementary cells with respect to the nominal geometry [21]. In particular, four different cases have been considered (Fig. 4.11): (a) MTJ with nominal geometry, (b) MTJ with defects and same area (i.e. the number of

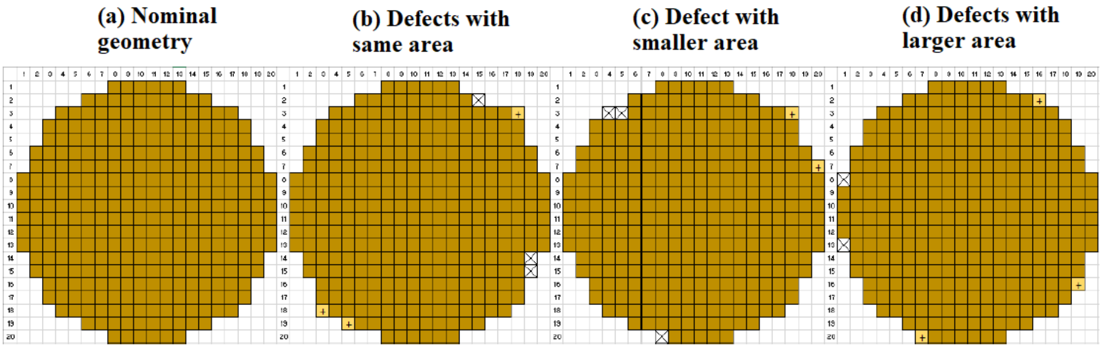

(a) Nominal geometry
(b) Defects with same area
(c) Defect with smaller area
(d) Defects with larger area
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Figure 4.11: Top-view of the FL geometry in the cases of (a) nominal geometry and in presence of defects: (b) same area, (c) smaller area, and (d) larger area.

added cells equals the number of removed ones), (c) MTJ with defects and smaller area (i.e. the number of added cells is lower than the number of removed ones), and (d) MTJ with defects and larger area (i.e. the number of added cells is higher than the number of removed ones). Table 4.3 reports the switching probabilities of the four considered cases obtained by micromagnetic simulations where an external in-plane field of 40 mT perpendicular to the J_{SHE} current has been also applied [7]. Data reported in Table 4.3 demonstrate that the presence of defects in the MTJ planar geometry can lead to a light detrimental effect on switching probabilities, leading them further away from the ideal value (i.e. 50%).

Table 4.3: Switching probabilities in the case of defects in the planar geometry of SOT-MRAM devices.

Case	**$P_{up\text{-}to\text{-}down}$ [%]**	**$P_{up\text{-}to\text{-}up}$ [%]**	**$P_{down\text{-}to\text{-}up}$ [%]**	**$P_{down\text{-}to\text{-}down}$ [%]**
Nominal geometry	52	48	52	48
Defects with same area	48	52	49	51
Defects with smaller area	61	39	45	55
Defects with larger area	46	54	48	52

Statistical simulations have been then repeated considering SOT-MRAM devices with different switching probabilities due to the presence of defects, according to the data of Table 4.3. In this regard, Table 4.4 shows statistical results obtained from such analysis for the PUF circuits with conventional (i.e. w/o 2-bit XOR operations) and enhanced (i.e. w/ 2-bit XOR operations) sensing scheme. Here, randomness and uniqueness results are reported in terms of the deviation from the ideal value (i.e. 0.5 corresponding to 50% when expressed as percentage).

Table 4.4: Summary results of statistical simulations.

PUF circuit	**δ_{rand}***	**δ_{uniq}***
Conventional sensing scheme (w/o 2-bit XOR)	2×10^{-2}	8×10^{-4}
Enhanced sensing scheme (w/ 2-bit XOR)	1×10^{-4}	2×10^{-5}

* δ_{rand} = |0.5 – randomness| and δ_{uniq} = |0.5 – uniqueness| are the deviation from the ideal value (i.e. 0.5)

Again, we can observe from Table 4.4 that the implementation of the enhanced sensing scheme enabling CiM 2-bit XOR operations allows reaching randomness and uniqueness much closer to the ideal value with respect to the circuit based on the conventional sensing scheme.

V. SUMMARY AND CONCLUSIONS

This thesis has presented the design of a Physical Unclonable Function (PUF) circuit based on Spin-Orbit-Torque Magnetic Random Access Memory (SOT-MRAM) devices and including a Computing-in-Memory (CiM) strategy to enhance the quality and security of the PUF device. Spintronic-based SOT-MRAMs along with Spin-Transfer-Torque (STT)-MRAMs are widely considered as promising candidates for the next generation of Systems on Chip with on-chip non-volatile storage implemented at nanoscaled technological nodes. CiM is regarded as a promising approach to deal with the processor-memory data transfer bottleneck in modern computing systems. In particular, CiM is particularly suitable for emerging resistive non-volatile memories like STT- and SOT-MRAMs where the simultaneous enabling of multiple bitcells in the memory array allows directly computing logic functions of the stored bits.

In the designed PUF implementation, the PUF secret is stored into a random state of a matrix of three-terminal SOT-MRAM devices consisting of a Magnetic Tunnel Junction (MTJ) with a perpendicular free layer (FL) placed on the top of a heavy metal (HM) strip. Such random state is obtained by applying a large enough current into the HM, which leads the FL magnetization in-plane. Once the current is removed, due to the effect of stochastic thermal fluctuations, the FL magnetization evolves to either up or down perpendicular direction with quite similar probability close to 50%. This gives rise to write randomly a bit “0” or “1” into the MTJ. Then, by following CiM approach, an enhanced sensing circuitry is exploited in the SOT-MRAM based PUF design to perform bitwise XOR logic operations during the reading operation.

The PUF architecture has been designed into a commercial circuit design tool (Cadence Virtuoso) by using a hybrid CMOS/spintronics modeling approach. Transistor models provided by a 0.8V/1.8V 18-nm FinFET technology have been considered, along with a Verilog-A based compact model aimed at integrating the behavior of SOT-MRAM devices into the circuit simulator.

Electrical and statistical simulations have been performed to evaluate energy characteristics for both writing and reading operations and security metrics of the designed PUF circuit. Obtained results prove that, although at the cost of higher read energy consumption, the implementation of 2-bit XOR operations allows reaching randomness and uniqueness values very close to the ideality (i.e. 50%), thus significantly increasing the quality and the security of the PUF response.

References

[1] S. A. Wolf, J. Lu, M. R. Stan, E. Chen, D. M. Treger, “The promise of nanomagnetics and spintronics for future logic and universal memory,” *Proc. of the IEEE*, vol. 98, no. 12, pp. 2155–2168, 2010.

[2] Liu, Enlong, “Materials and designs of magnetic tunnel junctions with perpendicular magnetic anisotropy for high-density memory applications,” in PhD Thesis, Nov 2018.

[3] N. Talati, R. Ben-Hur, N. Wald, A .Haj-Ali, J. Reuben, “MMPU—a real processing-in-memory architecture to combat the Von Neumann bottleneck,” in Springer Series in Advanced Microelectronics, 2019, pp. 191–213.

[4] G. Santoro, G. Turvani, and M. Graziano, “New logic-in-memory paradigms: An architectural and technological perspective,” in Micromachines, vol. 10, May 2019, pp. 368.

[5] C. Herder, M. Yu, F. Koushanfar and S. Devadas, “Physical unclonable functions and applications: A tutorial,” in Proceedings of the IEEE, vol. 102, 2014, pp. 1126–1141.

[6] Holcomb, D. E., Fu, K., “Bitline PUF: Building native challenge-response PUF capability into any SRAM,” in Cryptographic Hardware and Embedded Systems – CHES, 2014, pp. 510–526.

[7] G. Finocchio, T. Moriyama, R. De Rose, G. Siracusano, M. Lanuzza, V. Puliafito, M. Carpentieri, “Spin–orbit torque based physical unclonable function,” in Journal of Applied Physics , 128(3), 2020.

[8] S. Jain, A. Ranjan, K. Roy and A. Raghunathan, “Computing in memory with spin-transfer torque magnetic ram,” in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 26, no. 3, March 2018, pp. 470–483.

[9] De Rose R, Lanuzza M, Crupi F, Siracusano G, Tomasello R, Finocchio G, Carpentieri M (2017), “Variability-Aware Analysis of Hybrid MTJ/CMOS Circuits by a Micromagnetic-Based Simulation Framework,” IEEE Trans. Nanotechnol., vol. 16, no. 2, pp. 160-168.

[10] De Rose R, Lanuzza M, d’Aquino M, Carangelo G, Finocchio G, Crupi F, Carpentieri M (2017), “A Compact Model with Spin-Polarization Asymmetry for Nanoscaled Perpendicular MTJs,” IEEE Trans. Elec. Dev., vol. 64, no. 10, pp. 4346-4353.

[11] De Rose R, Lanuzza M, Crupi F, Siracusano G, Tomasello R, Finocchio G, Carpentieri M, Alioto M (2018), “A Variation-Aware Timing Modeling Approach for Write Operation in Hybrid CMOS/STT-MTJ Circuits,” IEEE Trans. Circuits Syst. I Regul. Pap., vol. 65, no. 3, pp. 1086-1095.

[12] De Rose R, d'Aquino M, Finocchio G, Crupi F, Carpentieri M, Lanuzza M (2019), "Compact Modeling of Perpendicular STT-MTJs With Double Reference Layers," IEEE Trans. Nanotechnol., vol. 18, pp. 1063-1070.

[13] Cadence Inc. (2016), "Generic 0.8V/1.8V Finfet/Multi Patterned 8 Metal Process Design Kit (PDK)," Available: https://support.cadence.com/.

[14] Wang Z, Zhao W, Deng E, Klein J-O, Chappert C (2015), "Perpendicular-anisotropy magnetic tunnel junction switched by spin-Hall-assisted spin-transfer torque," J. Phys. D: Appl. Phys., vol. 48, no. 6, 065001.

[15] Wang Z, Zhang L, Wang M, Wang Z, Zhu D, Zhang Y, Zhao W (2018), "High-Density NAND-Like Spin Transfer Torque Memory With Spin Orbit Torque Erase Operation," IEEE Elec. Dev. Lett., vol. 39, no. 3, pp. 343-346.

[16] S. Z. Peng et al., "Magnetic tunnel junctions for spintronics: Principles and applications," in Wiley Encyclopedia of Electrical and Electronics Engineering, Dec. 2014, pp. 1–16.

[17] Wicht, B., Nirschl, T., Schmitt-Landsiedel, D., "Yield and speed optimization of a latch-type voltage sense amplifier," in IEEE Journal of Solid-State Circuits, 39(7), 2004, p. 1148–1158.

[18] B. Halak, "Physically unclonable functions," Springer, 2018, pp. 24–25.

[19] L. Daihyun, J.W. Lee, B. Gassend, G.E. Suh, M.V. Dijk, S. Devadas, "Extracting secret keys from integrated circuits," IEEE Trans. Very Large Scale Integr. VLSI Syst. 13, 2005, pp- 1200-1205.

[20] Finocchio G, Carpentieri M, Martinez E, Azzerboni B (2013), "Switching of a single ferromagnetic layer driven by spin Hall effect," *Appl. Phys. Lett.*, vol. 102, 212410.

[21] Finocchio G, Consolo G, Carpentieri M, Romeo A, Azzerboni B, Torres L (2006), "Trends in spin-transfer-driven magnetization dynamics of CoFe⁄AlO⁄Py and CoFe⁄MgO⁄Py magnetic tunnel junctions," *Appl. Phys. Lett.*, vol. 89, 262509.