UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Perpendicular STT-MTJs with Double Reference Layers and its Application to Downscaled Memory Cells

Esteban José Garzón Córdova

Marco Lanuzza, Ph.D.
Director de Trabajo de Titulación

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

Quito, 21 de mayo de 2019

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Perpendicular STT-MTJs with Double Reference Layers and its Application to Downscaled Memory Cells

Esteban José Garzón Córdova

Firmas

Marco Lanuzza, Ph.D.

Director del Trabajo de Titulación

Lionel Trojman, Ph.D.

Director del Programa de Maestría en Nanoelectrónica

César Zambrano, Ph.D.

Decano del Colegio de Ciencias e Ingenierías

Hugo Burgos, Ph.D.

Decano del Colegio de Posgrados

Quito, 21 de mayo de 2019

© 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 el Art. 144 de la Ley Orgánica de Educación Superior.

Firma del estudiante:	
Nombre:	Esteban José Garzón Córdova
Código de estudiante:	00139265
C. I.:	1715952816
Lugar, Fecha	Quito, 21 de mayo de 2019

DEDICATORIA

A cada miembro de la familia Garzón-Córdova, a mi padre José a mi madre Ana Paulina y hermano Nicolás que ha sido una ayuda incondicional y pilar fundamental para alcanzar mis metas.

AGRADECIMIENTOS

Al departamento de ingeniería electrónica de la Universidad San Francisco de Quito y en especial a los responsables del programa de maestría por la beca de excelencia otorgada, que me permitió explorar nuevas fronteras y alcanzar logros. A los docentes de la Universidad San Francisco de Quito por abastecerme de sus conocimientos a lo largo de mis estudios que a su vez me ayudaron a triunfar en mis metas profesionales y académicas.

Especial agradecimiento a mi familia que gracias a su apoyo ningún logro hubiera sido posible.

RESUMEN

El diseño del chip presenta problemas debido al escalamiento de dispositivos a medida que el nodo tecnológico llega a sus límites físicos. La ruta para el desarrollo de nodos de 7nm en adelante se ha trazado, y superar los problemas de potencia y disipación de energía se ha convertido una parte fundamental para el diseño de chips. Las memorias en el diseño de chips tienen un papel fundamental y a su vez conforman un componente crucial que define el rendimiento del sistema. Para abarcar dichos problemas, se ha realizado investigación en el campo de las memorias MRAM, lo que ha conllevado a resultados significativos para memorias no volátiles con menos potencia en operaciones. Estos dispositivos se denominan Magnetic Tunnel Junction (MTJ) donde se proponen diferentes diseños para abarcar con las demandas de los nuevos nodos tecnológicos.

Esta tesis presenta el análisis de un arreglo de memoria STT-MRAM de 128 × 128 mediante el uso de dos tipos de dispositivos, el Single barrier (SB) MTJ y el Double barrier (DB) MTJ. Se explica los fundamentos del flujo de corriente MTJ y fenómenos importantes, como la resistencia a la magneto del túnel (TMR) y la anisotropía magnética perpendicular (PMA). El objetivo es estudiar el comportamiento de un STT-MRAM donde se muestran las ventajas, desventajas y el equilibrio entre los rendimientos de SB y DB. Para cada tipo de dispositivo, se toma un conjunto de cuatro configuraciones y, a través de un análisis determinístico y estadístico, se elegirá la configuración óptima en términos de energía. Además, la estructura de la celda (bitcell) es una combinación de las tecnologías FinFET y MTJ y, en consecuencia, se utiliza un modelo híbrido. En el diseño híbrido, el enfoque general para el diseño de circuitos cambia. En esta tesis, combinamos dos modelos diferentes, el modelo MOS proporcionado por modelos comerciales y el modelo MTJ representado en un código Verilog-A debido a la ausencia de modelos comerciales.

Palabras clave: Magnetic Tunnel Junction (MTJ), Spin-transfer torque magnetic RAM (STT-MRAM), Double Barrier (DB), Single Barrier (SB), Perpendicular Magnetic Anisotropy (PMA), FinFET, modelo compacto, espintrónica, memorias no volátiles (NVM).

ABSTRACT

Chip design presents problems due to scaling as the technology node reaches to the physical limits. The roadmap to 7nm technology node and beyond is already traced and overcome the problems in power and energy dissipation have become a fundamental part in the chip design. Memories on chip design take place a fundamental role and it became a crucial component, which defines the performance of the system. To encompass the problems a lot of research in MRAM has been done, leading to significant positive results focused on nonvolatile memories with less power in operations. These devices are called Magnetic Tunnel Junction (MTJ) where different designs are proposed in order to accomplish the new technology node demands.

This thesis presents the analysis of an array of 128×128 STT-MRAM by using two type of devices, the Single Barrier (SB) MTJ and the Double Barrier (DB) MTJ. The fundamentals of the MTJ current flow and important phenomena such as the Tunnel MagnetoResistance (TMR) and the Perpendicular Magnetic Anisotropy (PMA) will be explained. The objective is to study the behavior of an STT-MRAM where it is shown the advantages, disadvantages, and trade-off between SB and DB performances. For each type of device, a set of four type of configurations is taken, and through a deterministic and statistical analysis, the optimal configuration in terms of energy will be chosen. Besides, the bitcell structure is a combination of the FinFET and MTJ technology and in consequence, a hybrid model is used. In the hybrid design, the general approach for circuit design changes. In this thesis, we combine two different models, the MOS model provided by the foundry and the MTJ model represented in a Verilog-A code due to the absence of commercial models.

Keywords: Magnetic Tunnel Junction (MTJ), Spin-transfer torque magnetic RAM (STT-MRAM), Double Barrier (DB), Single Barrier (SB), Perpendicular Magnetic Anisotropy (PMA), FinFET, compact model, spintronics, memories.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

CHAPTER 1 - INTRODUCTION

1.1 Current chip design and new trends on memory technology

Nowadays, approximately 2.5 exabytes (EB) of digital information is produced daily, a quantity equivalent to billions of electronic devices. Roughly, a big percentage of the total data in the world was generated in the last years. Although, every year the amount of information that a single person manipulates is increasing due to electronic devices. For this reason, the electronic devices, information, and communication technology have become essential for our society giving the semiconductor industry one of the top places in the market.

Current integrated circuit (IC) design presents notorious challenges due to scaling. As the technology node reaches its physical limits the concern in power and energy consumption have become an important problematic on chip design. Scaling is not necessarily bad, it helped the circuits to be faster; nevertheless, it causes an increase on power consumption and in consequence a reduction of battery life, which is crucial in portable devices [19]. Although, the circuits become denser, leading to more power consumption and especially an increase in leakage when transistors are in standby mode. On the other hand, on IC design, the most common memories used nowadays are based on charge storage and a great part of the used area on the chip is due to the memory circuit, which is directly related to a great part of the power consumption on the chip. This makes the power-area a fundamental design metric. Furthermore, due to scaling, the charge storage memories are facing problems such as low-density improvements or lack of robustness because of variability and high power consumption in standby mode. As the density increases, the memory block becomes crucial and affect directly to a significant fraction of the total energy budget of the chip [18, 19]. In order

to face these problems in chip design, different solutions have been presented along the years. Solutions focusing on power and energy consumption such as parallelism, stack effect, logic optimization, etc. Moreover, focusing on memories, new topologies using traditional CMOS technology had been explored; nevertheless, the problem with leakage in standby mode is always present, placing the memories as an essential problematic on current chip design.

Memories technologies are in constant research where new proposals are emerging to face the problems mentioned before. The traditional memory hierarchies have speed gaps between their levels as the frequency increases [20]. To fill these gaps and ensure a good performance with less power consumption, a new restructuring of the memory hierarchy is proposed [18, 20]. The current and new memory hierarchies are shown in figure 1. Here we see potential possibilities using non-volatile (NV) logic and memories based on Spintronics technologies.

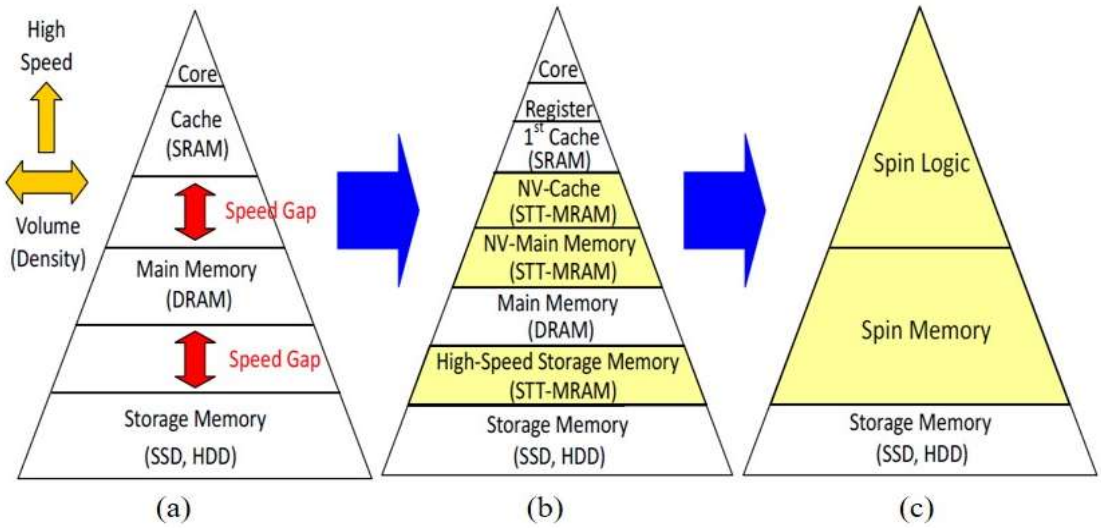

Figure 1: Current and new memory hierarchies [20]. (a) Representation of the conventional memory hierarchy. (b) - (c) Expected memory hierarchy.

As it is illustrated in figure 1, the core or processing units are placed on top. Therefore, as it is getting closer to the core, the memory becomes faster and it size decreases. On the contrary, when it is far from the core the volume increases and the

speed decreases. In this path to the core, the old trend leaves a speed gap on the memory hierarchy (see figure 1-a) due to the technology node scaling, which limits the scaling related to memories. In other words, memory devices cannot be scaled without losing performance. Furthermore, to reach the core, register files are used, followed with the embedded cache modules that are based on Static Random-Access Memory (SRAM). SRAM presents good performance in speed and easy integration due to the use of purely CMOS technology. Nevertheless, it exhibits limitations in area integration produced by the size of the bit cell, which typically is composed of six transistors, and consequently, the cost will be higher. Despite the use of the CMOS technology node, the physical effects and thermal dissipation of the transistors determinates the system performance.

Continuing with the hierarchical pyramidal shape, it is seen the introduction of a Dynamic Random-Access Memory (DRAM). Comparing DRAM and SRAM, DRAM uses less area because a transistor and a capacitor build the bit cell [31]. Some drawbacks on DRAM are in power consumption and fabrication process. In terms of power consumption, it is not efficient when large DRAMs are built, due to the two stages presented when a bit needs to be saved [30] [31]. On the contrary, the fabrication process used on DRAM is different from the one used by SRAM [32]. Thus, the DRAM is not suitable for embedded memories. There is still a notorious problem presented on both memories, SRAM and DRAM, which is the high standby power dissipation. This problem gets worse because of the volatility of the memory, needing a constant power supply to preserve the information. Finally, for all the reasons mentioned, SRAM and DRAM will not work for the future computer systems.

At the end of the hierarchy, it is seen the non-volatile memories, which includes the massive storage such as Hard Disk Drive (HDD) and Solid State Drive (SSD) or flash memories like NAND or NOR. These memories in comparison with the previous ones generate a gap of different orders of magnitude according to the speed point of view. Due to these gaps, the performance of a computer system degrades. To face and follow the requirements of a system and fill the remaining gaps, new technologies such as MRAM based memories are placed on the field. Different proposals have emerged, each one with their own limitations. Following conventional approaches, we have new memory technologies like Phase-Change Random Access Memory (PCRAM) and Resistive Random Access Memory (RRAM); and with the Spintronics approach, we have the MRAM, which includes the Spin-Transfer Torque MRAM (STT-MRAM) and the Spin-Orbit Transfer MRAM (SOT-MRAM). Among all these technologies, the most promising in terms of scalability and power consumption are STT-MRAM and SOT-MRAM, making them suitable for embedded memories (see figure 1-b-c). The reader may notice that the SOT-MRAM is not exhibiting on the hierarchy (b) however, it is expected its replacement in the first cache memory blocks where the STT-MRAM is not suitable to use. STT-MRAM is a potential replacement for L3 cache memory, while the SOT-MRAM has recently emerged as another potential prospective that could be introduced in the SRAM application domain by replacing the L1 and L2 cache memories [40]. To highlight the features of the different emerging memory technologies, table 1 exhibits a comparison between them. Note the SOT-MRAM data was demonstrated for the first time the full integration on 300mm wafer on June 2018 by IMEC, reason why there is a lack of information.

Table 1: Comparison between different emerging and established memories [32] [40].

	Emerging Memory				Established Memory	
	SOT-MRAM	STT-MRAM	PCMS	RRAM	DRAM	Flash NAND
Non-Volatile	Yes	Yes	Yes	Yes	No	Yes
Endurance (Nb cycles)	High (5x1010)	High (1012)	Medium (108)	Low (106)	High (1015)	Low (105)
2016 latest technological node produced (nm)	-	40 nm	20 nm	130 nm	IX nm	15 nm
Cell size (cell size in F2)	-	Medium (6-12)	Not Specified	Medium (6-12)	Small (6-10)	Very small (4)
Read latency (ns)	Very Fast (0.21 ns)	Fast (10-20 ns)	Fast (50-100 ns)	Medium (250 ns)	Very fast (ns)	Slow (100,000 ns)
Power consumption	300 (pJ)	Medium (50 pJ/bit)	Medium	Medium (6nJ/bit)	Low	Very High
2016 price ($/Gb)	-	High ($3000-$200/Gb)	Low (<$0.5/Gb)	High ($100/Gb)	Low (<$1/Gb)	Very Low (<$0.05/Gb)
Suppliers	-	Everspin	Micron/Intel	Adesto	Samsung, Micron, SK Hynix	Samsung, Micron, Toshiba, SK Hynix, Intel

The market for memories is given by the application of the memory. For instance, if the users need to save information for future use, they will need an HDD or a NAND memory. On the other hand, if users need a fast service where the data is used for a short time, a fast memory such as SRAM or a stable memory like DRAM is needed. However, the main goal is to fill the technology gaps and look for a potential replacement of SRAM or DRAM because of the scaling reasons mentioned before. The potential prospects are based on Magnetoresistive Random Access Memory (MRAM) where STT-MRAM and SOT-MRAM present a low power consumption and speed improvements as was shown in figure and table 1.

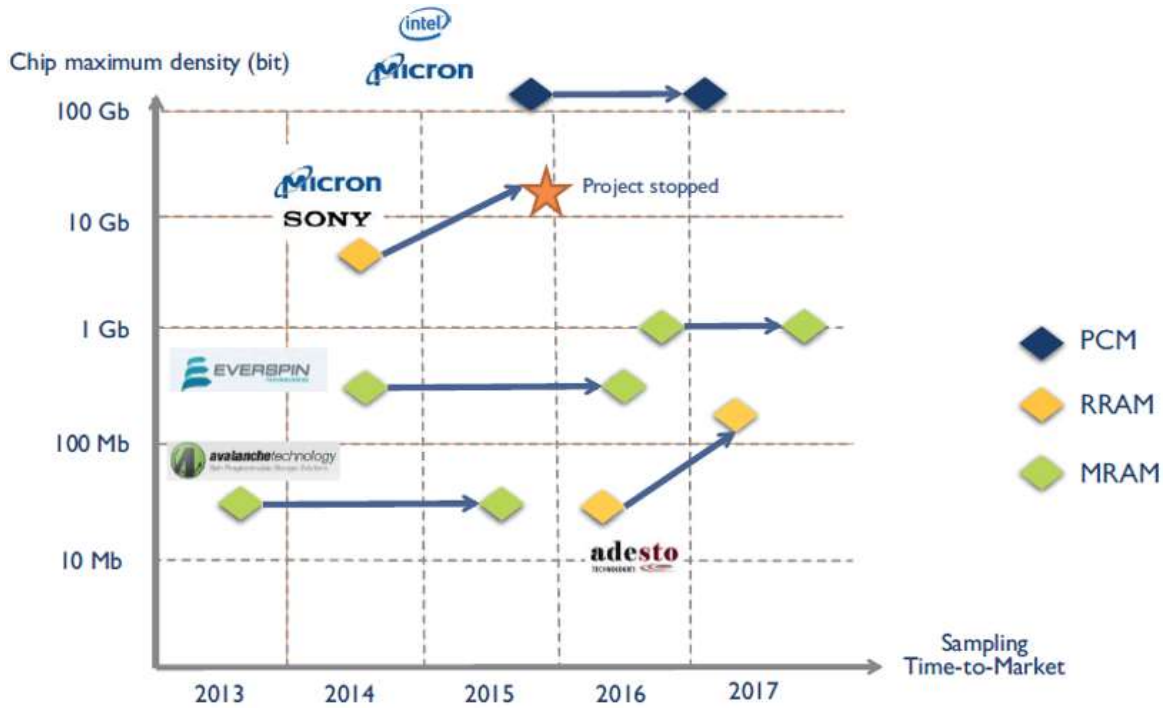

Figure 2: Chip density vs. sampling time-to-market. The list of companies is not exhaustive (Data updated until June 2017) [21].

As it was told before, the focus is on new NV memories. Different technologies are present, such as Phase-Change Memory (PCM), Resistive Random Access Memory (RRAM) and the Magnetoresistive Random Access Memory (MRAM) of which we will be talking along this thesis. Currently, NV technologies are present in specific niches of the market because of the limited density. The market of NV memories is increasing considerably, it is expected to reach

$3.9 billion by 2022 with improves in cost and density [21]. On the other hand, foundries such as GlobalFoundries, Samsung, TSMC, UMC, and SMIC are giving a big step to introduce MRAM and RRAM technologies by 2018-2019 [21, 22]. In figure 2, we can see the evolution of the technologies mentioned previously. It is taken into account the 'on-chip density vs. sampling time-to-market' where clearly the MRAM technology is taking the lead by a couple of companies. Thus, the memory business shows good conditions to adapt the NV technology on the market.

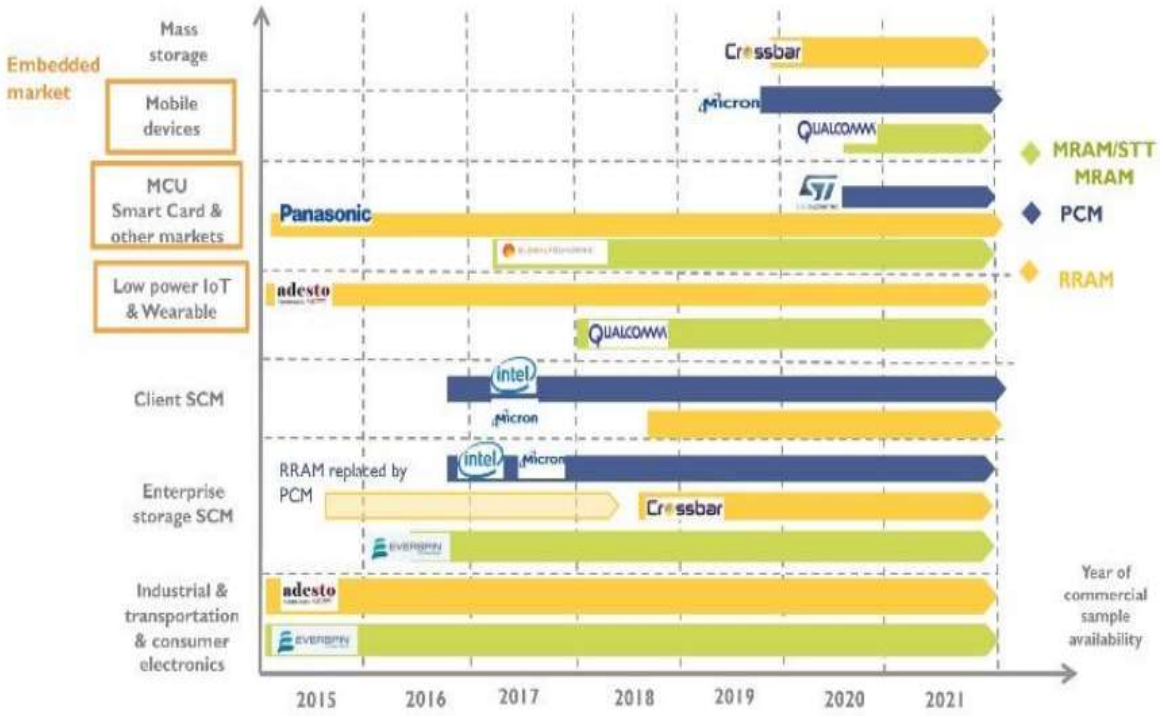

Figure 3: Market applications of NV memories – Data updated until July 2016 [23].

Several applications for the new technologies on NV memories have raised in the past years. Figure 3 shows that before 2015 the MRAM applications were limited for industrial, transportation and specific consumer electronics. Nevertheless, the impact on the market increased considerably, reaching to the embedded market like mobile devices or low power IoT and wearable. For this reason, on niche market applications the MRAM technology will be suitably introduced before 2021 with a potential replacing of DRAM by the MRAM [23]. The

increase in commerce and market of this technology will depend on companies and foundries. Nevertheless, variations of MRAM technology such as Spin-Transfer Torque MRAM (STT-MRAM) is expected to lead in the market by 2021 [23].

In the following sections is briefly explained the MRAM working principle and an overview of it, mentioning the most important discoveries and developments until now.

1.2 Spintronic based memories

Devices that involve the effect of the electron spin are called Spintronic devices, which use the electron properties to process digital information. As it was mentioned previously, the classical approach is on charge-based devices where the use of an electrical charge is needed to manipulate information. In Spintronics, the spin electron property, which is related to its angular momentum, is used. The focus is on the manipulation of the electron spin in different metals and semiconductor materials. Furthermore, the electron spin can be changed depending on the magnetization so a device can exhibit a big or low resistance state.

The aim is the development of memories based on Spintronics, principally MRAM applications. MRAM is a type of device in charge of store digital data in stable magnetic states, which are found on Magnetoresistive devices [1]. Besides, these devices are based on the phenomenon known as Giant Magnetoresistance (GMR), which is defined as the variation of the electrical resistance due to the change in an applied magnetic field [25]. According to the value of the device resistance, we can write, read or hold its data. In addition, it has two stable states, high resistance and low resistance that gives the value of magnetoresistance. These devices are built to have a magnetoresistance value very large in order to get a high performance with a good distinction of the cell state [1, 2]. On the other hand, we have two generations of MRAM, the first one is called Toggle MRAM where the cell is programmed by using magnetic fields [8]. However, due to scalability problems, a second generation emerged.

The second MRAM generation uses the Spin-Transfer Torque (STT) and it is based on Magnetic Tunnel Junction (MTJ) with CMOS technology [9]. The main difference between these two is that the first MRAM uses an external magnetic field while the second one has to induce a current on the device.

Another potential device has emerged, it is known as SOT-MRAM. As it was mentioned previously, it exhibits a good performance capable to replace the SRAM cache memories at the higher levels placing the SOT-MRAM as the next generation MRAM technology [40]. The STT and SOT MRAMs are based on MTJ with CMOS technology. The main difference relays on the read and write operation where on the STT happens in the same path, while SOT has the write and read operations decoupled.

Figure 4: Spin-Transfer Torque MRAM EMD3D256M – Everspin Technologies 2017

1.3 Roadmap of magnetoresistive random access memory (MRAM) technologies

MRAM technology has passed through several important moments along its development. In table 1, we can see the roadmap of MRAM that includes the milestones of the technology. It is seen that potential designs have been launched and all started with the Giant Magnetoresistance (GMR) discovery in 1989 by IBM, to later introduce tunneling materials, reaching a new perspective by using Magnetic Tunnel Junction devices. Different

variations and approaches were presented. One of them is known as Spin-Transfer Torque MRAM (STT-MRAM) and will be analyzed in the next chapter. Furthermore, figure 4 is shown a sample of one of Everspin's Spintronic memory device (STT-MRAM) launched on the market.

Table 2: Milestones in MRAM technology [24].

Year	Event
1989	Magnetoresistive effect discovered by IBM in thin-film structures. The Giant Magnetoresistance discovery (GMR)
2000	Beginning of the MRAM development program.
2003	Introduction of 128 Kbit MRAM chip build with 0.180 (µm) technology.
2004	Different companies like Infineon and Toshiba develop MRAM cells prototypes up to 16 Mbit. MRAM became as part of standard product for companies such as Freescale.
2005	The first development of MRAM using Spin-Transfer Torque (STT) by Renesas Technology and MRAM cells runs at 2 GHz.
2006	The fastest, high-density MRAM developed by Toshiba and NEC. In this year, the sale of the first MRAM chip (MR2A16A) begins.
2007	First MRAM device that drives the road to 1Gbit of capacity. Moreover, more research on STT technology is done. Freescale starts to sell 4 Mbit MRAM.

2008	Companies like Freescale has sold over a million MRAM chips. Samsung starts the development of STT-MRAM predicting a chip release in 2012. Freescale founds a new company called Everspin. Everspin technologies release the memories for consumer applications.
2009	Big companies such as Airbus starts to use Everspin's MRAM as part of the flight control. Thus, the interest in this memories increases and Everspin rises the MRAM for consumer applications.
2010	New Tunnel-Magnetoresistance element available for STT-MRAM permitting a 10 Gbit capacity. Strong research on improving STT-MRAM.
2011	More companies like BMW and Toshiba start to use MRAM for their own applications. Samsung designed and developed a perpendicular MTJ at 17nm technology. Everspin technologies make an agreement with Cadence in order to establish memory models.
2012	First STT-MRAM chip announced by Everspin launched in 2013. Toshiba developed an STT-MRAM low power consumption.
2013	Toshiba builds a computer architecture based on STT-MRAM.

	The launch of the new SATA SSD using STT-MRAM (Chip series: EMD3D064M Spin-Torque MRAM).
2014	Everspin signed an agreement with GlobalFoundries and sold close to 40 million MRAM chips. On the other hand, Everspin starts the STT-MRAM production while Toshiba developed a microprocessor cache memory based on STT-MRAM.
2015	Start of 32/64Mbit STT-MRAM samples by Everspin.
2016	Start of 256Mb STT-MRAM samples by Everspin. Demonstration of 11nm STT-MRAM junction by IBM. 4 Gb STT-MRAM prototype by Toshiba. Furthermore, the first time demonstrated a perpendicular MTJ at 8nm by IMEC researchers.
2017	Start of 1Gb perpendicular MTJ STT-MRAM samples by Everspin.
2018	January: Commercial production of the first 40nm 256Mb perpendicular MTJ STT-MRAM. February: Ultra-small 10nm MTJ developed in Tohoku University. August: IMEC exhibits the manufacturing of Spin-Orbit torque MRAM on 300mm silicon wafers.

CHAPTER 2 – SPIN-TORQUE MAGNETIC TUNNEL JUNCTION (STT-MTJ)

As it was discussed the previously, actual technology trends force the research of new technologies and solutions for the current chip design problems, specific approaches based on alternative memory devices. In this chapter is presented an explanation of the fundamentals of MRAM devices and a description of the STT-MRAM structures and configuration used in this thesis.

2.1 Magnetoresistance tunnel junction (MTJ) fundamentals

The basic device on the STT-MRAM design is called MTJ. It has two important phenomena; the first one is called the Tunnel MagnetoResistance (TMR), which is considered the readout signal of the device. The second one is the Perpendicular Magnetic Anisotropy (PMA). Before the explanation of those phenomena, it is important to understand the principle of the magnetoresistance effect. Furthermore, an explanation of the MTJ write and read operations, followed by the analytical model description, are presented in this section.

- **Magnetoresistance effect**

Phenomenon researched by Humphrey at the 80's where an electrical resistance variation of a material is presented according to an applied electric field [3]. The concept mostly relays on the properties of a magnet and its behavior due to external influences. Considering a simple magnet, it is known that the electrons presented in the magnet have a spin configuration (North and South Pole). For instance, in figure 5 we can see these configurations where the electrons spin states can be changed by flipping the poles of the magnet. Note that a reference point is needed in order to know the appropriate state.

Figure 5: Two configurations of a magnet: North-South (NS) and South-North (SN) where "black layers" represent the poles.

Taking two magnets and getting closer one to each other the phenomenon of attraction or repulsion can be seen (See figure 6). This phenomenon is related to high and low resistance. Thus, we can consider it in the use of electrical circuits. For instance, consider a simple electrical circuit showed in figure 7, when the magnets feel attraction with opposite poles a 'low resistance behavior' is presented; on the contrary, if we force the magnets to touch each other with same poles, the effect of repulsion is represented or a 'high resistance behavior' in response to that touching. Therefore, in order to take advantage of these effects, we consider ferromagnetic materials, which are used nowadays in the research and development of new memory technologies such as MRAM.

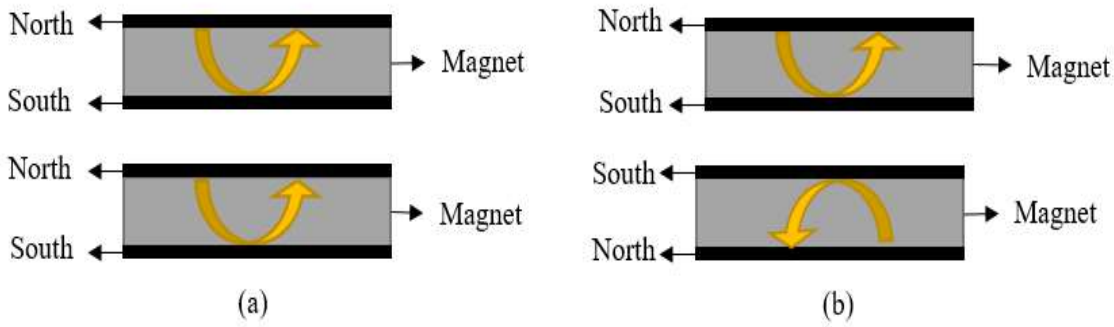

Figure 6: (a) Configuration NS-NS: A magnetic attraction is felt. (b) Configuration NS-SN: A magnetic repulsion is felt.

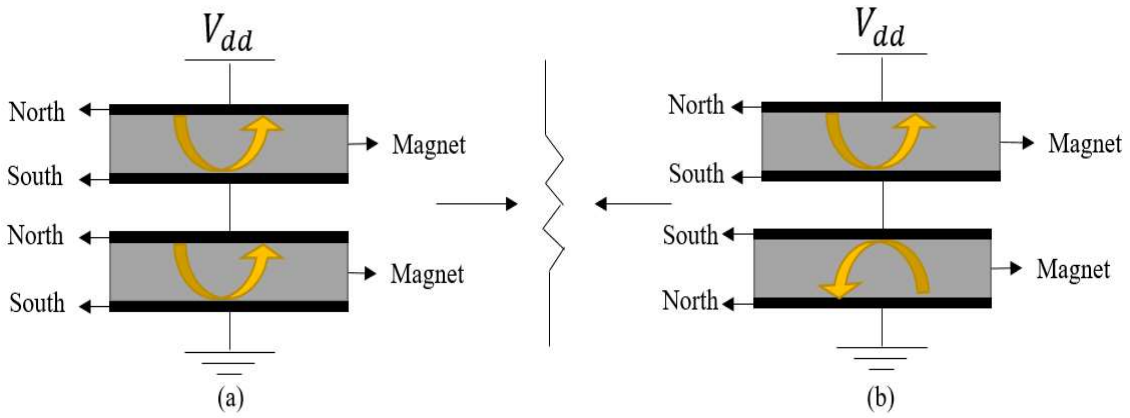

Figure 7: (a) Low resistance and (b) high resistance.

- **Overview of conventional magnetoresistance tunnel junction (MTJ) structure – Tunnel magnetoresistance (TMR) effect**

The magnetoresistance tunnel junction can be classified according to the type of insulator. A basic MTJ structure is presented in figure 8 where it is composed by two ferromagnetic layers and an insulator. The first one is the pinned layer (PL): this layer is also known as reference layer and it is fundamental to make a reference point when a change of the electron spin is done. Then we have the insulator layer whose general function is to define the type of magnetoresistance structure. For instance, in the case of the MTJ from figure 8, we use an insulator and it is called a TMR device. On the contrary, if we use a non-magnetic metal it is known as the Giant Magnetoresistance (GMR) device. Furthermore, both configurations are related to quantum mechanics effects. Finally, we have the Free Layer (FL) that gives the spin configuration such as Parallel (P) or Anti-Parallel (AP). In summary, the PL and FL act as polarizer and analyzer respectively while the electrons are passing through a thin oxide tunnel barrier, which is commonly used.

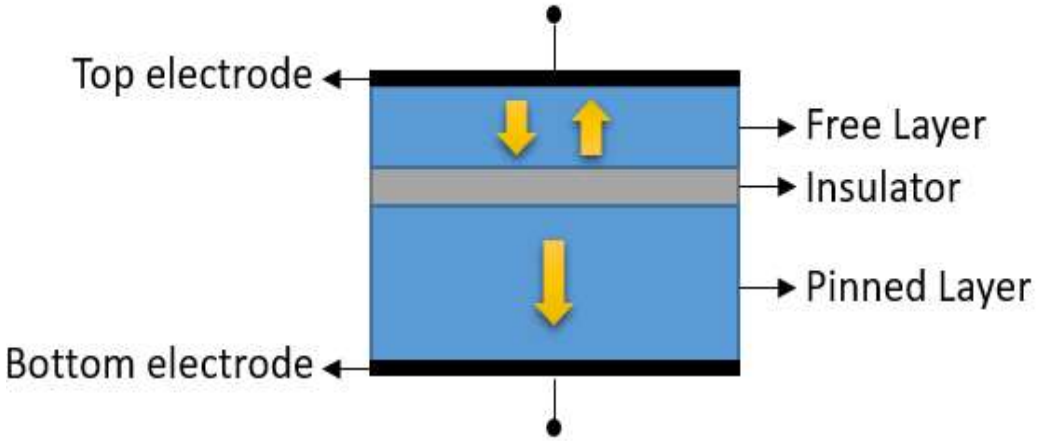

Figure 8: MTJ basic configuration.

The free layer has an anisotropic energy barrier, which allows the change from one spin configuration to another. Figure 9 presents the behavior of the free layer spin configuration according to the energy needed to flip from one state to another. Starting from zero degrees (Parallel state), the spin cannot change its state unless it has enough energy to jump to the next state (Anti-Parallel state) of 180 degrees. Thus, figure 9 shows clearly, how the spin can be retained. This can be understood as the retention of the spin by an energy barrier E_B. Furthermore, it is shown that $E_{B1} = E_{B2}$ which is not always true due to the presence of high order effects. When the MTJ is subject to high order effects the energy barrier is the minimum energy between E_{B1} and E_{B2}.

In order to characterize MTJ devices we take into account the Magnetoresistive Ratio:

$$MR = \frac{R_H - R_L}{R_L} \tag{1}$$

Where R_H and R_L represent the MTJ high and low resistance respectively, which also corresponds to the stable states of Parallel or Antiparallel. As the MR increases, the better the MTJ is, and a better differentiation between states is done. Furthermore, we have seen

previously that two different types of structures are present: the GMR and TMR. Between these two, the most suitable structure is the TMR that exhibits an MR greater than 100% [10]. Nevertheless, having a high MR is not enough due to different problems that this structure has to overcome in order to get a good performance on the circuit design.

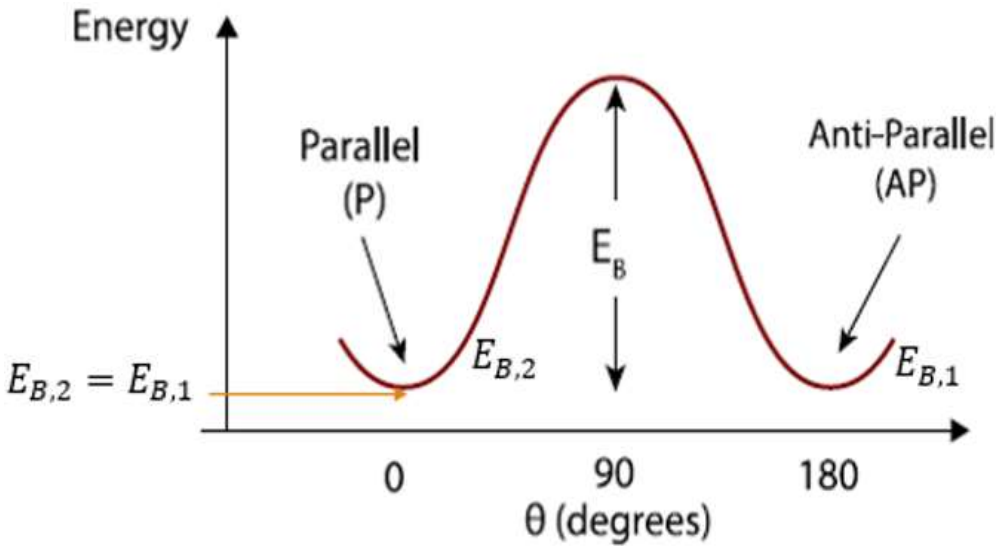

Figure 9: Energy and spin configuration [6]. The degrees represent the angle between the PL magnetization and FL magnetization.

- **Perpendicular magnetic anisotropy (PMA) MTJ**

The last aspect of the MTJ model is the fact that ferromagnetic materials have a different type of magnetization. It is found that the magnetization on the magnetic layer can be “in-plane magnetic anisotropy” (IMA) or “perpendicular magnetic anisotropy” (PMA) phenomenon that allows the non-volatile data retention. In figure 10 is shown the configuration mentioned. We can see that the IMA have a greater area than the PMA so, in order to achieve high integration density and a low switching critical current (enhancing the writing efficiency), the PMA is used [6].

PMA arises at the interface between two different layers. These layers are MgO and CoFeB, which help considerably the development of non-volatile memories based on perpendicular MTJ (p-MTJ) [34]. Hence, the core of an STT-MRAM is the P-MTJ based on the

interfacial effect due to CoFeB/MgO. The interface improves the anisotropy and data retention, allowing the scaling of the device and an increase of the energy barrier when the device is summited to thermal agitation [34] [36]. Improving the performance is also subjected to the addition of metallic capping layers in the MTJ, which have a positive enhancement in the features of the device [33]. This complicates the fabrication process and as a result, the price increases in orders of magnitude compared with typical consumer application memory devices as we have seen in chapter 1.

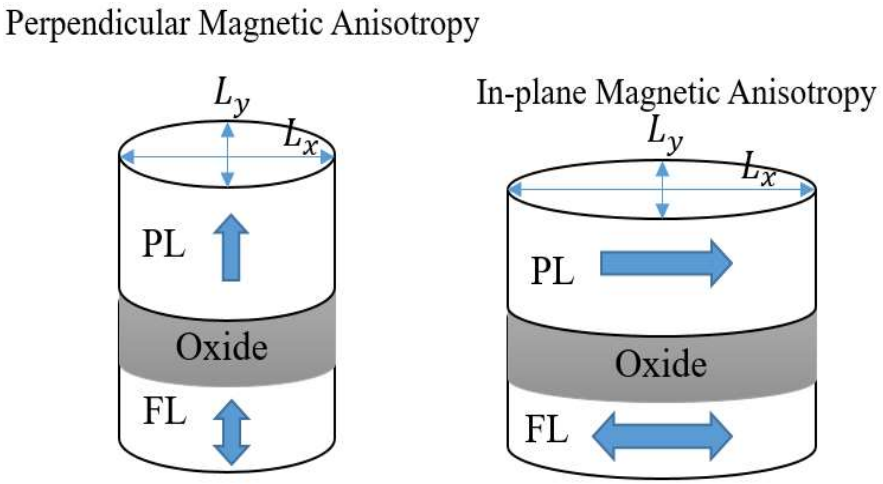

Figure 10: “In-plane Magnetic Anisotropy” (IMA) and “Perpendicular Magnetic Anisotropy” (PMA). It is shown the IMA presents $L_x > L_y$ while PMA has a circular cross-section that makes it suitable for integration.

P-MTJ can be done by one FL and one PL wherein the presence of high temperatures and a high current is flowing in the direction of the PL there will be a large field from the PL and can vary the property of the FL [35]. In consequence, a non-desired change of the information storage can be presented. To face this problem, an alternative two PL in antiparallel direction and located after the FL are built [35]. This is an alternative structure that has been published a couple of years ago [36]. Nevertheless, no matter the structure topology, to overcome the different stability problems in variability or reliability the fabrication process plays an important role and it is getting very complex. Nowadays, to

construct an MTJ it is built by more or less 15 to 20 layers (some of them are only a few atoms) [34] [36]; a bunch of them will represent the different three MTJ layers we have mentioned before. However, this material science and fabrication perspective are out of the topic of this thesis.

- **MTJ reading and writing**

Understanding the read and write operation is fundamental to know the MTJ device and how it is related to memory devices. Figure 11 exhibits the typical resistive behavior where it is defined two well-defined logic states (High and Low resistance states). Therefore, the reading is done by applying a low bias voltage and seeing the resistance value. Due to the ferromagnetic material, it is not presented a material relaxation that causes a resistance drift that can affect the data storage [35] [37]. The MTJ device has no-presence of this effect so the states remain always the same and the lifetime of the memory is highly reliable. On the contrary, the writing operation uses the spin transfer torque. A current flows through the device and it has two possibilities. The current can simply pass through the device without changing the state; or, if enough current is presented, according to the critical current of the device, the change of state can be done. It is based on the spin momentum transfer, which is explained in detail in the next subsection. However, a brief schematic is shown in figure 12, in which there is a representation of the reading and writing cycle.

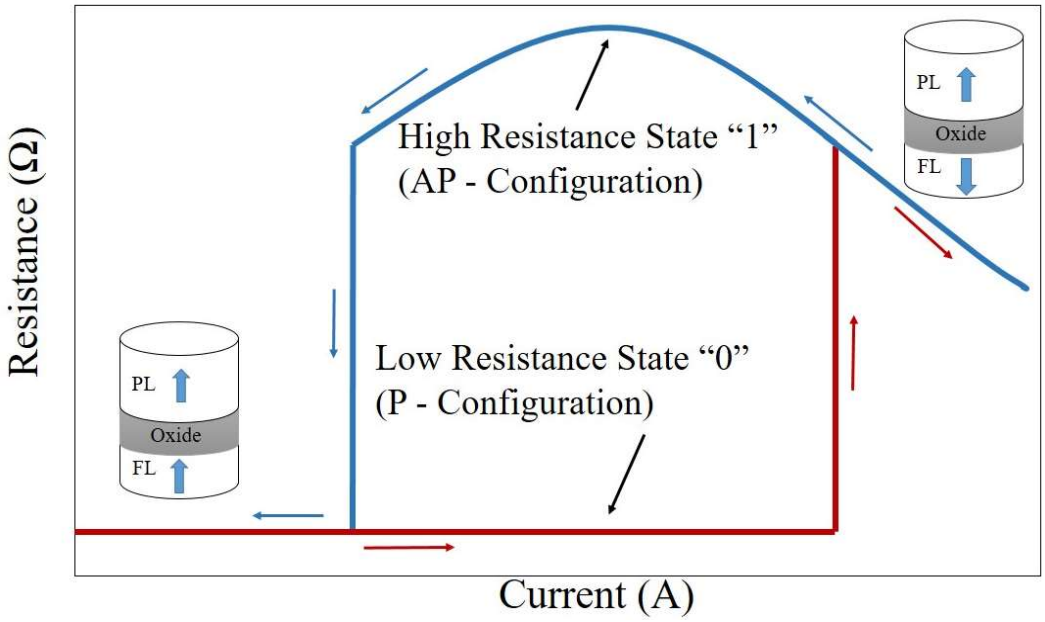

Figure 11: MTJ reading states – Hysteresis resistance characteristic.

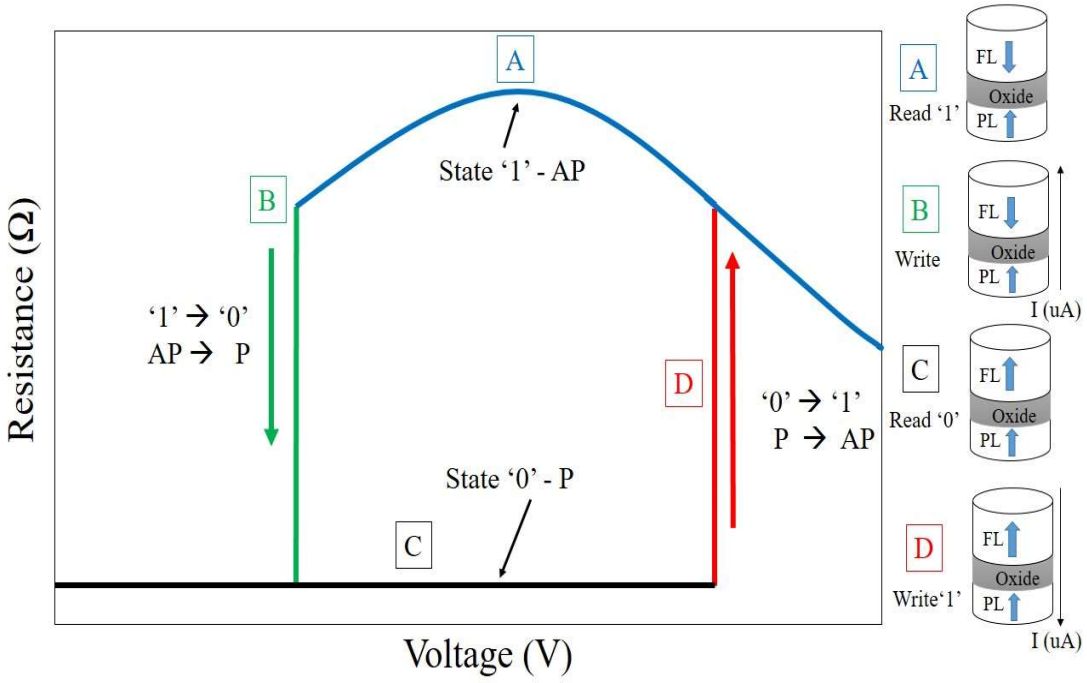

Figure 12: Read and write stages for an MTJ. Note that in the states B and D the device is submitted to a current so the state change can be done.

- **Spin-Transfer Torque (STT) structure**

Different models have been done and tested along the development of MRAM devices.

One of them is based on the change from one state to another by an induced magnetic field $\boldsymbol{H}$;

this is known as Field-Induced Magnetization Switching (FIMS) [6] [26]. A current circulating through a wire generates this magnetic field and with the appropriate value, the switch can happen. Nevertheless, due to a significant amount of wires/interconnections (increasing due to scaling) near to the MTJ a non-desired switching may occur. Thus, FIMS is no longer considered and to encompass the scalability problems STT structure was developed.

STT is focused on another type of switching. The main difference between the last one is that FIMS is based on GMR or TMR where the spin orientation is switched by the magnetic field generated by an electric current. On the contrary, STT uses the spin-polarized current for the change in the magnetization state [27]. To describe the model first is considered a flow of electrons from the PL to the FL. It is known that the ferromagnetic material, which corresponds to the magnetic PL, has a strong polarization capable of polarizing the electron spin. Later the electrons tunnel through the barrier of the material and a torque happens on the FL magnetization causing the alignment of the FL magnetization, $\boldsymbol{m}$, with the PL magnetization, $\boldsymbol{m_p}$. On the contrary, if electrons are traveling from the FL to PL, electrons will try to align to $\boldsymbol{m}$. Figure 13 shows the possible change of states according to the direction of an applied current.

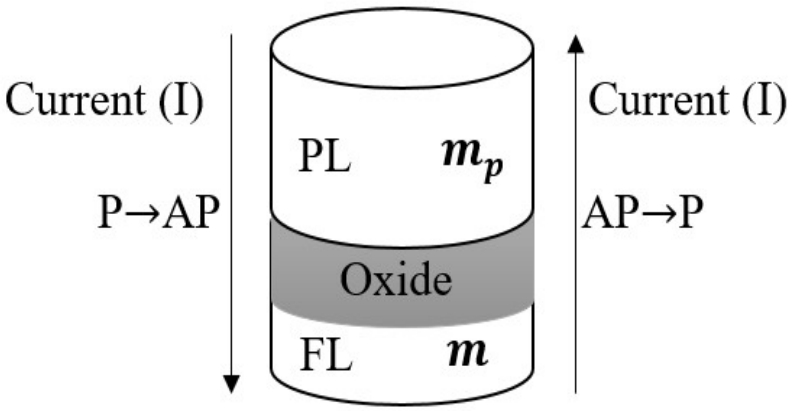

Figure 13: Sketch of the physical MTJ structure considering the direction of the current and the corresponding switching state.

Now, we have learned that the magnetization happens in the FL. In order to understand its behavior, a dynamic magnetization model based on Landau–Lifshitz–Gilbert–Slonczewski (LLGS) equation is used. The LLGS equation takes in account the effects of the STT and has to be solved in order to know the state of the MTJ; the LLGS equation in its simpler form can be written as following [6]:

$$\frac{\partial \mathbf{m}}{dt} = -|\gamma_0|\,\mathbf{m}\times\mathbf{h}_{\mathbf{eff}} + \alpha\left(\mathbf{m}\times\frac{\partial \mathbf{m}}{dt}\right) + \boldsymbol{STT} \tag{1}$$

Where $\boldsymbol{m}$ and is the magnetization of the FL, γ_0 is the gyromagnetic ratio, α is Gilbert's damping coefficient, $\boldsymbol{h_{eff}}$ is the effective magnetic field felt by the magnetic material and $\boldsymbol{STT}$ is the Spin-Transfer Torque term. The solution of the equation is presented in the literature and it is solved by numeric integration; the expression is given by (2) [12]. For simulation purposes, the MTJ modeling can be based on a micromagnetic analysis or an analytical compact model, which is explained in the next chapter; these models are described by the following equation [11], [12], [29]:

$$\frac{d\mathbf{m}}{d\tau} = -\mathbf{m}\times\left[\mathbf{h}_{\mathbf{eff}} - \alpha\frac{d\mathbf{m}}{d\tau} - \beta\frac{\mathbf{m}\times\mathbf{m}_{\mathbf{P}}}{1+c_p\mathbf{m}\cdot\mathbf{m}_{\mathbf{P}}} + \mathbf{h}_{\mathbf{th}}\right] \tag{2}$$

Where $\boldsymbol{m_p}$ is the magnetization of the PL, $\tau = \gamma_0 M_S t$ is the time, M_S is the saturation magnetization, β is the normalized injected current density, $\mathbf{h_{th}} = \nu\boldsymbol{\chi}$ is the thermal field where $\boldsymbol{\chi}$ describes a white Gaussian noise, $\nu = \sqrt{(2\alpha k_B T)/(\mu_0 M_S^2 V_{FL})}$ is the intensity of thermal fluctuations, μ_0 is the vacuum permeability, T is the temperature, k_B is the Boltzmann constant, V_{FL} is the volume of the free layer and $c_p = \eta^2$ is the spin-torque asymmetry description where η is the spin polarization factor. Although, it is important to mention the influence of the temperature. The temperature is always presented and affects directly the

MTJ switching time by changing the reference angle between the magnetization of the PL and FL [6] [11].

2.2 Single vs. double MTJ

Different structures are analyzed in order to improve the STT-MRAM devices. These structures are known as single barrier (SB) and double barrier (DB) MTJs. As we have seen so far, a single barrier MTJ was described. This is the classical device composed by an FL and PL and depending on the direction of the current we can switch it form the P to AP state or the AP to P state. On the other hand, two pinned layers, the Top Pinned Layer (P_{LT}) and Bottom Pinned Layer (P_{LB}), two oxide layers (the top oxide layer, $t_{ox,t}$, and bottom oxide layer, $t_{ox,b}$) and an FL, compose the double barrier MTJ configuration. The pinned layers are oriented on opposite direction between each other, thus taking advantage of the torque strength as the electrons flow through the device. In addition, whether it is an SB or DB, it is always considered as a two-state device associated with a low resistance and a high resistance as it was mentioned before. Figure 14 shows the SB and DB MTJ typical configurations.

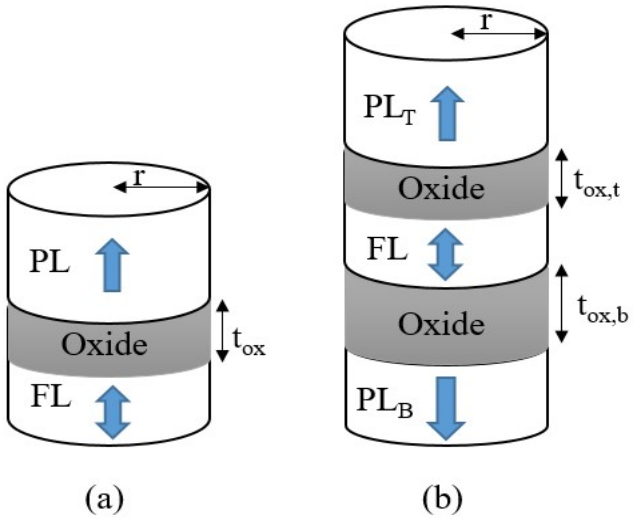

Figure 14: (a) Single Barrier MTJ. (b) Double Barrier MTJ. Both structures are presented with PMA. Note the oxides on the DB are not equal sizing.

The operation principle between SB and DB MTJ are the same. Considering the DB structure, the incident electrons oriented in the same direction as the pinned layer can tunnel through the first oxide. Then the FL, which is on the opposite direction of P_{LT}, will exert a torque to flip with one spin state to another. In this operation, the electrons are favored from PL_B and it is said that the torque is stronger causing a faster switching. In other words, we need less current to make switching. Thus, the principal characteristic of the DB MTJ is the presence of low switching currents. Besides, in the case of SB, the P to AP transition is the slowest because when the electrons enter from the FL we have less polarization efficiency. While in the case of DB it is presented pinned layers on top and bottom terminals, so electrons will always have the presence of a PL no matter in which direction are entering. Another advantage is the use of a low voltage supply for these structures, so they can be considered as low power solutions. Nevertheless, the problem relays on high writing currents either for DB or for SB.

2.3 Spin-Transfer Torque (STT) MRAM

A generic STT-MRAM circuit structure is composed by an SB or DB MTJ and an access transistor. The MTJ can be configured in two ways: a bottom pinned configuration (also known as Standard Configuration - SC) and top pinned configuration (known as Reverse Configuration - RC) as it is shown in figure 15-(a) [6]. The access transistor enables the access to the MTJ by applying a V_{DD} on the transistor's gate. For instance, considering a write operation, the word line of an N-type transistor (WLn) that is connected to the gate, it will be charged to V_{DD} and a current will flow between the bit line (BL) and source line (SL). As it was mentioned in the previous chapter the read and write operations are not decoupled, so they happen on the same path. This leads to a source degeneration ($V_{GS} < V_{DD}$) during one of the write operations when the current is driven by the transistor from SL to BL.

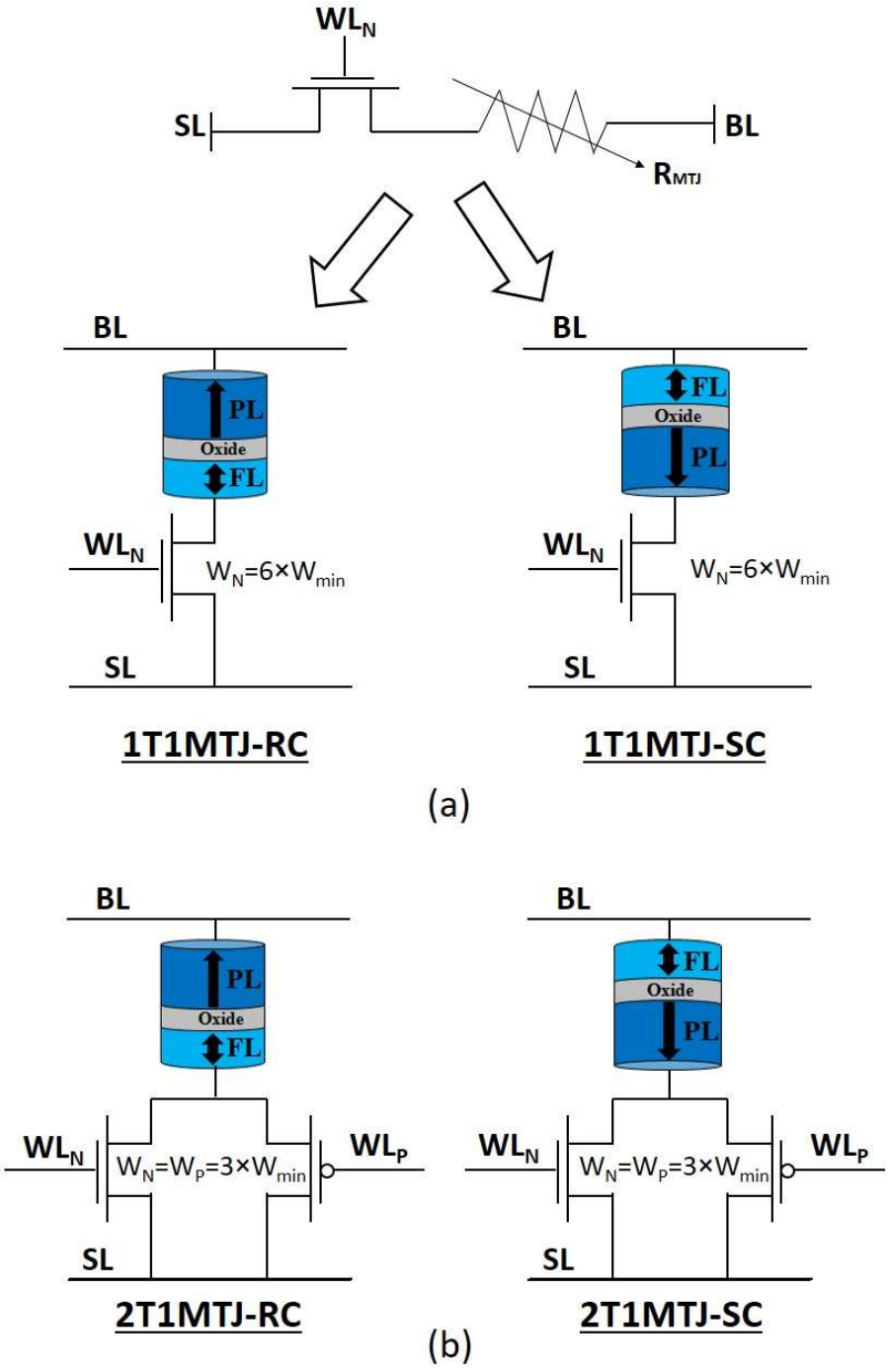

Figure 15: (a) 1T1MTJ-RC and 1T1MTJ-SC. (b) Two-transistor RC (2T1MTJ-RC) and SC (2T1MTJ-SC). Note that in this example the width of transistors is equally balanced according to the 1T1MTJ configuration.

The effect of source degeneration causes a reduction of the write current I_{write}. The source degeneration presented on the bitcell is presented in figure 16, where it is seen that in the case of a DB MTJ, the RC is considered when the PLT is connected to BL and the SC when PLB is connected to BL. In figure 16, due to the current direction, note the source terminal

becomes the terminal of the transistor that is connected to the MTJ. To reduce, control and tolerate the effect of source degeneration, a two-transistor configuration is done (see figure 15 – (b)). Thus, four types of configurations are presented where each one represents an STT-MRAM bit cell.

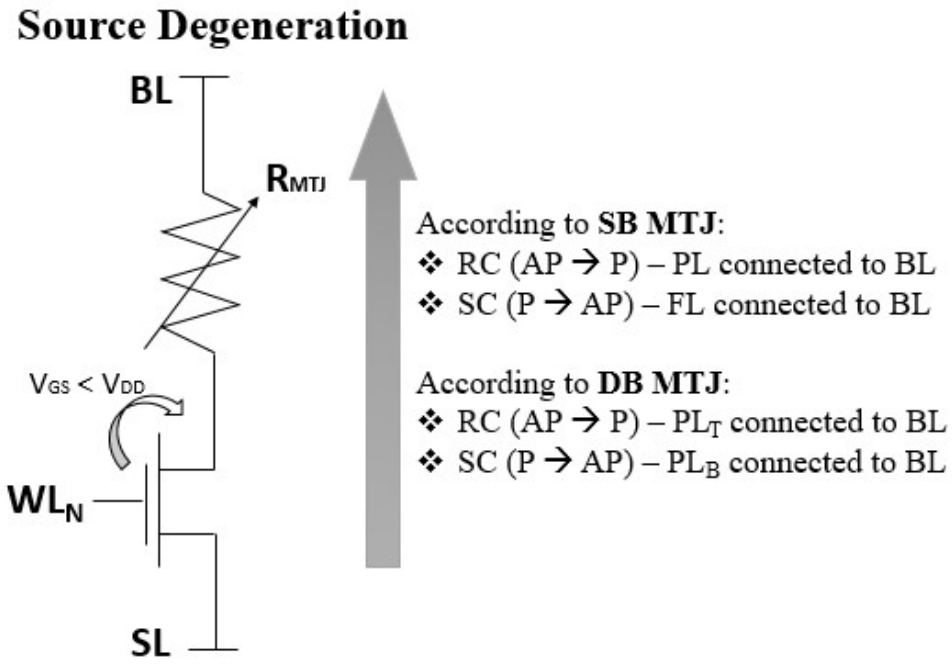

Figure 16: Source degeneration cases for SB and DB MTJ. The arrow represents the flow of electrons from SL to BL.

A typical STT-MRAM array is presented in figure 17 where for each bit cell a 1T1MTJ-RC is used. We notice that the array is similar to the SRAM or DRAM arrays. For this reason, we need a column decoder (for the BL and SL) and a row decoder (for the WL). The WL drives N cells so a write driver is needed to have a good slew rate of the signal. Recalling the 1T1MTJ and 2T1MTJ designs, for the case of 2T, we will have two WL, which means that we have to consider two buffers for each cell. Another important aspect to consider is the two separate drives, one for reading and the other for writing; because the writing needs a maximum current while reading a small current [41].

As we have learned so far, to write in a bit cell is sufficient to apply a current above the critical current of the MTJ, so the change of a state is done. On the contrary, for reading, we mentioned that it is necessary sense the value of the MTJ resistance. For this, a voltage or current sensing can be used. Furthermore, to perform the reading, WL is connected to V_{DD} and a read current will flow in the bit cell. Thus, for the case of a voltage sensing, a voltage drop (V_{drop}) is generated between the BL and SL. So, in order to know the current state of the bit cell, a sense amplifier (SA) is used. The SA will compare between the V_{drop} and the V_{REF} giving an AP ("1") state when $V_{drop} > V_{REF}$ and a P ("0") state when $V_{drop} < V_{REF}$.

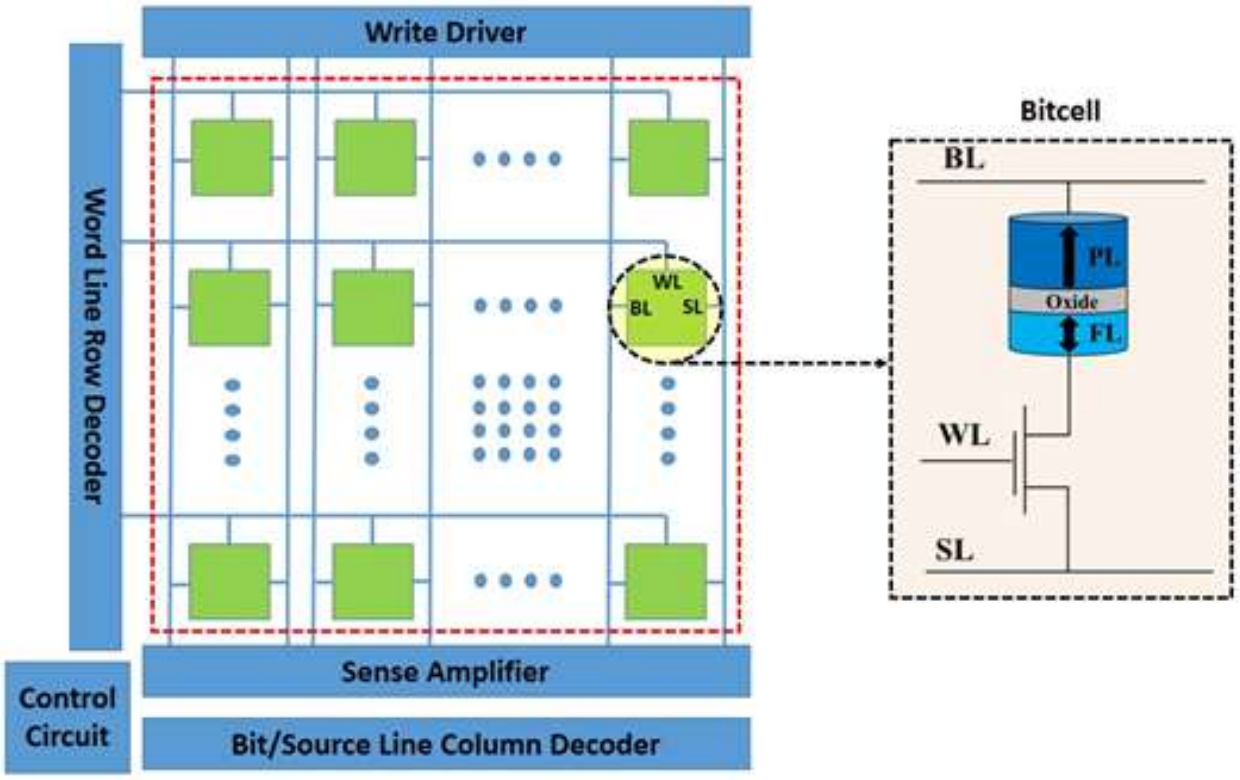

Figure 17: STT-MRAM array considering 1T1MTJ bit cell [29]. The read and write drives considered in a single "Write Driver" block.

In the following chapter, we will see that for the reading analysis it is used a voltage sense amplifier where a fixed current is applied to the cell. This leads to knowing the available sensing margin of the bit cell, which is defined as following [41]:

$$V_{SM} = V_{BL,AP} - V_{BL,P} = I_{read} \cdot (R_{bitcell,AP} - R_{bitcell,P}) \quad (3)$$

CHAPTER 3 - HYBRID CMOS/MTJ MEMORY DESIGN

This chapter provides the approach and simulation methodologies for the analysis of the STT-MRAM. It is mentioned the approach used for knowing the MTJ behavior, which is a requirement to simulate the MTJ circuits. Furthermore, a brief description of the simulation structure is presented. In the end, a comparison between the single barrier and double barrier MTJ is shown.

3.1 Simulation methodology

All the simulations are done in Cadence® – Virtuoso®. Bellow, we will explore the different possible approaches that can be used to accomplish the simulation. Both methodologies are built with a Verilog-A code. However, only one of them is used because of the less computational effort that the simulation exhibits.

- **MTJ circuit approach**

The aim is to build a hybrid circuit design between CMOS and MTJ technologies. It is considered the FinFET technology, where the first step is to take the FinFET model such as the nominal or Monte Carlo that is available by the foundry; in this model, we can modify certain parameters of the device like the number of fingers, length, etc. Then to this model (FinFET model) is associated with a compact model for the MTJ, which is written in Verilog-A. This Verilog-A model describes the behavior of the MTJ. It is possible to use two models, the first is by using a microspin approximation, which translates to a Look-Up Table (LUT) methodology [38], and the second one is the use of an analytical compact model. The microspin approximation is based on a micromagnetic analysis, which includes the appropriate MTJ behavior and a proper statistical distribution for the MTJ switching delay is seen [11] while the analytical compact model that is described on the literature [29]. As part of Verilog-A model,

it is also possible to modify the dimensions and different parameters of the MTJ device like the dimensions, temperature, resistance, etc. In the end, the combination of these two models, FinFET and MTJ, allows the simulation of the hybrid circuit design. This process is mandatory when it is required the introduction of an MTJ model because until now it is not commercially available and in consequence, there are not commercial models. The same happens when we are trying to use tunnel FETs.

- **Look-up table (LUT) based methodology**

As it was mentioned before, to properly achieve the correct MTJ behavior, and in consequence the model, a Verilog-A file is needed. This file contains all the code necessary to calculate the current of the MTJ either SB or DB. In addition, a LUT containing the necessary data to calculate the switching time is used. Figure 1 exhibits a block diagram description for this kind of approach.

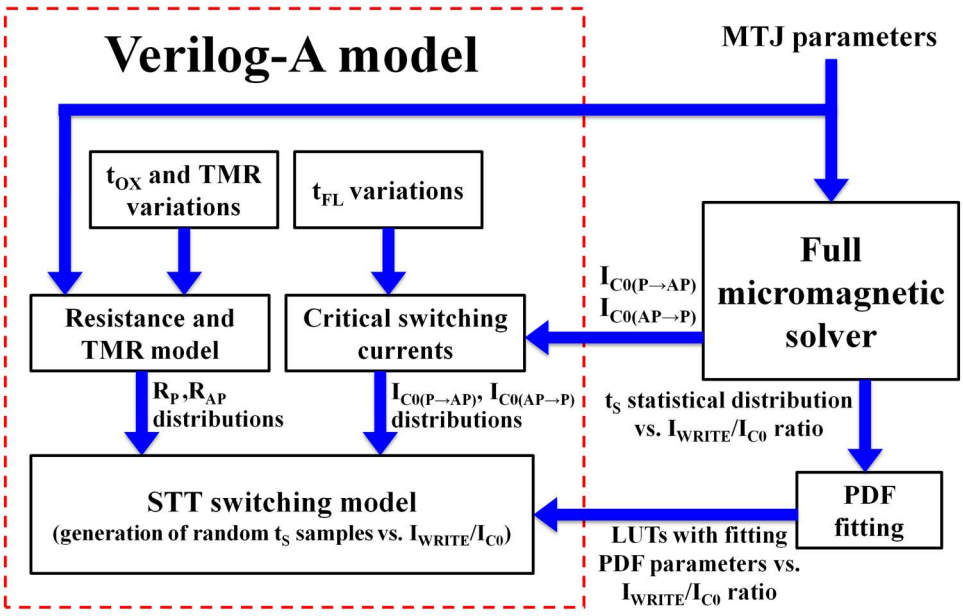

Figure 18: LUT-based methodology block diagram [38].

The LUT is a discretized table, which contains the data of the switching time according to a specific current. The look-up table is made for the transitions AP to P and vice versa.

Besides, for each current value, it is presented different moments such as mean, standard deviation, and the skewness. However, we will have values that do not match on the table. For instance, if we have a current value between 10 and 12 μA, the simulation calculates the value by linear interpolation. This is the reason whereby a good trend of the curves is needed; if not, the value obtained will have a notorious error. Then, taking the values of the three momentums (mean, standard deviation and skewness), the skew-normal distribution (distribution that fits better the micro-magnetic data) is built. Finally, a random sample is taken from the distribution and we get our switching time. As we can deduce, this will considerably increase the calculation time in the simulations.

- **Deterministic and transient analysis**

Two type of simulations are presented, the deterministic and statistical analysis. Considering the LUT based method, when the normal transient analysis is done, also named deterministic analysis, the MTJ will switch to the mean value for a certain current called from a LUT. On the other hand, on the statistical or Monte Carlo analysis, we need to generate N samples or N executions and for each sample, the simulator delivers a unique sample of the switching time relative to the chosen and described distribution in the code. For instance, simulating a memory cell where the applied current is 26 μA, we get that the MTJ will switch, in mean, at 1.5 ns. Making the deterministic analysis is said that applying the current, after 1.5 ns, the change of state is done. On the contrary, making a Monte Carlo Analysis, for each sample we will have different values. Now, we need to determinate the distributions, if the Gaussian distribution is chosen, the simulator will pick up the mean and sigma. Then, for each sample, a statistical random value is created from the obtained distributions as we mentioned previously. The same occurs when we choose diverse distributions like Erlang, but in that case, more moments will be present.

Distributions such as Gaussian or Erlang have their own function on Verilog-A. On the other hand, for the case of the skew-normal distribution, it is necessary to implement it. For this, an alternative method is applied to generate samples based on the skew-normal distribution. However, that topic is beyond the scoop of the simulations presented in this thesis.

- **Analytical compact model**

For a correct modeling of the switching process is necessary to establish a pertinent model. The LUT-based methodology presented above presented a high computational effort due to the characterization through the micromagnetic simulations [29]. The established analytical compact model overcome the issue due to the computational effort and it is easy to integrate into the simulator.

In the figure shown below is the complete block description of the analytical compact model. It is a generic block, which is valid for SB and DB configurations. The only variation between those two is the change of the "Resistance and bias-dependent TMR" block and the "Analytical Formulation". It considers five important effects that act on the switching behavior of an MTJ device. These effects are the MTJ process variations, the spin-torque asymmetry in the switching process, the temperature dependence, the thermal heating or cooling and the voltage-dependent because of the perpendicular magnetic anisotropic [29]. One of the most important parameters between them is the switching process, which entails a statistical switching model divided into two regimes, the thermal activation regime, and the fast switching regime. The first one is for injected currents (I_{MTJ}) below the critical current (I_c) and follows the Nèel-Brown model. This can be understood as the current for reading data. On the

contrary, for currents above the critical current, which means a writing operation, it is used an extended analytical formulation [29].

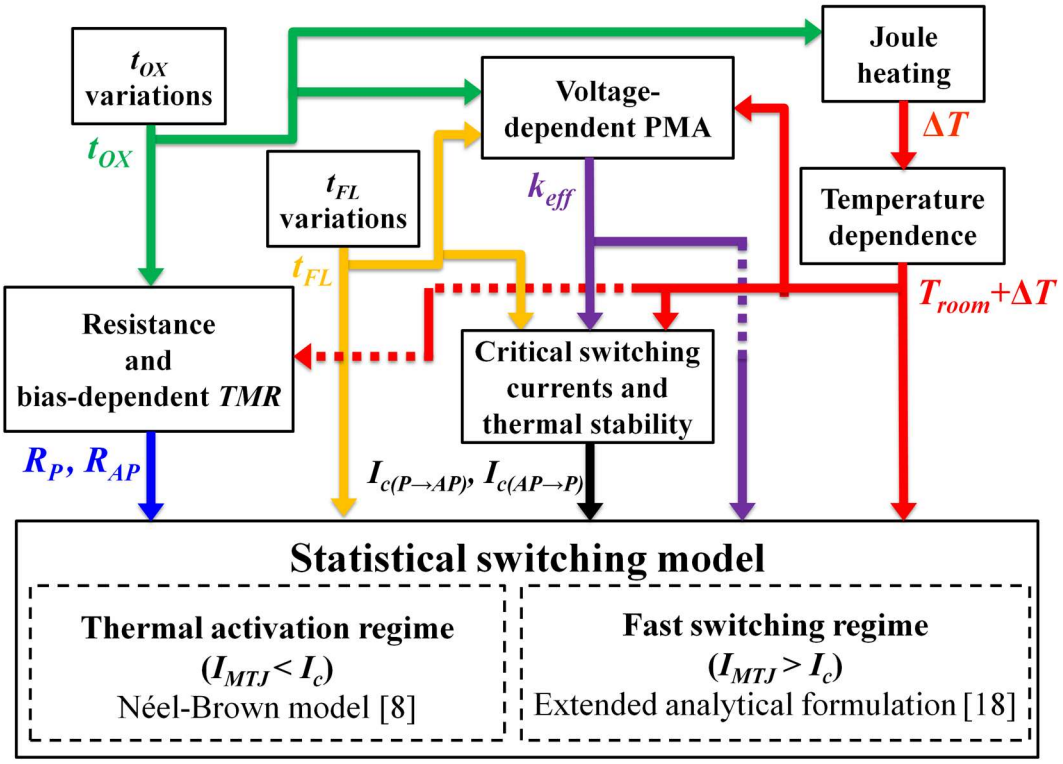

Figure 19: Analytical compact model block description [29].

The compact model gives the statistical distribution for the transition AP-P and P-AP. Furthermore, is capable of distinguishing a deterministic and stochastic behavior by giving an initial magnetization angle, making possible a deterministic simulation for the MTJ switching [29]. So far the analytical formulation for the fast switching fits well for currents slightly higher the critical current; however, there is not a model capable of describing the region between the thermal activation regime and the fast switching regime.

3.2 Model validation

The SB and DB follow compact models in order to get the appropriate behavior of the STT switching activity. The model validation is done by comparing the micromagnetic simulations and the analytical predictions [43]. Once the validation is done the respective

simulation and compassion between SB and DB MTJs is presented later. Note that for validation purposes, to follow the experimental data, the validation is done for certain values of MTJ physical parameters. Later, it is shown the MTJ parameters that are used for the MTJ.

To validate the model, two validations are presented. The first one is the resistance and TMR model validation by comparing the models with the experimental data presented in [43]. On the other hand, the second validation is the comparison with the analytical STT switching model with a full micromagnetic solver. Remember we have SB and DB MTJs, which means that each one has to be validated. In the following, only the DB MTJ model validation is done. The figure 20 exhibits the first validation considering the following parameters: $t_{ox,t}$ = 0.8 nm, $t_{ox,b}$ = 0.75 nm, a bias voltage for the TMR V_H = 0.5 V, $TMR_{T(0)}$ = 140%, $TMR_{B(0)}$ = 80%, a resistance-area product top RA_T = 100 Ω·μm^2 and resistance-area bottom RA_B = 50 Ω·μm^2 [43].

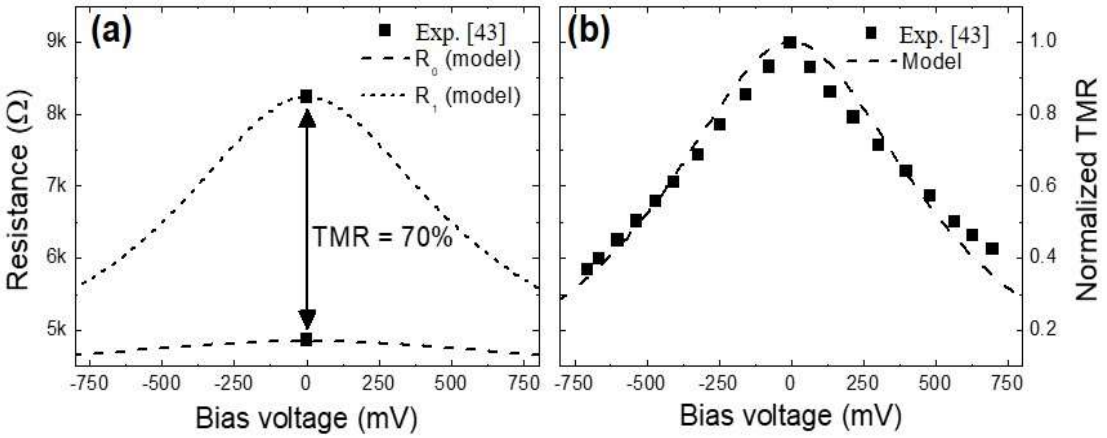

Figure 20: Resistance and TMR model validation in comparison with the experimental data [43].

On the contrary, for the STT switching model validation of the DB, figure 21 justifies it with three different MTJ radius (r = 12 nm, 10 nm and 7 nm). Furthermore, it is considered the following parameters: a saturation magnetization $M_S = 10^6$ A/m, α = 0.03, K_u = 1.1×106 J/m^3, a free layer thickness t_{FL} = 1.2 nm and η = 0.67 [43]. We can see that the moments follow the results of the micromagnetic solver.

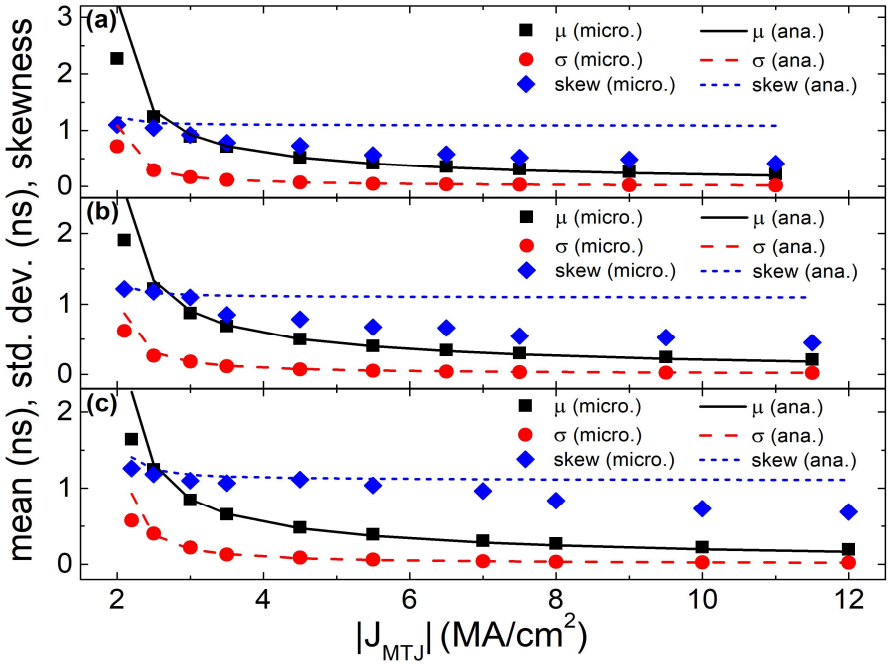

Figure 21: Data validation for a DB-MTJ with Mean value (μ), standard deviation (σ) and skewness (skew) of the switching time as a function of the MTJ current density [43]. Three dimensions considered: (a) r = 12 nm, (b) r = 10 nm, and (c) r = 7 nm.

3.3 Simulation structure and CMOS/MTJ parameters

Through this chapter, we had established the simulation methodology and its model validations for an SB and DB according to the literature. Now it is shown the hybrid CMOS/MTJ parameters used for the analysis and the simulation framework, which will be used in the rest of the thesis.

- **CMOS/MTJ parameters**

Starting with the MTJ model, table 3 specifies the parameters for an SB and DB MTJs. In order to match with the CMOS technology node, which in our analysis is 28 nm, the MTJ radius is chosen as r = 14 nm. In addition, from table 3, the current and thermal stability were fitted according to the experimental data informed on [44]; the obtained values can be extracted from figure 21. Moreover, it is included a variability in percentage for several parameters; these variations are included in the MTJ compact model.

Table 3: SB and DB MTJ parameters for a single dimension of r = 14 nm.

SB	DB			
Parameter		**Description**	**Value**	**Units**
M_S		Saturation magnetization (300 K)	1×10^6	A/m
α		Gilbert damping factor	0.05	--
r		MTJ radius	14	nm
surface		MTJ surface (variability)	1.23 (5%)	μm^2
t_{ox} (σ/μ)	--	Oxide thickness (variability)	0.85 (1%)	nm
--	$t_{ox,t}$ (σ/μ)	Top Oxide thickness (variability)	0.85 (1%)	nm
--	$t_{ox,b}$ (σ/μ)	Bottom Oxide thickness (variability)	0.65 (1%)	nm
t_{FL} (σ/μ)		FL thickness (variability)	1.2 (1%)	nm
RA	--	Resistance-area product	5.0	$\Omega\cdot\mu m^2$
--	RA_t	Top Resistance-area product	5.0	$\Omega\cdot\mu m^2$
--	RA_b	Bottom Resistance-area product	1.0	$\Omega\cdot\mu m^2$
R_P	--	SB-MTJ resistance in P state	8.12	kΩ
R_{AP}	--	SB-MTJ resistance in AP state	20.3	kΩ
--	R_0	DB-MTJ resistance in P state at V = 0 V	9.38	kΩ
--	R_1	DB-MTJ resistance in AP state at V = 0 V	20.8	kΩ
TMR_0	--	TMR ratio (300 K and 0 V)	150% (3%)	--
--	$TMR_{0,T}$	Top TMR ratio (300 K and 0 V)	150% (3%)	--
--	$TMR_{0,B}$	Bottom TMR ratio (300 K and 0 V)	150% (3%)	--
Δ^*		Thermal stability	59.14	--
V_H		Bias voltage for $TMR = 0.5\times TMR(0)$	0.5	V
η		Spin-polarization factor	0.67	--
$N_{x,y}$		In-plane demagnetizing factor	0.042356	--
N_z		Perpendicular demagnetizing factor	0.915288	--
k_{eff}		Effective anisotropy (300 K and 0 V)	0.5276	--
$J_{c(P\rightarrow AP)}{}^*$	--	P→AP critical current density	6.53	MA/cm^2
$I_{c(P\rightarrow AP)}{}^*$	--	P→AP critical current	40.21	μA
$J_{c(AP\rightarrow P)}{}^*$	--	AP→P critical current density	2.48	MA/cm^2
$I_{c(AP\rightarrow P)}{}^*$	--	AP→P critical current	15.3	μA
--	$J_{c(AP\leftrightarrow P)}{}^*$	AP↔P critical current density	1.8	MA/cm^2
--	$I_{c(AP\leftrightarrow P)}{}^*$	AP↔P critical current	11.08	μA

T_{room}	Room temperature	300	K

*This data was fitted according to the experimental data reported on [44]

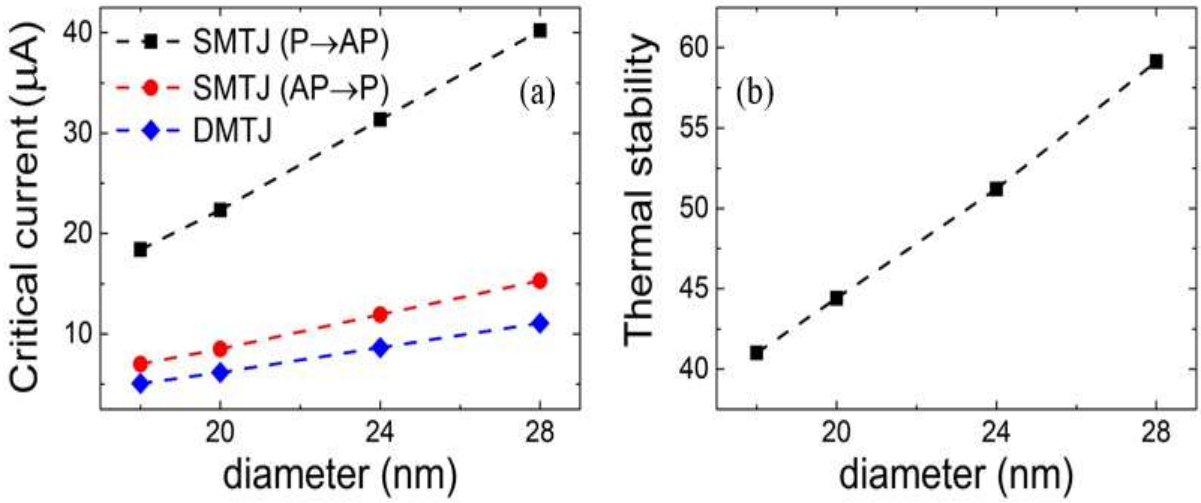

Figure 22: Fitted data according to [44]. (a) Critical current vs. MTJ diameter. (b) Thermal stability vs. MTJ diameter.

For the MOS technology case, it is used the available FinFET technology node. The table 4 summarizes the parameters of the FinFET used independently if it is an NMOS or PMOS device. In table 4, only the most important parameters are mentioned. In the case of *nfin* or *m* values, they are used as default for all the analysis. On the other hand, *nf* is the only parameter that varies on the simulations. According to it, the transistor area, and in consequence the bitcell area, changes.

Table 4: FinFET parameters used for access transistor in memory cells.

Parameter	Description	Value	Units
L	Gate Length	28	nm
nfin *	Number of Fins per Finger	2	--
nf **	Number of Fingers	1	--
m	Multiplier – Number of parallel MOS devices	1	--

* Corresponds to the width of each finger and it is expressed in integer units.
** Corresponds to the number of gate fingers presented in the layout.

Now that we have the parameters of our hybrid model, we can establish the STT-MRAM layout parameters. Generally, when measuring the bitcell area, the MTJ size is not taken in account, instead, the area is limited by the transistor dimensions or the metal pitch [41]. The layout parameters are used to calculate the minimum technology feature size (F) which will be later used to represent the cell area. Thus, the F is defined as [45]:

$$F = \frac{1}{2}\left(W_{\min,M1} + S_{\min,M1}\right) \tag{4}$$

Where $W_{min,M1}$ is the minimum width of the Metal-1 layer and $S_{min,M1}$ is the minimum spacing of the Metal-1 layer. For the technology used we have that F = 32 nm[†]. Additionally, in table 5 are shown the values considering the F for a default transistor; that means, the values are calculated considering the parameters exhibited in table 4.

Table 5: STT-MRAM layout parameters

Parameter	Description	Value	Units
$W_{min,bitcell}$	Minimum bitcell width	4.94	F
$H_{min,bitcell}$	Minimum bitcell height	5.94	F
$A_{bitcell}$	Minimum bitcell area $W_{min,bitcell} \bullet H_{min,bitcell}$	29.32	F^2

- **Simulation structure**

As we have mentioned previously, the simulations are done based on the analytical compact model described in section 3.1. With this model, the memory designed is summited to a deterministic and statistical analysis, where the process variations are studied by using Monte Carlo simulations. The process variations for the MTJ are included in the analytical model written in Verilog-A while in the case of the FinFETs, the foundry provides the statistical models.

[†] For Cadence-confidentiality reasons the $W_{min,M1}$ *and* $S_{min,M1}$ are not shown.

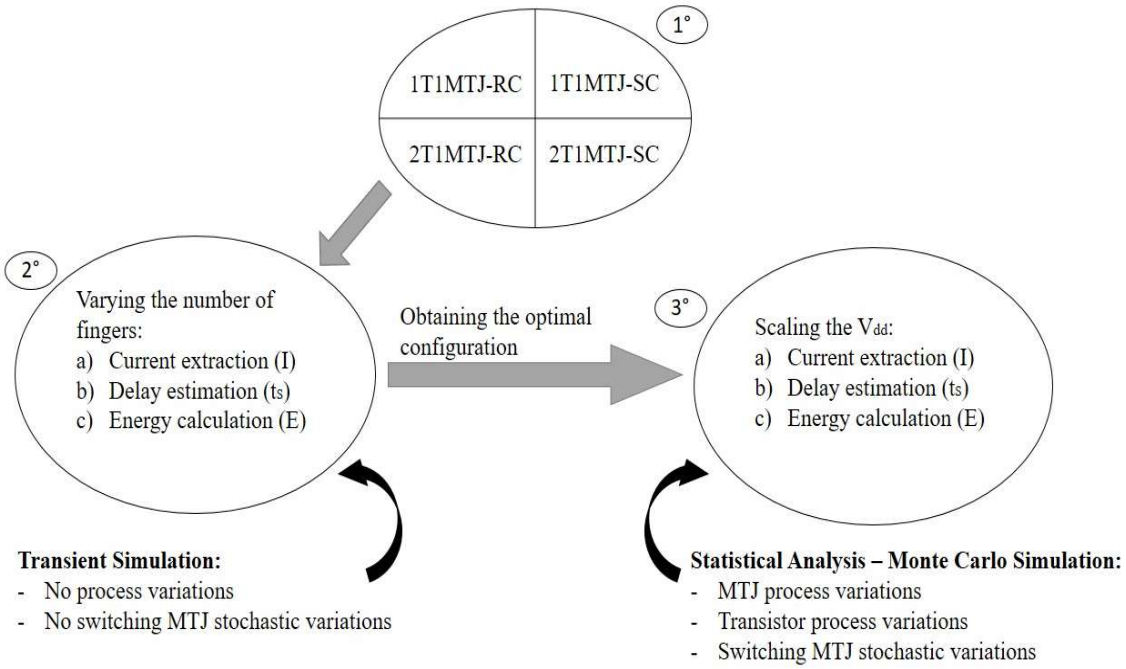

Figure 23: General workflow for the writing analysis of the STT-MRAM independently if it is SB of DB MTJ.

Two kinds of studies are done, the writing and reading. The writing study is described by a general workflow exhibited in figure 23. We start with the four-bitcell configurations mentioned previously and with a transient analysis, the optimal configuration by looking for the best energy option is chosen. Then, the optimal configuration is analyzed by scaling the V_{dd} in order to find the minimum energy point. Furthermore, all this workflow process is done by using an SB or DB MTJ.

Finally, for the reading study, we start with the optimum configuration for SB and DB. Unlike the previous workflow, this consists only in the calculation of the available sensing margin of the STT-MRAM. In figure 24 is shown the corresponding workflow. Note the start point is the optimal configurations obtained in the writing analysis. The writing and reading workflows presented contains a general perspective of what it is done and in the next chapter, we will see in detail all the analysis done for the single 28nm technology node.

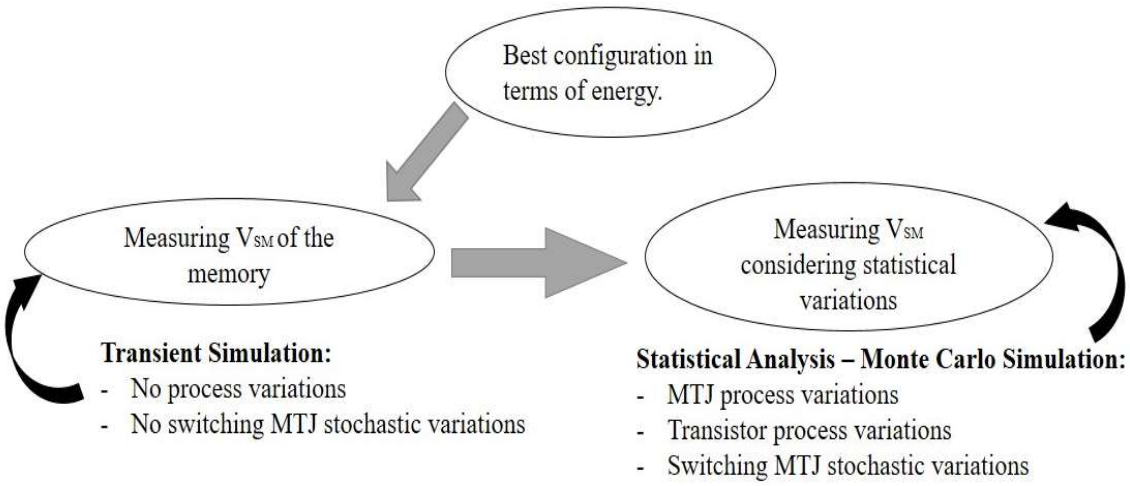

Figure 24: General workflow of the STT-MRAM reading analysis for SB and DB.

CHAPTER 4 – STT-MRAM ANALYSIS

This chapter provides the simulations and main results of the thesis. As it was mentioned in the previous chapter, from the foundry, it is taken the 28 nm node technology and a Verilog-A code was made for a 28 nm MTJ. For this node, the analysis of four configurations for SB and DB MTJ is done. It is important to mention that the critical current of SB and DB is considered very high, setting the STT-MRAM in a very pessimistic case where the MTJ is considered with a low damping.

4.1 STT-MRAM writing analysis

- **Initial considerations & preliminary analysis**

The topology under test is a 128×128 memory array. For simulation purposes a single bitcell is built with the corresponding buffer lines and the peripheral capacitances for each line (WL, BL and SL), thus it is represented our 128×128 memory block as depicted on the example of figure 25. The terminal T1 (T2) represents the PL (FL) and the terminal named "State" is to know in which state is the MTJ. The buffers were sized strong enough to have the same rise time in all the cases. Moreover, the capacitance values depend on the number of fingers (*nf*) of the access transistor. Hence, if we increase the area of the bitcell, the capacitance will increase and the peripheries will see a greater capacitance.

There is synergy between the buffers and capacitance design. First, the capacitances have to be designed. We have a memory block of 128x128 bit cells, so the WL sees 128 transistor's gate, the SL sees 128 source terminals of the transistor while the BL is always connected to the access transistor. It is true that in the 2T configuration it is used two transistors but at the end, each one sees 128 WL. In table 6 is listed all the capacitance values for the 1T and 2T cases extracted from the 28 nm transistor considering the FinFET default values listed in table 4 of the previous chapter. In the 1T case the $C_{WL} = (C_{gs} + C_{gd})\times 128$, $C_{SL} =$

$(C_{sd} + C_{sg})\times 128$ and $C_{BL} = C_{SL}/10$; for this last one, the C_{BL}, we do not know the capacitive effect of the MTJ so it is lowered a decade according to the C_{SL}. On the other hand, for the 2T design the $C_{WLn} = (C_{gsn} + C_{gdn})\times 128$, $C_{WLp} = (C_{gsp} + C_{gdp})\times 128$, $C_{SL} = 2(C_{sd} + C_{sg})\times 128$ and $C_{BL} = C_{SL}/10$. Now, the buffer design takes into account the capacitance values. The reader may notice that the SL buffer has less drive strength (the half) in comparison with the WL buffer; this is because the C_{SL} is more or less the half of the C_{WL}. Finally, the BL buffer is 10 times less than the SL buffer as depicted in figure 25. On the other hand, when using 2T, BL and SL remain the same as in the 1T case while the WL buffer is divided in two similar buffers with WLp buffer smaller than the WLn buffer according to the C_{WLn} and C_{WLp} values illustrated in table 6. With all these considerations, the rise time in the lines will be the same, assuring a good simulation environment for the memory analysis.

Table 6: Capacitance values extracted from the 28 nm transistor.

According to 1T NMOS				
Parameter	**Description**	**NMOS**	**PMOS**	**Units**
		Value		
C_{WL}	Word line capacitance	35.14	23.72	fF
C_{SL}	Source line capacitance	17.59	11.86	fF
C_{BL}	Bit line capacitance	1.759	1.186	fF
According to 2T				
Parameter	**Description**	**Value**		**Units**
C_{WLn}	Word line NMOS capacitance	35.14		fF
C_{WLp}	Word line PMOS capacitance	23.72		fF
C_{SL}	Source line capacitance	35.17		fF
C_{BL}	Bit line capacitance	3.517		fF

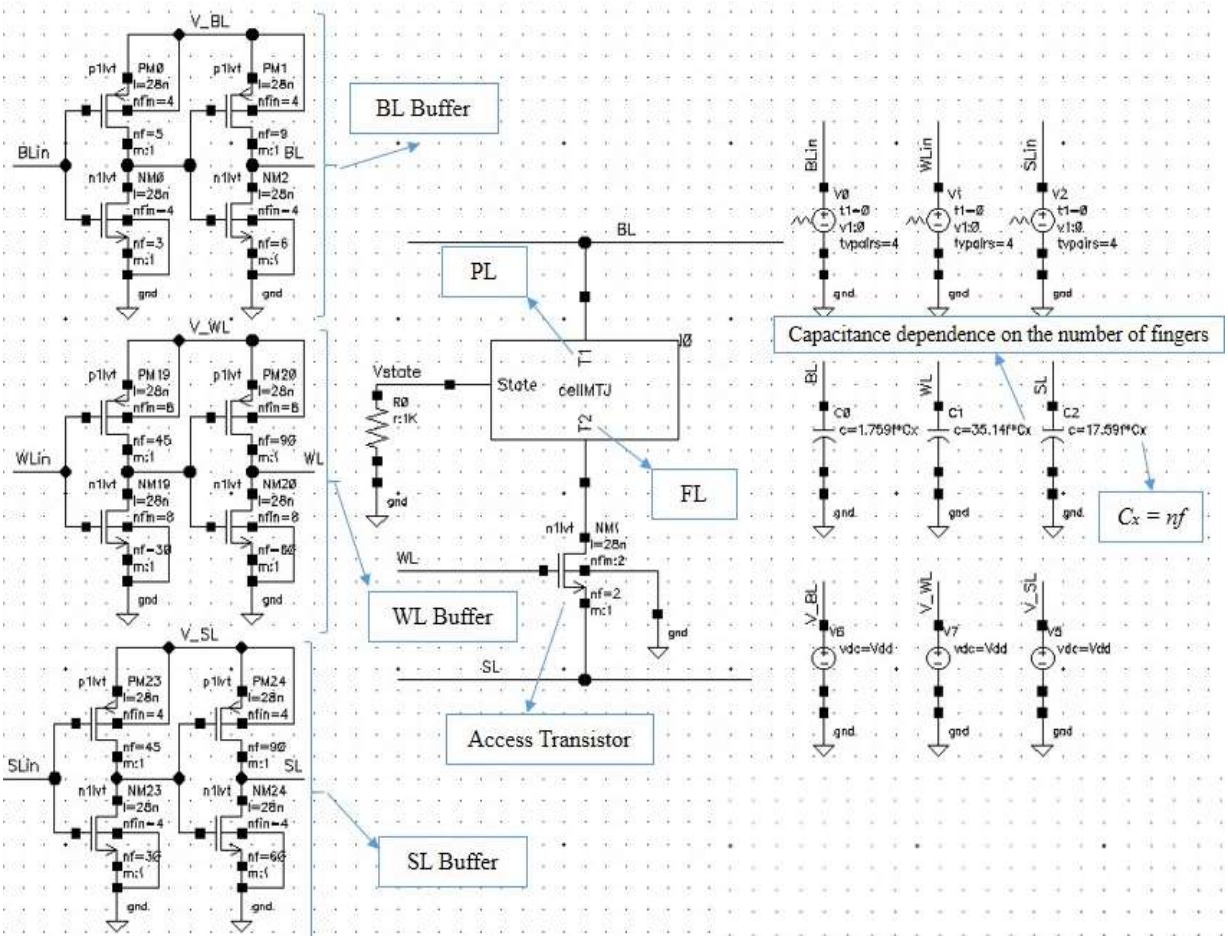

Figure 25: Writing simulation schematic representing the 128×128 STT-MRAM using the 1T1MTJ-RC bitcell. For this configuration, the P – AP transition is done by a positive pulse from BL to SL, while for the AP – P transition a positive pulse from SL to BL is used. Note we have three generators, one for each line, so the contribution of the energy is calculated independently.

The last consideration is in the energy calculation. We need to take into consideration the periphery circuit. In the following, a brief explanation why; for instance, when the writing is done, we access a line inside of the 128×128 STT-MRAM, three lines are considered. The WL energy it is divided for the number of MTJs presented in the line (taking into consideration the case where all the 128 are switched on). On the contrary, for the SL and BL contains the transient signal, which will travel through the access line and will arrive at the bitcell considering all the energy of the SL and BL buffers; in other words, all the energy of the lines is needed to write into the bitcell.

Before the writing and reading study, a preliminary analysis is done in order to know a general behavior of the MTJ. As we have noticed from table 3, SB and DB have the same FL and PL (referring to the PL_T in the DB case) and as we will see in the following results, for the SB case, it is not presented the result. Hence, the only parameter that is varying in this pre-analysis is the PL_B ($t_{ox,b}$) on the DB, which is shown in figure 26. This tells us that one barrier is making more resistive than the other one and by changing the $t_{ox,b}$, the total resistance of the MTJ is changing. This allows an increase on the TMR and in consequence the current, which is an advantage. Thus, it can be said that the DB can be adjusted by varying the smallest oxide layer. Obviously, we cannot go as far as we want, a breakdown of the MTJ can occur; however, this is beyond the scope of our analysis.

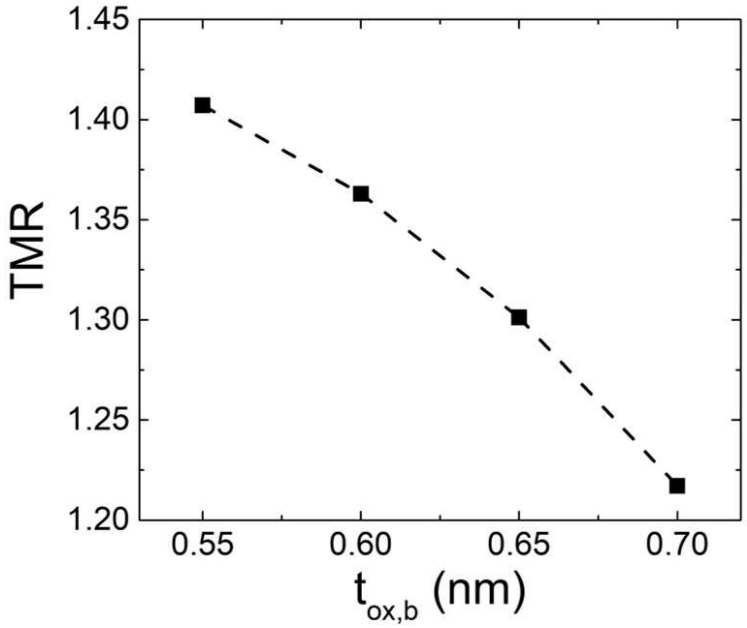

Figure 26: TMR vs. $t_{ox,b}$ of the DB MTJ.

- **Writing deterministic analysis**

All the analysis follows the general workflow shown in figure 23. The purpose is to vary the integration density with the number of fingers, which is later translated in cell-area units (F^2). The first step is to start with a deterministic simulation and extract the current as the integration density varies. In figure 27 are reported those results, we can see that the SB does

not have a good performance for small transistors, or in other words, the performance decreases for a reduced number of fingers. For 2T1MTJ configurations exhibits a better performance than the 1T1MTJ configurations.

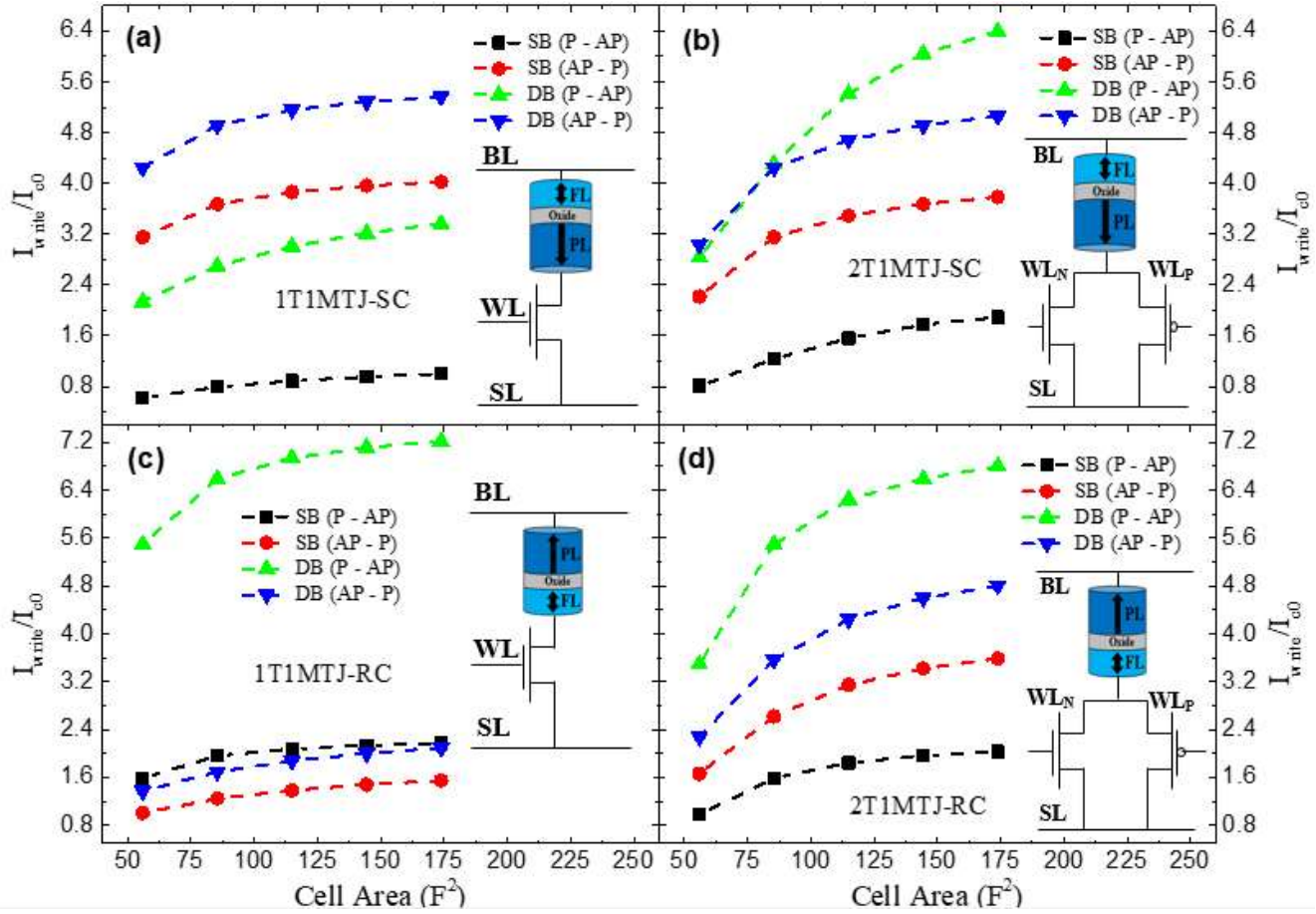

Figure 27: I_{write}/I_{c0} vs. Cell Area for the different SB and DB configurations.

In general, with these results we can understand how much area is necessary to have a certain current. Furthermore, the DB starts from two or three times the critical current, which confirms one of the problems mentioned in previous chapters, which was the presence of high writing currents. Last but not least, for SB and DB a bunch of simulations was included parallel to this deterministic or transient analysis. These simulations were summited to the variation of the oxide (oxides) for the SB (DB) and it was repeated the same calculations mentioned before. The current increases as the oxide/s decreases and the expected

performance improvement of the TMR illustrated in the preliminary analysis, and in consequence, the bitcell improvement is seen.

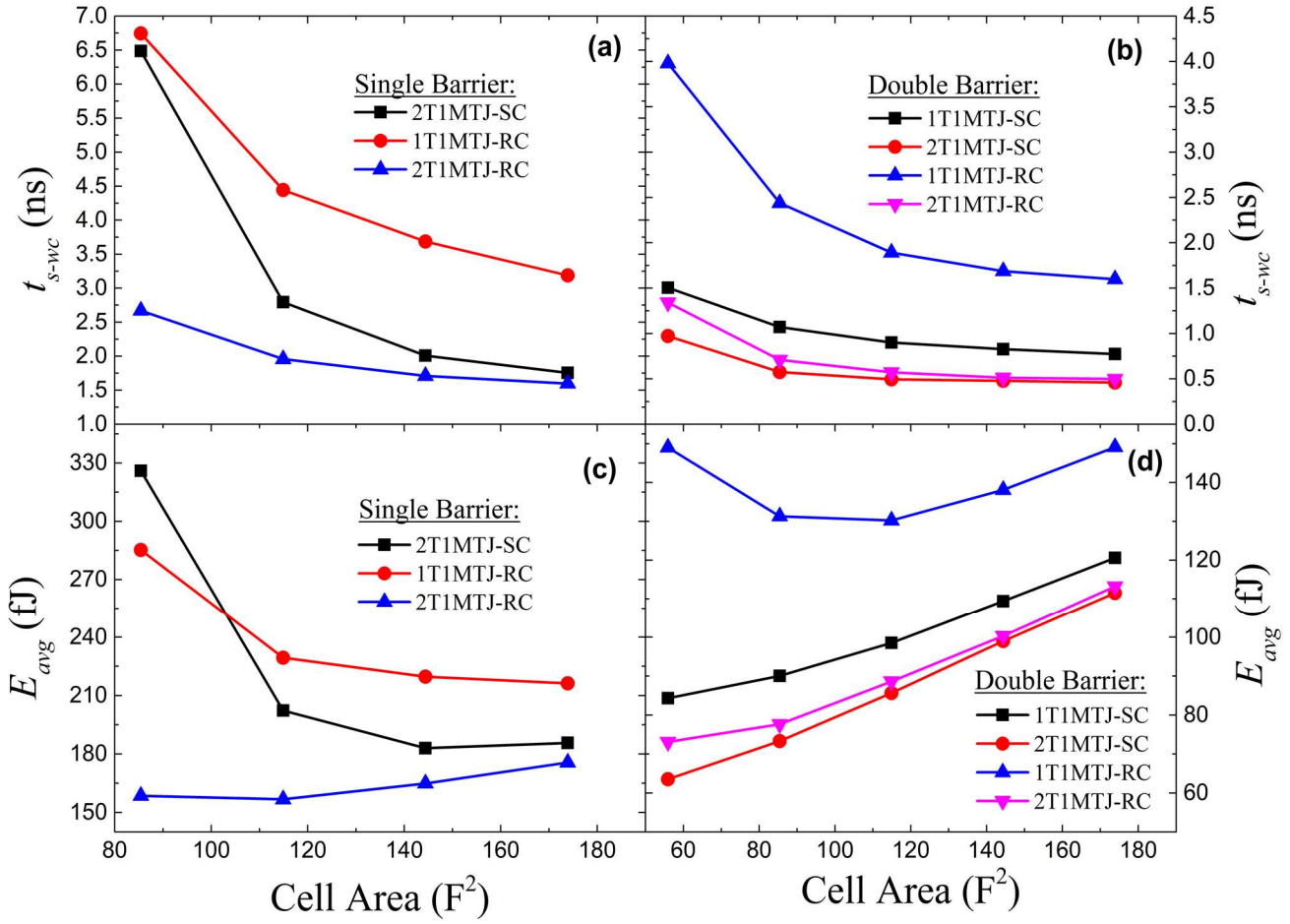

Figure 28: Deterministic analysis for (a)-(b) Worst case delay ($t_{s\text{-}wc}$) vs. Cell Area and (c)-(d) Average energy (E_{avg}) vs. Cell Area for the different SB and DB configurations.

Now for each current value, which corresponds to a certain cell area, the switching delay or switching time (t_s) of the bitcell has to be calculated. The t_s calculation for the different configurations, and with the help of the current previously calculated in comparison with the critical current of the SB or DB, will show us which configurations will write and which do not. To obtain the t_s a calculation external to cadence is done, where with a MATLAB script (See Annexes) the t_s is calculated for the AP – P and P – AP transitions. With the MATLAB script a switching time CDF is done, which shows the switching probability error. Now, we are interested in being on the tail of the distribution, where we have a t_s much greater than the t_s

mean, so it is chosen a write error rate (WER) value of WER = 1×10^{-6}, which tell us that we have an error probability of (1 – P) or (1 – 1×10^{-6}). Note that t_s is calculated for the AP – P, and P – AP transitions, so at this point, it is necessary to consider the maximum delay between these two. This maximum delay is considered the worst-case delay, $t_{s\text{-}wc}$. Figure 28 a-b shows the worst-case delay for SB and DB and we can notice that our initial configuration testing has been reduced.

Finally, the average energy is calculated for the worst-case delay (at a WER = 1×10^{-6}) between the two transitions (AP – P and P - AP) and the results are exhibited on figure 28 c-d. It is important to mention that the energy calculated is considering the peripheries of the STT-MRAM as it was mentioned before. These Results give us the optimal configuration in terms of energy. The best configurations are 2T1MTJ-RC and 2T1MTJ-SC for SB and DB respectively. In this stage, the dimension is fixed, which is to say that the memory integration capacity is defined. For instance, for the SB (optimal energy point) and DB we choose an area cell of 115 F^2 to compare in terms of area parity. In addition, for the DB (optimal energy point) case, an area cell of 56 F^2 is chosen. The reader can see that for the SB case it is not presented the corresponding area cell like the DB or can be said that the SB does not scale for a maximum integration capacity. In the case of maximum capacity, the SB does not write as we can see in the first analysis of figure 27.

- **Writing statistical analysis**

Now, the statistical analysis is done by using Monte Carlo simulations. For the last two configurations chosen and with an integration capacity defined, the V_{DD} is scaled as it is shown in figure 29. The analysis made is the same as before, where the t_s and later the worst-case delay are calculated for later obtain the average energy. It is seen that scaling the V_{DD} the current decreases and in consequence, the delay increases. If we fix a voltage (e.g VDD = 0.75

V), we can see the DB is better in terms of speed. At the same time, the average energy scales with the V_{DD} where a point of minimum energy can be found where the best option is the DB. The summary of this results is listed in table 7. Here it is seen at parity of area, the DB has also a better energy performance with a write energy saving of 67.2% when we move from SB to DB. Furthermore, we even see energy savings in the case of maximum integration (56 F^2) for the DB.

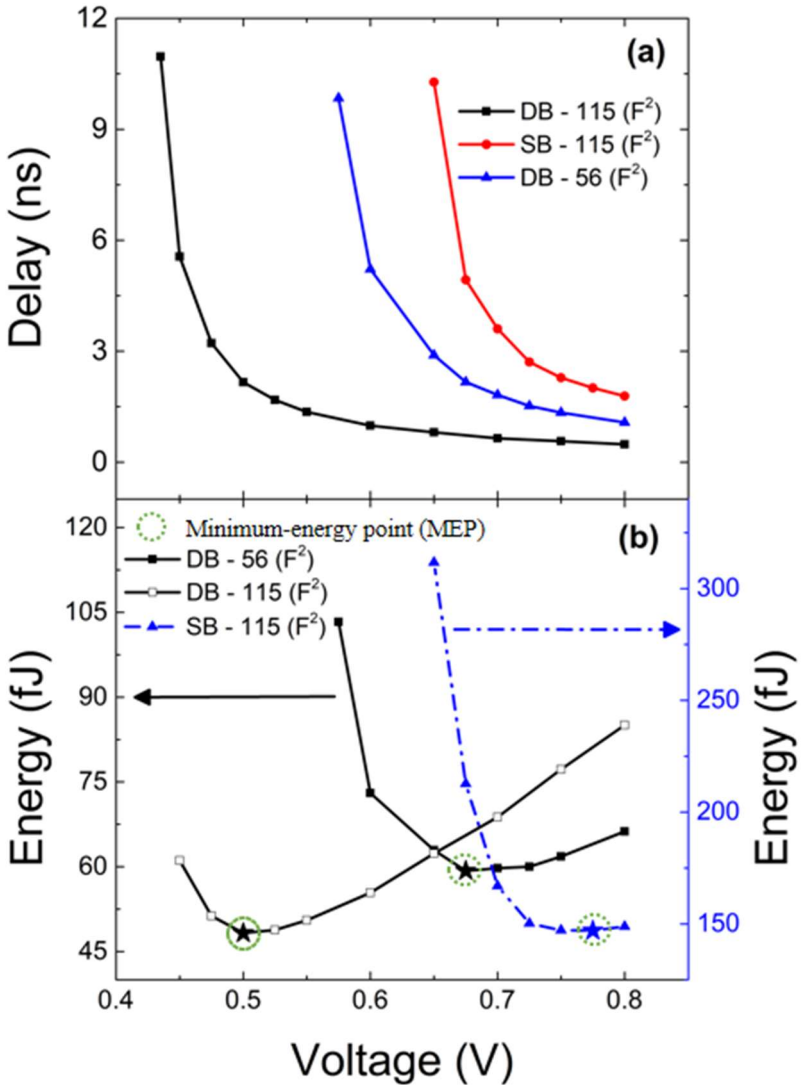

Figure 29: Monte Carlo simulation for (a) Delay vs. V_{DD} and (b) Energy vs. V_{DD} for the different cell area.

Table 7: Minimum energy points for SB and DB optimal configurations at the 28 nm technology node.

Bitcell Type	Area (F^2)	V_{MEP} (V)	Write Delay (ns) @ MEP	Write Energy (fJ) @ MEP
DB	56	0.650	2.89	62.94
DB	115	0.500	2.16	48.27
SB	115	0.775	2.01	147.03

During the writing analysis, we have covered three important aspects, the delay, the energy, and the integration capacity. As a result, we got two optimal configurations, the 2TRC for the SB and 2TSC for the DB. Now, the last aspect to consider on the STT-MRAM is its behavior in the reading operation. In the following section, we will see the robustness in reading for the V_{DD} = 0.8 V.

4.2 STT-MRAM reading analysis

- **Initial considerations**

Unlike the previous case where a transient analysis was done, the reading study is summited to a DC analysis. As a typical reading in memories, the WL is always on while the SL will be connected to ground. Then a reference current (I_{READ}) has to be sent to the bitcell as it is shown in the schematic in figure 30, which follows the brief reading methodology details mentioned in chapter 2. The current sent does not have to pass the critical current; otherwise, we will write on the bitcell. Moreover, following the general workflow mentioned in figure 24, all the results are presented with Monte Carlo simulations. In the simulation environment, in comparison with the previous one (see figure 25), the peripherals such as the buffers for each line, are neglected.

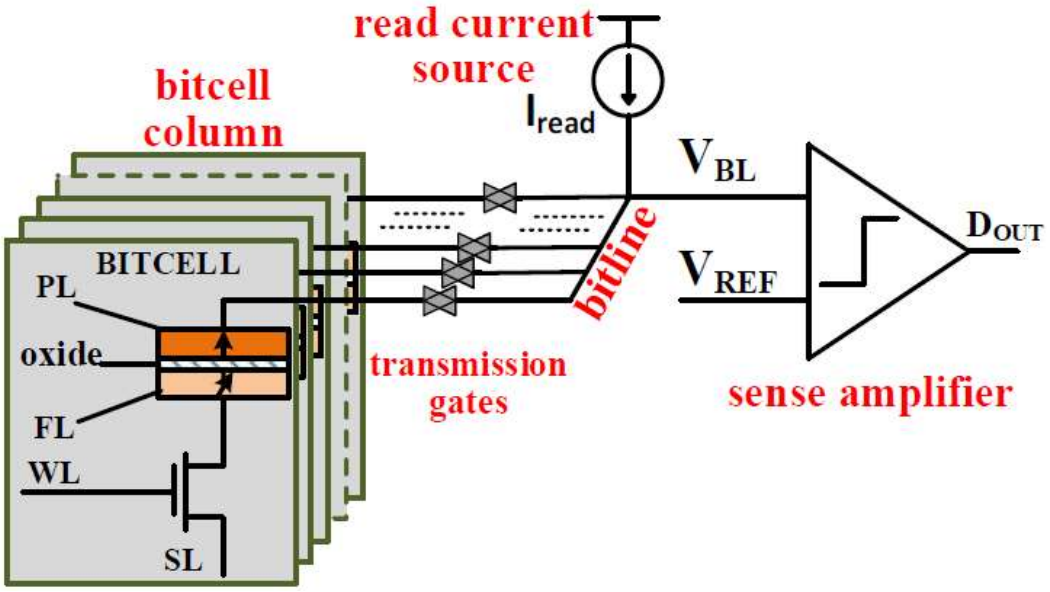

Figure 30: Typical reading simulation schematic according to [41].

- **Reading statistical analysis**

By applying the I_{READ} on the bitcell, the bitline voltage (V_{BL}) is calculated and then compared with a voltage reference (V_{REF}) to finally obtain the stored value in the bitcell; for instance, If we are in the AP (R_{AP}) state, it is read a value greater than V_{REF}, otherwise it is read the P (R_P) state. Now it is necessary to know the appropriate value of I_{READ}. This value should assure a low read disturbance probability (P_{DR}), which is defined as the probability to disturb or flip the bitcell after a reading event, and it is expressed as [42]:

$$P_{DR}(I_{read}) = 1 - \exp\left(\frac{-t_p}{\tau_0 \cdot e^{\Delta(1-I_{read}/I_{c0})}}\right) \quad (5)$$

Where τ_0 is the attempt time typically at 1 ns, Δ is the thermal stability, I_{c0} is the critical current and t_p is the duration of the read event. To ensure a low P_{DR}, the I_{READ} should be in the range of the tens of *uA* [41]. In the STT-MRAM practical design, a P_{DR} is fixed so the I_{READ} is established [41]. By looking the eq. 5, the t_p can be modified in order to get less P_{DR}. Typically, this can be used in the STT-MRAM designs but will cause a longer write pulse [42].

Two stages need to be taken. The first one is the P_{DR} analysis, which is done independently of the type of configuration. In figure 31 can be seen the P_{DR} in function of I_{READ} for SB and DB. Note the SB has two critical currents; however, the smallest one does not have to be considered. It is seen when we increase the I_{READ}, the flipping probability increases. Nevertheless, we cannot make the reading current smaller as we want because there is a trade-off with the reading sensibility.

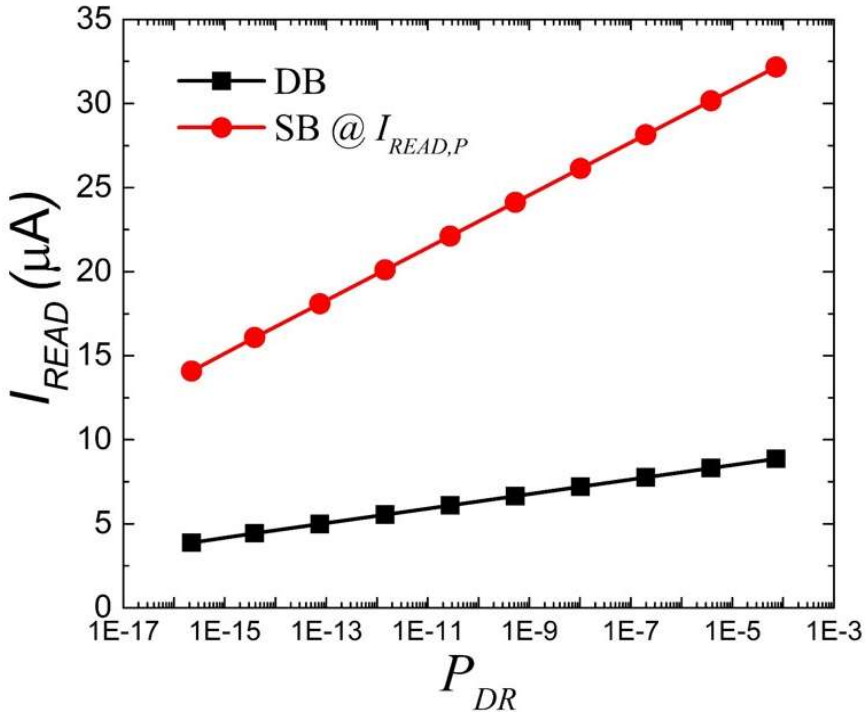

Figure 31: I_{READ} *vs.* P_{DR} for the SB and DB bitcells.

The second stage is the calculation of the $V_{SM} = V_{AP} - V_P$ by measuring the V_{BL}, which corresponds to the AP and P voltage. The I_{READ} according to figure 31 is used to send a current into the bitcell where we got the available sensing margin for different values of each I_{READ}. Including all the process variations into the bitcell, for the P and AP voltages, it is had a Gaussian distribution (see figure 32) where the V_{REF} is placed between these Gaussian distributions. Note the V_{REF} is closer to the P distribution because the AP standard deviation is greater than the P standard deviation [41].

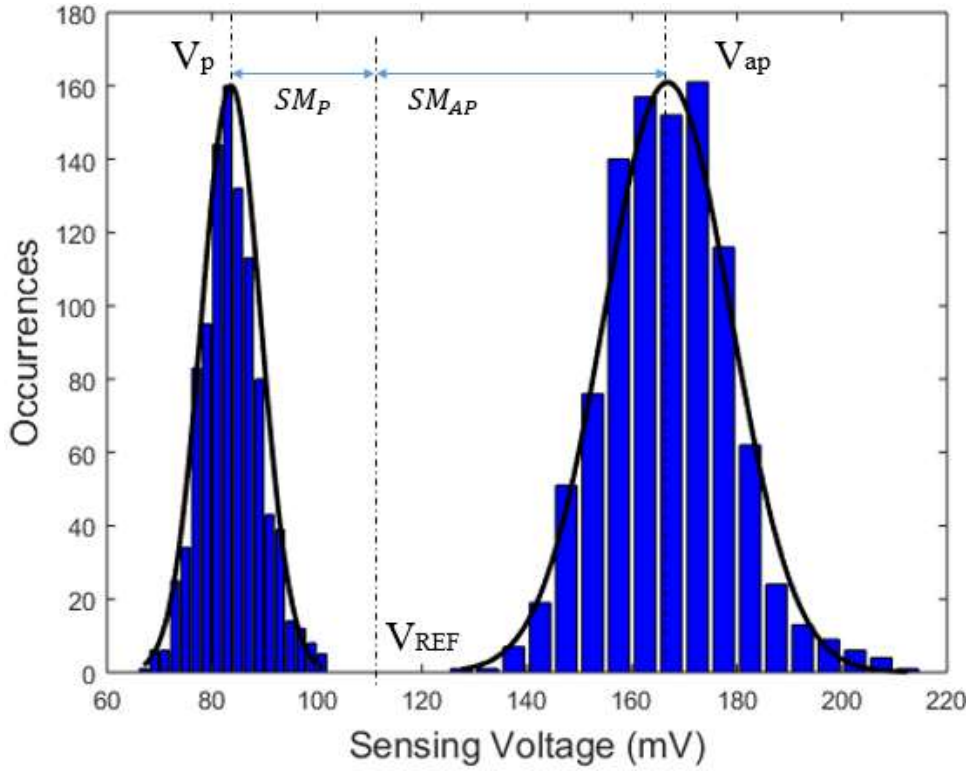

Figure 32: Example of probability distributions for the AP and P sensing voltages. It is considered a DB with area = 155 (F^2) at a P_{DR} = 2E-7 is considered.

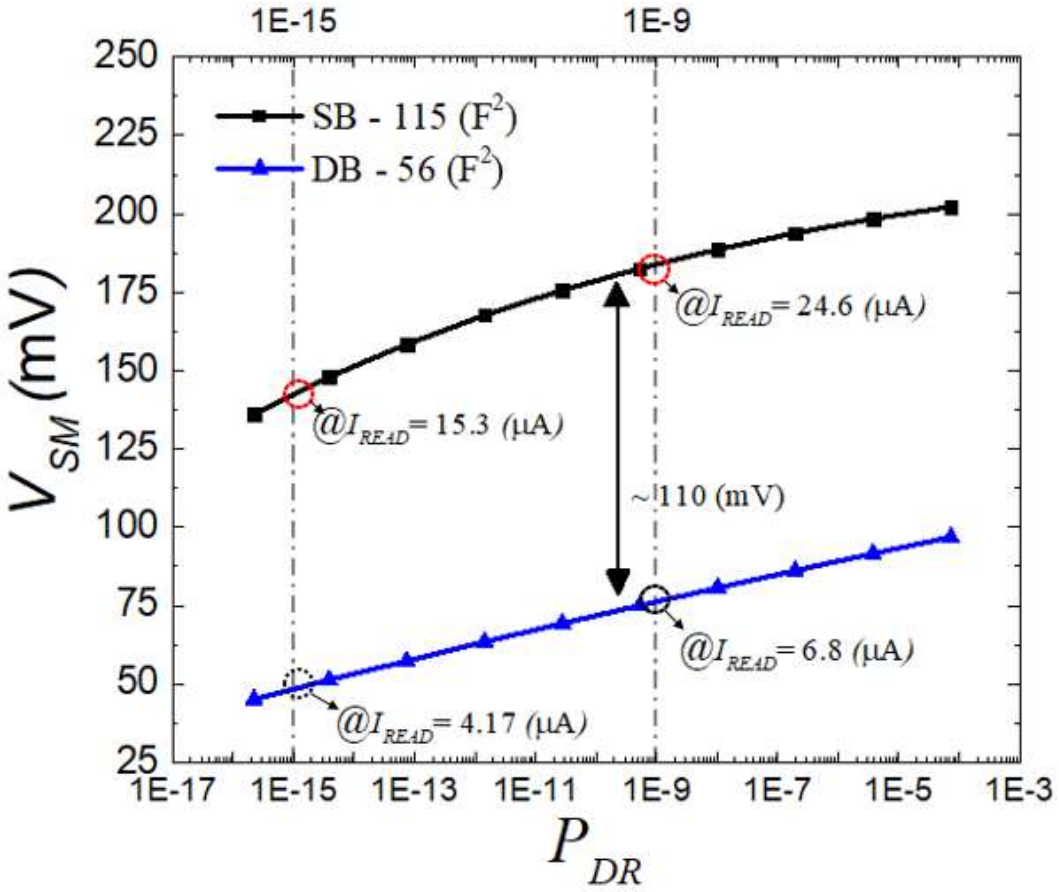

Figure 33: V_{SM} *vs.* P_{DR} considering process variations.

In figure 33 is exhibited the V_{SM} with the process variations. Now, comparing the SB and DB, for a fixed P_{DR} we have a fixed V_{SM} where the SB presents a lower flipping probability due to the higher sensing margin. Clearly, it is seen that if the P_{DR} decreases, the I_{READ} will decrease and the SM gets worse. Moreover, the DB exhibits lower read currents and in consequence, the DB will present problems in the reading sensitivity due to the limitations and complexity in reading small sensing margins.

4.3 Writing and reading analysis summary

Along the reading and writing analysis, we noticed the advantages and limitations of the STT-MRAM based on SB or DB MTJs. We started with the writing analysis where we have taken a set of 4 configurations, which for each one the area size is varied by modifying the number of fingers. With the variation of the cell variation, we have shown the behavior of the delay and energy in terms of writing. All the exhibited results gave us the STT-MRAM writing behavior as the area size was modified. At the end of the writing analysis, we have chosen the optimal configurations that are 2T1MTJ-RC (Single Barrier) and 2T1MTJ-SC (Double Barrier). For the SB, the RC is the best because the source degradation is presented in the AP – P transition where the critical current is smaller than the P – AP case. Hence, the current extracted (in AP - P) is higher enough to avoid a higher switching time (at a WER of 1×10^{-6}) that will cause an increase in the average energy consumption. On the contrary, for the DB, the SC is the best due to the source degradation presented in the P – AP transition where the resistance is smaller, leading a smaller energy consumption. Another important observation is the fact that the DB starts with 4 to 5 times the critical current, so in terms of area parity, we can use a lower V_{DD}, and in terms of V_{DD} parity, we can use smaller area in the DB. Finally, with those two optimal configurations, the reading analysis is done. In this case, the P_{DR} is essential to set a disturbance probability so later with the corresponding I_{READ} (where in the

case of the SB the high current case is chosen), the V_{SM} is obtained by applying the pertinent variations on the MTJ and access transistors.

In terms of writing, the DB has only advantages over the SB such as speed and energy. However, through the reading analysis, we have learned that DB is the worst due to a higher disturbance probability. Another reason is the small critical current, which in consequence the available sensing margin will be small, causing reading sensing problems. To overcome this limitation in the DB, we can increase the DB critical current but this negatively affects in the writing. However, the reading problems are generally managed with the reading circuits and methodologies.

CONCLUSIONS

Due to scaling, an increase in energy and power consumption is presented in the design of the latest devices, suffering problems in leakage, reliability and variability. Moreover, it is known the memories use a significant area on the chip, which is translated to high density and in consequence high power dissipation. For this reason, the concern in power consumption has increased and the metrics in new designs are focused on power and energy. In addition, we have seen that research on MRAM shows potential designs to face the actual technology problems. This has led to significant positive results in building potential designs and devices, such as STT-MRAM, which is combined with CMOS technology to get compatibility in the chip-semiconductor production.

The basic structure used in the STT-MRAM memories is named MTJ and it can be divided into two promising designs the SB and DB MTJs. The MTJ conductance depends on its magnetization, being this the key feature of the MTJ structure. In order to characterize its behavior, the Cadence – Virtuoso® design environment was used. Unfortunately, the MTJ devices are not commercially available yet, so in consequence, there are not commercial models for simulations. For this reason, in order to get the MTJ behavior, the implementation of Verilog-A compact model in Cadence is done and used along the analysis. As we have seen, a hybrid CMOS-MTJ circuit is needed to achieve promises results in the switching time of a memory device.

The STT-MRAM behavior was divided in two analyses. The first is the characterization of the memory in terms of writing performance where was reported the behavior considering parameters such us velocity (delay), energy and area occupation. According to writing analysis, and taking into account only the SB, by varying the integration capacity we prove that it is not good in writing for small transistors or certain configurations, especially for the 1T1MTJ cases.

On the contrary, the DB always have a better writing performance due to the lower critical currents. Thus, using the area variation, the MEP is found scaling 3 to 4 times the critical current. In the end, from the SB and DB, two configurations had remained as the best in terms of energy. As it was analyzed, the 2TRC is better in the SB because of the different critical currents in the two transitions, while the dominant factor of why 2TSC is better in the DB is the different resistances. Last but not least, when the t_{ox} decreases, the current increases and the TMR is improving, which in general enhances the MTJ performance.

On the other hand, the second analysis was the read access performance where the only parameter that matters is the disturbance probability, which gives the reading robustness. A low read disturbance probability has to be fixed in order to assure the appropriate reading current. The reading current is a compromise between two factors, the fact that for a lower current, we could see an error in sensing, and the fact that a flip of the bitcell can happen. Here it was seen the flipping probability is always higher in the DB because of the lower critical currents. Thus, the critical current has to be large or the reading current has to be small to guarantee a low flipping probability. This leads to the fact the DB is bad in reading due to the small critical current. However, in general for either SB or DB, if the critical current increases, the write performance will degrade.

In summary, it was shown the advantages of the DB over the SB MTJs in the use of memories. It is concluded that in area parity, the DB gains in writing but loose in reading when compared with the SB. Furthermore, the DB will always exhibit a better performance in speed and energy in the writing behavior, but it gets worse in reading sensing due to the small currents. In the end, the lost in reading compensates a good writing performance.

REFERENCES

[1] D. Apalkov, B. Dieny and J.M. Slaugther, “Magnetoresistive Random Access Memory,” *Proceedings of the IEEE*. Vol. 104, No. 10, pp. 1796 – 1798, October 2016.

[2] S. Tehrani *et al.*, “Magnetoresistive Random Access Memory Using Magnetic Tunnel Junctions,” *Proceedings of the IEEE*, Vol. 91, No. 5, pp. 703 – 714, May 2003.

[3] Y. Makino *et al.*, “Magnetoresistance, stress effects, and a self-similar expansion model for the magnetization process in amorphous wires,” *IEEE Transactions on Magnetics*, Vol. 25, No. 5. September 1989.

[4] S.Z. Peng *et al.*, “Magnetic Tunnel Jucntion for Spintronics: Principles and Applications,” John Wiley & Sons, Inc. DOI: 10.1002/047134608X.W8231. 2014.

[5] Industry view of 1st Generation MRAM, 2017 Microchip Technology [Online] Availabe: http://www.microchip.com/memory

[6] X. Fong *et al.*, “Spin-Transfer Torque Memories: Devices, Circuits, and Systems,” in *Proceedings of the IEEE*, Vol. 104, pp.1449-1488, July 2016.

[7] S. Oh *et al.*, “Bias-voltage dependence of perpendicular spin-transfer torque in asymmetric MgO-based magnetic tunnel junctions,” *Nature Physics* 5, pp. 898-902, October 2009.

[8] A. Thomas, *et al.*, “A 4-Mb 0.18-µm 1T1MTJ Toggle MRAM With Balanced Three Input Sensing Scheme and Locally Mirrored Unidirectional Write Drivers,” *IEEE Journal of Solid-State Circuits*, Vol. 40, No. 1, pp. 301-303, January 2005.

[9] N.D. Rizzo, *et al.*, “A Fully Functional 64 Mb DDR3 ST-MRAM Built on 90nm CMOS Technology,” *IEEE Transactions on Magnetics*, vol. 49, no.7, pp. 4441, July 2013.

[10] L. Loong *et al.,* “Strain-enhanced tunneling magnetoresistance in MgO magnetic tunnel junctions,” *Institute of Materials Research and Engineering – Scientific Reports.* DOI:10.1038/srep06505. September 2014.

[11] R. De Rose *et al*., “Variability-Aware Analysis of Hybrid MTJ/CMOS Circuits by a Micromagnetic-Based Simulation Framework,” *IEEE Transactions on Nanotechnology,* Vol. 16, no. 2, pp. 160-167, March 2017.

[12] R. De Rose *et al*., “A Variation-Aware Timing Modeling Approach for Write Operation in Hybrid CMOS/STT-MTJ Circuits,” *IEEE Transactions on Circuits and Systems-I: Regular Papers,* Vol. 65, no. 3, pp. 1086-1093, March 2018.

[13] H. Sato *et al.*, “Perpendicular-anisotropy CoFeB-MgO magnetic tunnel junctions with a MgO/CoFeB/Ta/CoFeB/MgO recording structure,” *Applied Physics Letters*, Vol. 101, 022414, July 2012.

[14] G. Hu *et al.*, “STT-MRAM with double magnetic tunnel junctions*,” IEEE International Electron Devices Meeting*, No. 15, pp. 668-671. 2015.

[15] G. Wang *et al*., "Compact Modeling of High Spin Transfer Torque Efficiency Double-Barrier Magnetic Tunnel Junction*," IEEE/ACM International Symposium on Nanoscale Architectures*, No. 17, pp. 49-54. July 2017.

[16] G. Jan *et al.,* "High Spin Torque Efficiency of Magnetic Tunnel Junctions with MgO/CoFeB/MgO Free Layer," *Applied Physics Express,* No.5, 093008, pp. 1-3, September 2012.

[17] M. Carpentieri *et al.,* "Micromagnetic Analysis of Statistical Switching in Perpendicular Magnetic Tunnel Junctions with Double Reference Layers," *IEEE Magnetics Letters,* Vol. 7, June 2016.

[18] T. Endoh *et al.,* "An Overview of Nonvolatile Emerging Memories – Spintronics for Working Memories," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, Vol. 6, No. 2, June 2016.

[19] K. Gavaskarl & U.S. Ragupathy, "An efficient Design and Comparative Analysis of Low Power Memory Cell Structures," *Green Computing Communication and Electrical Engineering (ICGCCEE) – International Conference*, DOI: 10.1109/ICGCCEE.2014.6922280, March 2014.

[20] T. Endoh *et al.,* "Restructuring of Memory Hierrarchy in Computing System with Spintronics-Based Technologies," *Symposium on VLSI Technology Digest of Technical Papers*, pp. 89-90. June 2012.

[21] Yole Développement, "Emerging Non-Volatile Memory 2017 – Market & Technology report," June 2017.

[22] Yole Développement, "Is the emerging non-volatile memory (NVM) market ready for take-off?," Press Report, June 2017.

[23] Yole Développement, "Storage-class memory will be the clear go-to market for emerging non-volatile memory in 2021," Non-Volatile Memory Report, July 2016.

[24] "MRAM History," *https://www.mram-info.com/history,* August 2018.

[25] L. Chang *et al.,* "A brief introduction to Giant Magnetoresistance," *Ohio State University*, Columbus, OH 43210.

[26] L. Torres & N. Bruchon. "On the Use of Magnetic RAMs in Field-Programmable Gate Arrays," *International Journal of Reconfigurable Computing,* DOI:10.1155/2008/723950, November 2008.

[27] I.K. Yanson *et al.,* "Current-field diagram of magnetic states of a surface spin valve in a point contact with a single ferromagnetic film," *Low Temperature Physics*, Vol. 39, No. 3, 2013.

[28] A. Giordano *et al.,* "Semi-implicit integration scheme for Landau-Lifshitz-Gilbert-Slonczewski," *Journal of Applied Phyics*, Vol. 111, 2012.

[29] R. De Rose *et al*., "A Compact Model with Spin-Polarization Asymmetry for Nanoscaled Perpendicular MTJs," *IEEE Transactions on Electron Devices,* Vol. 64, No. 10, pp. 4346-4353, 2017.

[30] H. Mulaosmanovic *et al,* "Working Principles of a DRAM Cell Based on Gated-Thyristor Bistability," *IEEE Electron Device Letters,* Vol. 35, No. 9, pp. 921-923, 2014.

[31] S. Okhonin *et al,* "A Capacitor-Less 1T-DRAM Cell," *IEEE Electron Device Letters,* Vol. 23, No. 2, 2002.

[32] Yole Développement, "Emerging Non-Volatile Memory 2016 – From Technologies to Market report," July 2017.

[33] S. Peng *et al,* "Origin of interfacial perpendicular magnetic anisotropy in MgO/CoFe/ metallic capping layer structures," *Nature Scientific Reports,* December 2015.

[34] T. Liu *et al,* "Thermally robust Mo/CoFeB/MgO trilayers with strong perpendicular anisotropy," *Nature Scientific Reports,* July 2014.

[35] L. Thomas, "Basic Principles, Challenges and Opportunities of STT-MRAM for Embedded Memory Applications," TDK – Headway Technologies, USA, May 2017.

[36] S. Ikeda, "Perpendicular-anisotropy CoFeB-MgO based magnetic tunnel junctions scaling down to 1X nm," *IEEE International Electron Devices Meeting,* 2014.

[37] J. Li, B. Luan and C. Lam, "Resistance Drift in Phase Change Memory," *IEEE International Reliability Physics Symposium (IRPS),* 2012.

[38] R. De Rose *et al*., "Impact of voltage scaling on STT-MRAMs through a variability-aware simulation framework*," International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD),* July 2017.

[39] A. Sarkar *et al*., "Low Power VLSI Design: Fundamentals," Walter de Gruyter GmbH & Co KG, 2016.

[40] IMEC, "Imec demonstrates manufacturability o state-of-the-art spin-orbit torque MRAM devices on 300mm Si Wafers," *Press releases*, Leuven, June 2018.

[41] T. Quang Kien, "STT-MRAMS Circuit Techniques for Enhanced Robustness in Low Power Embedded Applications," Department of Electrical and Computer Engineering, National University of Singapore, 2017.

[42] D. Apalkov *et al.,* "Spin-Transfer Torque Magnetic Random Access Memory (STT-MRAM)," *ACM Journal on Emerging Technologies in Computing Systems,* Vol. 9, No. 2, Article 13, May 2013.

[43] R. De Rose *et al*., "A Compact Model for Perpendicular Spin-Transfer-Torque Magnetic Tunnel Junctions with Double Reference Layers," *IEEE,* Waiting for Publication.

[44] Y. Zhang *et al*., "Compact Model of Subvolume MTJ and Its Design Application at Nanoscale Technology Nodes," *IEEE Transactions on Electron Devices,* Vol.62, No. 6, June 2015.

[45] K. Jeong & A. Kahng, "A Power-Constrained MPU Roadmap for International Technology Roadmap for Semiconductors," *UCSB VLSI CAD Laboratory - ISOCC*, University of California, San Diego, November 2009.

[46] A. Chintaluri, “Analysis of Defects and Fault Models in Embedded Spin-Transfer Torque (STT) MRAM Arrays,” *Georgia Institute of Technology*, Georgia, May 2016.

ANNEXES INDEX

ANNEX A: 28 NM MATLAB SCRIPT FOR SWITCHING DELAY OF P-AP AND AP-P TRANSITIONS

MATLAB script – 28nm DB switching delay for P-AP and AP-P transitions

Define physical constants

```
e = 1.6e-19;            % elementary charge [C]
mub = 9.274e-24;        % Bohr magneton constant [J*T^-1]
kb = 1.3806488e-23;     % Boltzmann constant [J*K^-1]
mu0 = 4*pi*1e-7;        % vacuum permeability [H/m]
```

Define technology and device parameters

```
alpha = 0.05;           % Gilbert damping coefficient
gamma = 1.76*1e11;      % gyromagnetic constant [Hz/T]
Ms = 1e6;               % saturation magnetization in the free layer [A/m]
ku = 0.88e6;            % interfacial perpendicular anisotropy [J*m^-3]
tfl = 1.2e-9;           % thickness of the free layer [m]
r = 14e-9;              % MTJ radius of the surface [m]
T = 300;                % room temperature [K]
nu = 0.67;              % spin polarization factor
Nperp = 0.0423558;
Nz = 0.9152884;
cp = -nu^4;             % parameter which controls the asimmetry of the spin-torque
haz = 0;                % external field
g = 2;                  % Lande' factor
```

Initial calculations

```
surface = pi*r^2;                                   % MTJ surface [m^2]
Vfl = surface*tfl;                                  % volume of the free layer [m^3]
keff = Nperp + (2*ku/(mu0*Ms^2)) - Nz;              % effective anistoropy
Hk_eff = (Nperp-Nz)*Ms+(2*ku/(mu0*Ms));             % effective anisotropy field [A/m]
ku_eff = (mu0*Ms^2*(Nperp-Nz)/2)+ku;                % [J/m^3]
E = mu0*Ms*Hk_eff*Vfl/2;                            % [J]
delta = E/(kb*T)                                    % thermal stability
```

Switching behavior

```
betacrit = alpha*(1+cp)*(keff+haz);                         % normalized critical
current
Ic0 = betacrit*(e*gamma*mu0*Ms^2*Vfl/(mub*4*nu*g))          % calculation of the
critical current
Jc0 = Ic0/(surface*1e4)

mu = (mu0*Ms^2*Vfl)/(kb*T);                                 % parameter defined by
D'Aquino
nPts = 1e3;                                                 % number of points
considered
theta = linspace(0+0.001,+pi/3,nPts);                       % tilting angle with
respect to z-axis (varies between 0 and pi/3)
peq = mu*keff*theta.*exp(-mu*(keff/2)*(theta.^2));          % PDF of tilting angle
```

```
figure(1)                                                    % plot PDF of tilting
angle
plot(theta,peq)
title('PDF of theta')
xlabel('theta')

mz0 = cos(theta);                                            % initial state of the
magnetization
mzf = -0.9;                                                  % final state of the
magnetization for P->AP switching
%Jmtj = 3e6
%Imtj = Jmtj*(surface*1e4)
Imtj = 65.1e-6;
beta = Imtj*(mub*4*nu*g)/(e*gamma*mu0*Ms^2*Vfl);             % normalized bias
current
nF = beta/betacrit                                           % ratio between the
normalized injected bias current and the normalized critical current

% Define parameters for ts formula coming from the resolution of the integral
h = -beta/(alpha*(1+cp))

ts = (1/alpha)*[(1/(2*(h-keff)))*log((1+mzf)./(1+mz0)) - (1/(2*(h+keff)))*log((1-
mzf)./(1-mz0)) - (keff/(h^2-keff^2))*log((h+keff*mzf)./(h+keff*mz0))];

% Plot ts as a function of mz0
figure(2)
plot(mz0,ts)
title('ts')
xlabel('mz0=cos(theta)')

% Perform a simple numerical inversion of ts to obtain the element g^-1(ts) used
for the computation of the switching time PDF
ts2 = linspace(min(ts),max(ts),nPts);
%ts2 = fliplr(ts);
mz2 = interp1(ts,mz0,ts2,'pchirp');
%mz2 = fliplr(mz0);

% Plot mz2 as a function of ts2 (should be equal to the numerical inversion of ts)
figure(3)
plot(ts2,mz2)
title('mz2')
xlabel('ts2')

tsunnorm = ts2/(gamma*mu0*Ms);      % denormalization of ts [s]

% Calculate the switching time PDF from D'Aquino formulation
tsPDF_daq = mu*keff.*(exp(-mu*(keff/2).*(acos(mz2)).^2)).*abs(alpha.*(keff*mz2 +
h).*(1 - (mz2).^2));  % Proposed by D'AQUINO (corrected with respect to Giulio's
code)
tsPDF_daq = tsPDF_daq/trapz(tsunnorm,tsPDF_daq);

% Calculate numerically the switching time CDF from D'Aquino PDF
tsCDF_num = cumsum(tsPDF_daq);
tsCDF_num = tsCDF_num/sum(tsCDF_num);
tsCDF_num = tsCDF_num/max(tsCDF_num);
```

```
% Calculate the switching time CDF from D'Aquino formulation
tsCDF_daq = exp(-mu*(keff/2).*(acos(mz2)).^2);

% Calculate numerically the switching time PDF from D'Aquino CDF
tsPDF_num = diff(tsCDF_daq)./diff(tsunnorm);
new_tsunnorm = tsunnorm(1:end-1)+diff(tsunnorm)./2;

% Plot numerical vs. D'Aquino PDF
figure(4)
plot(tsunnorm*1e9,tsPDF_daq,'r');
hold on;
plot(new_tsunnorm*1e9,tsPDF_num,'b');
xlabel('Time [ns]');
legend('DAQUINO PDF','Numerical PDF');

% Plot numerical vs. D'Aquino CDF
figure(5)
plot(tsunnorm*1e9,tsCDF_daq,'r');
hold on;
plot(tsunnorm*1e9,tsCDF_num,'b');
xlabel('Time [ns]');
legend('DAQUINO CDF','Numerical CDF');

% Generate random samples from CDF through inverse sampling method
num = 1e7;
rng(0)
rnd = rand(num,1);
r_ts = interp1(tsCDF_daq,tsunnorm*1e9,rnd,'linear',0);

% Plot histogram
figure(6)
hist(r_ts,100);
title('Histogram of the switching time')
xlabel('Time [ns]');

% Calculate the moments of the histogram
mean_ts = (sum(r_ts)/num)*1e-9
std_ts = std(r_ts)*1e-9
skew_ts = skewness(r_ts)
kurt_ts = kurtosis(r_ts)

rng(0)
r_pears = pearsrnd(mean_ts,std_ts,skew_ts,kurt_ts,num,1);
r_pears = sort(r_pears);
figure(7)
hist(r_pears,100)
title('Histogram of the switching time (Pearson)')
xlabel('Time [ns]');
mean_pears = sum(r_pears)/length(r_pears)
sigma_pears = std(r_pears)
skew_pears = skewness(r_pears)
kurt_pears = kurtosis(r_pears)
[f_pears,y_pears]=ecdf(r_pears);
figure(8)
```

```
plot(tsunnorm*1e9,tsCDF_daq,'r');
hold on
plot(y_pears*1e9,f_pears,'b')
xlabel('Time [ns]');
legend('DAQUINO CDF','PEARSON')
WER = 1-f_pears;
figure(9)
loglog(y_pears*1e9,WER)
xlabel('Time [ns]');

ts_05 = interp1(tsCDF_daq,tsunnorm*1e9,0.5,'linear',0)
WER_target = 1e-6;
ts_WER = interp1(WER,y_pears*1e9,WER_target,'linear',0)
```

MATLAB script – 28 nm SB Switching delay for the AP-P transition

Define physical constants

```
e = 1.6e-19;              % elementary charge [C]
mub = 9.274e-24;          % Bohr magneton constant [J*T^-1]
kb = 1.3806488e-23;       % Boltzmann constant [J*K^-1]
mu0 = 4*pi*1e-7;          % vacuum permeability [H/m]
```

Define technology and device parameters

```
alpha = 0.05;           % Gilbert damping coefficient
gamma = 1.76*1e11;       % gyromagnetic constant [Hz/T]
Ms = 1e6;                % saturation magnetization in the free layer [A/m]
ku = 0.88e6;             % interfacial perpendicular anisotropy [J*m^-3]
tfl = 1.2e-9;            % thickness of the free layer [m]
r = 14e-9;               % MTJ radius of the surface [m]
T = 300;                 % room temperature [K]
nu = 0.67;               % spin polarization factor
Nperp = 0.0423558;
Nz = 0.9152884;
cp = nu^2;               % parameter which controls the asimmetry of the spin-torque
haz = 0;                 % external field
g = 2;                   % Lande' factor
```

Initial calculations

```
surface = pi*r^2;                                % MTJ surface [m^2]
Vfl = surface*tfl;                               % volume of the free layer [m^3]
keff = Nperp + (2*ku/(mu0*Ms^2)) - Nz;           % effective anistoropy
Hk_eff = (Nperp-Nz)*Ms+(2*ku/(mu0*Ms));          % effective anisotropy field [A/m]
ku_eff = (mu0*Ms^2*(Nperp-Nz)/2)+ku;             % [J/m^3]
E = mu0*Ms*Hk_eff*Vfl/2;                         % [J]
delta = E/(kb*T)                                 % thermal stability
```

P->AP switching behavior

```
betacrit_p = alpha*(1+cp)*(keff+haz);                          % normalized critical
current for P->AP transition
betacrit_ap = alpha*(1-cp)*(-keff+haz);                        % normalized critical
current for AP->P transition
Ic0p = betacrit_p*(e*gamma*mu0*Ms^2*Vfl/(mub*2*nu*g));      % calculation of the
critical current for P->AP transition (denormalization of betacrit) [A]
Jc0p = Ic0p/(surface*1e4)                                      % critical current
density for P->AP transition [A/cm^2]
Ic0ap = betacrit_ap*(e*gamma*mu0*Ms^2*Vfl/(mub*2*nu*g));    % calculation of the
critical current for AP->P transition (denormalization of betacrit) [A]
Jc0ap = Ic0ap/(surface*1e4);                                   % critical current
density for AP->P transition [A/cm^2]

mu = (mu0*Ms^2*Vfl)/(kb*T);                                    % parameter defined by
D'Aquino
nPts = 1e3;                                                    % number of points
considered
theta = linspace(pi-0.001,pi-pi/3,nPts);                       % tilting angle with
respect to z-axis (varies between 0 and pi/3)
peq = mu*keff*(pi-theta).*exp(-mu*(keff/2)*((pi-theta).^2));          % PDF of
tilting angle
% figure(1)                                                      % plot PDF of tilting
angle
plot(theta,peq)
title('PDF of theta')
xlabel('theta')

mz0 = cos(theta);                                              % initial state of the
magnetization
mzf = 0.9;                                                    % final state of the
magnetization for P->AP switching
%Jmtj = 3e6
%Imtj = Jmtj*(surface*1e4)
Imtj = 48.25e-6;
beta_ap = -Imtj*(mub*2*nu*g)/(e*gamma*mu0*Ms^2*Vfl);          % normalized bias
current
nF = beta_ap/betacrit_ap                                       % ratio between the
normalized injected bias current and the normalized critical current

% Define parameters for ts formula coming from the resolution of the integral
a = keff;
b = haz;
c = beta_ap/alpha;
d = cp;

epsi    = beta_ap/alpha;
C1      = sqrt(-4*epsi*cp*keff + 2*haz*cp*keff - (haz*cp)^2 - keff^2);
C2      = atan((keff + 2*keff*cp*mz0 + haz*cp)/C1);
C3      = atan((keff + 2*keff*cp*mzf + haz*cp)/C1);
C4      = log(abs(keff*mzf + keff*(mzf.^2)*cp + haz + haz*cp*mzf - epsi));
C5      = log(abs(keff*mz0 + keff*(mz0.^2)*cp + haz + haz*cp*mz0 - epsi));
```

```
ts = (-1/(2*alpha))*[keff*C4*C1 - keff*C5*C1 - log(1+mzf)*C1*haz +
log(1+mzf)*C1*epsi+...                    % row #1
        -log(abs(mz0-1))*C1*haz + 2*C3*keff^2 + 2*C3*epsi*cp^2*haz +
2*C3*haz*cp^3*keff+...                  % row #2
        +6*C3*epsi*cp*keff - 2*C3*haz*cp*keff - 2*C3*keff^2*cp^2 -
log(1+mzf)*C1*keff+...                    % row #3
        -2*C2*haz*cp^3*keff + 2*C2*haz*cp*keff - 6*C2*epsi*cp*keff -
2*C2*epsi*cp^2*haz - 2*C2*keff^2+...    % row #4
        +2*C2*keff^2*cp^2 + log(abs(mz0-1))*cp^2*C1*haz - log(abs(mz0-
1))*cp^2*C1*keff+...                    % row #5
        +log(abs(mz0-1))*cp*C1*epsi + log(1+mzf)*cp^2*C1*keff +
log(1+mzf)*cp^2*C1*haz+...                    % row #6
        -log(1+mzf)*cp*C1*epsi - log(1+mz0)*cp^2*C1*keff -
log(1+mz0)*cp^2*C1*haz+...                          % row #7
        +log(1+mz0)*cp*C1*epsi - log(abs(mzf-1))*cp^2*C1*haz + log(abs(mzf-
1))*cp^2*C1*keff+...              % row #8
        -log(abs(mzf-1))*cp*C1*epsi + keff*cp^2*C5*C1 - cp*C5*epsi*C1 -
keff*cp^2*C4*C1 + cp*C4*epsi*C1+... % row #9
        log(1+mz0)*C1*haz - log(1+mz0)*C1*epsi + log(1+mz0)*C1*keff+...
% row #10
        log(abs(mzf-1))*C1*haz - log(abs(mzf-1))*C1*keff - log(abs(mzf-
1))*C1*epsi+...                       % row #11
        log(abs(mz0-1))*C1*epsi + log(abs(mz0-1))*C1*keff]/...
% row #12
        [(haz + keff*cp - epsi + haz*cp + keff)*(haz - keff - epsi - haz*cp +
keff*cp)*C1];                 % row #13

% Plot ts as a function of mz0
figure(2)
plot(mz0,ts)
title('ts')
xlabel('mz0=cos(theta)')

% Perform a simple numerical inversion of ts to obtain the element g^-1(ts) used
for the computation of the switching time PDF
ts2 = linspace(min(ts),max(ts),nPts);
%ts2 = fliplr(ts);
mz2 = interp1(ts,mz0,ts2,'pchirp');
%mz2 = fliplr(mz0);

% Plot mz2 as a function of ts2 (should be equal to the numerical inversion of ts)
figure(3)
plot(ts2,mz2)
title('mz2')
xlabel('ts2')

tsunnorm = ts2/(gamma*mu0*Ms);      % denormalization of ts [s]

% Calculate the switching time PDF from D'Aquino formulation
tsPDF_daq = mu*keff.*(exp(-mu*(keff/2).*(pi-acos(mz2)).^2)).*abs(alpha.*(keff*mz2 +
haz - ((beta_ap/alpha).*((1 + cp*mz2).^-1))).*(1 - (mz2).^2));   % Proposed by
D'AQUINO (corrected with respect to Giulio's code)
tsPDF_daq = tsPDF_daq/trapz(tsunnorm,tsPDF_daq);

% Calculate numerically the switching time CDF from D'Aquino PDF
```

```
tsCDF_num = cumsum(tsPDF_daq);
tsCDF_num = tsCDF_num/sum(tsCDF_num);
tsCDF_num = tsCDF_num/max(tsCDF_num);

% Calculate the switching time CDF from D'Aquino formulation
tsCDF_daq = exp(mu*(-keff/2)*((pi-acos(mz2)).^2));

% Calculate numerically the switching time PDF from D'Aquino CDF
tsPDF_num = diff(tsCDF_daq)./diff(tsunnorm);
new_tsunnorm = tsunnorm(1:end-1)+diff(tsunnorm)./2;

% Plot numerical vs. D'Aquino PDF
figure(4)
plot(tsunnorm*1e9,tsPDF_daq,'r');
hold on;
plot(new_tsunnorm*1e9,tsPDF_num,'b');
xlabel('Time [ns]');
legend('DAQUINO PDF','Numerical PDF');

% Plot numerical vs. D'Aquino CDF
figure(5)
plot(tsunnorm*1e9,tsCDF_daq,'r');
hold on;
plot(tsunnorm*1e9,tsCDF_num,'b');
xlabel('Time [ns]');
legend('DAQUINO CDF','Numerical CDF');

% Generate random samples from CDF through inverse sampling method
num = 1e7;
rng(0)
rnd = rand(num,1);
r_ts = interp1(tsCDF_daq,tsunnorm*1e9,rnd,'linear',0);

% Plot histogram
figure(6)
hist(r_ts,100);
title('Histogram of the switching time')
xlabel('Time [ns]');

% Calculate the moments of the histogram
mean_ts = (sum(r_ts)/num)*1e-9
std_ts = std(r_ts)*1e-9
skew_ts = skewness(r_ts)
kurt_ts = kurtosis(r_ts)

rng(0)
r_pears = pearsrnd(mean_ts,std_ts,skew_ts,kurt_ts,num,1);
r_pears = sort(r_pears);
figure(7)
hist(r_pears,100)
title('Histogram of the switching time (Pearson)')
xlabel('Time [ns]');
mean_pears = sum(r_pears)/length(r_pears)
sigma_pears = std(r_pears)
skew_pears = skewness(r_pears)
```

```
kurt_pears = kurtosis(r_pears)
[f_pears,y_pears]=ecdf(r_pears);
figure(8)
plot(tsunnorm*1e9,tsCDF_daq,'r');
hold on
plot(y_pears*1e9,f_pears,'b')
xlabel('Time [ns]');
legend('DAQUINO CDF','PEARSON')
WER = 1-f_pears;
figure(9)
loglog(y_pears*1e9,WER)
xlabel('Time [ns]');

ts_05 = interp1(tsCDF_daq,tsunnorm*1e9,0.5,'linear',0)
WER_target = 1e-6;
ts_WER = interp1(WER,y_pears*1e9,WER_target,'linear',0)
```

MATLAB script – 28 nm SB switching delay for the P-AP transition

Define physical constants

```
e = 1.6e-19;            % elementary charge [C]
mub = 9.274e-24;        % Bohr magneton constant [J*T^-1]
kb = 1.3806488e-23;     % Boltzmann constant [J*K^-1]
mu0 = 4*pi*1e-7;        % vacuum permeability [H/m]
```

Define technology and device parameters

```
alpha = 0.05;          % Gilbert damping coefficient
gamma = 1.76*1e11;      % gyromagnetic constant [Hz/T]
Ms = 1e6;               % saturation magnetization in the free layer [A/m]
ku = 0.88e6;            % interfacial perpendicular anisotropy [J*m^-3]
tfl = 1.2e-9;           % thickness of the free layer [m]
r = 14e-9;              % MTJ radius of the surface [m]
T = 300;                % room temperature [K]
nu = 0.67;              % spin polarization factor
Nperp = 0.0423558;
Nz = 0.9152884;
cp = nu^2;              % parameter which controls the asimmetry of the spin-torque
haz = 0;                % external field
g = 2;                  % Lande' factor
```

Initial calculations

```
surface = pi*r^2;                                % MTJ surface [m^2]
Vfl = surface*tfl;                               % volume of the free layer [m^3]
keff = Nperp + (2*ku/(mu0*Ms^2)) - Nz;           % effective anistoropy
Hk_eff = (Nperp-Nz)*Ms+(2*ku/(mu0*Ms));          % effective anisotropy field [A/m]
ku_eff = (mu0*Ms^2*(Nperp-Nz)/2)+ku;             % [J/m^3]
E = mu0*Ms*Hk_eff*Vfl/2;                         % [J]
delta = E/(kb*T)                                 % thermal stability
```

P->AP switching behavior

```
betacrit_p = alpha*(1+cp)*(keff+haz);                          % normalized critical
current for P->AP transition
betacrit_ap = alpha*(1-cp)*(-keff+haz);                        % normalized critical
current for AP->P transition
Ic0p = betacrit_p*(e*gamma*mu0*Ms^2*Vfl/(mub*2*nu*g));         % calculation of the
critical current for P->AP transition (denormalization of betacrit) [A]
Jc0p = Ic0p/(surface*1e4)                                      % critical current
density for P->AP transition [A/cm^2]
Ic0ap = betacrit_ap*(e*gamma*mu0*Ms^2*Vfl/(mub*2*nu*g));       % calculation of the
critical current for AP->P transition (denormalization of betacrit) [A]
Jc0ap = Ic0ap/(surface*1e4);                                   % critical current
density for AP->P transition [A/cm^2]

mu = (mu0*Ms^2*Vfl)/(kb*T);                                    % parameter defined by
D'Aquino
nPts = 1e3;                                                    % number of points
considered
theta = linspace(0+0.001,+pi/3,nPts);                          % tilting angle with
respect to z-axis (varies between 0 and pi/3)
peq = mu*keff*theta.*exp(-mu*(keff/2)*(theta.^2));             % PDF of tilting angle
figure(1)                                                      % plot PDF of tilting
angle
plot(theta,peq)
title('PDF of theta')
xlabel('theta')

mz0 = cos(theta);                                              % initial state of the
magnetization
mzf = -0.9;                                                    % final state of the
magnetization for P->AP switching
%Jmtj = 10e6
%Imtj = Jmtj*(surface*1e4)
Imtj = 74,27e-6;
beta_p = Imtj*(mub*2*nu*g)/(e*gamma*mu0*Ms^2*Vfl);             % normalized bias
current
nF = beta_p/betacrit_p                                         % ratio between the
normalized injected bias current and the normalized critical current

% Define parameters for ts formula coming from the resolution of the integral
a = keff;
b = haz;
c = beta_p/alpha;
d = cp;

epsi    = beta_p/alpha;
C1      = sqrt(4*epsi*cp*keff - 2*haz*cp*keff + (haz*cp)^2 + keff^2);
C2      = atanh((keff + 2*keff*cp*mz0 + haz*cp)/C1);
C3      = atanh((keff + 2*keff*cp*mzf + haz*cp)/C1);
C4      = log(abs(keff*mzf + keff*(mzf.^2)*cp + haz + haz*cp*mzf - epsi));
C5      = log(abs(keff*mz0 + keff*(mz0.^2)*cp + haz + haz*cp*mz0 - epsi));


ts = (-1/(2*alpha))*[keff*C4*C1 - keff*C5*C1 - log(1+mzf)*C1*haz +
log(1+mzf)*C1*epsi+...                    % row #1
```

```
        -log(abs(mz0-1))*C1*haz - 2*C3*keff^2 - 2*C3*epsi*cp^2*haz -
2*C3*haz*cp^3*keff+...                  % row #2
        -6*C3*epsi*cp*keff + 2*C3*haz*cp*keff + 2*C3*keff^2*cp^2 -
log(1+mzf)*C1*keff+...                    % row #3
        +2*C2*haz*cp^3*keff - 2*C2*haz*cp*keff + 6*C2*epsi*cp*keff +
2*C2*epsi*cp^2*haz + 2*C2*keff^2+...    % row #4
        -2*C2*keff^2*cp^2 + log(abs(mz0-1))*cp^2*C1*haz - log(abs(mz0-
1))*cp^2*C1*keff+...                   % row #5
        +log(abs(mz0-1))*cp*C1*epsi + log(1+mzf)*cp^2*C1*keff +
log(1+mzf)*cp^2*C1*haz+...                    % row #6
        -log(1+mzf)*cp*C1*epsi - log(1+mz0)*cp^2*C1*keff -
log(1+mz0)*cp^2*C1*haz+...                          % row #7
        +log(1+mz0)*cp*C1*epsi - log(abs(mzf-1))*cp^2*C1*haz + log(abs(mzf-
1))*cp^2*C1*keff+...              % row #8
        -log(abs(mzf-1))*cp*C1*epsi + keff*cp^2*C5*C1 - cp*C5*epsi*C1 -
keff*cp^2*C4*C1 + cp*C4*epsi*C1+... % row #9
        log(1+mz0)*C1*haz - log(1+mz0)*C1*epsi + log(1+mz0)*C1*keff+...
% row #10
        log(abs(mzf-1))*C1*haz - log(abs(mzf-1))*C1*keff - log(abs(mzf-
1))*C1*epsi+...                       % row #11
        log(abs(mz0-1))*C1*epsi + log(abs(mz0-1))*C1*keff]/...
% row #12
        [(haz + keff*cp - epsi + haz*cp + keff)*(haz - keff - epsi - haz*cp +
keff*cp)*C1];                 % row #13

% Plot ts as a function of mz0
figure(2)
plot(mz0,ts)
title('ts')
xlabel('mz0=cos(theta)')

% Perform a simple numerical inversion of ts to obtain the element g^-1(ts) used
for the computation of the switching time PDF
ts2 = linspace(min(ts),max(ts),nPts);
%ts2 = fliplr(ts);
mz2 = interp1(ts,mz0,ts2,'pchirp');
%mz2 = fliplr(mz0);

% Plot mz2 as a function of ts2 (should be equal to the numerical inversion of ts)
figure(3)
plot(ts2,mz2)
title('mz2')
xlabel('ts2')

tsunnorm = ts2/(gamma*mu0*Ms);      % denormalization of ts [s]

% Calculate the switching time PDF from D'Aquino formulation
tsPDF_daq = mu*keff.*(exp(-mu*(keff/2).*(acos(mz2)).^2)).*abs(alpha.*(keff*mz2 +
haz - ((beta_p/alpha).*((1 + cp*mz2).^-1))).*(1 - (mz2).^2));   % Proposed by
D'AQUINO (corrected with respect to Giulio's code)
tsPDF_daq = tsPDF_daq/trapz(tsunnorm,tsPDF_daq);

% Calculate numerically the switching time CDF from D'Aquino PDF
tsCDF_num = cumsum(tsPDF_daq);
tsCDF_num = tsCDF_num/sum(tsCDF_num);
```

```
tsCDF_num = tsCDF_num/max(tsCDF_num);

% Calculate the switching time CDF from D'Aquino formulation
tsCDF_daq = exp(-mu*(keff/2).*(acos(mz2)).^2);

% Calculate numerically the switching time PDF from D'Aquino CDF
tsPDF_num = diff(tsCDF_daq)./diff(tsunnorm);
new_tsunnorm = tsunnorm(1:end-1)+diff(tsunnorm)./2;

% Plot numerical vs. D'Aquino PDF
figure(4)
plot(tsunnorm*1e9,tsPDF_daq,'r');
hold on;
plot(new_tsunnorm*1e9,tsPDF_num,'b');
xlabel('Time [ns]');
legend('DAQUINO PDF','Numerical PDF');

% Plot numerical vs. D'Aquino CDF
figure(5)
plot(tsunnorm*1e9,tsCDF_daq,'r');
hold on;
plot(tsunnorm*1e9,tsCDF_num,'b');
xlabel('Time [ns]');
legend('DAQUINO CDF','Numerical CDF');

% Generate random samples from CDF through inverse sampling method
num = 1e7;
rng(0)
rnd = rand(num,1);
r_ts = interp1(tsCDF_daq,tsunnorm*1e9,rnd,'linear',0);

% Plot histogram
figure(6)
hist(r_ts,100);
title('Histogram of the switching time (analytical)')
xlabel('Time [ns]');

% Calculate the moments of the histogram
mean_ts = (sum(r_ts)/num)*1e-9
std_ts = std(r_ts)*1e-9
skew_ts = skewness(r_ts)
kurt_ts = kurtosis(r_ts)

rng(0)
r_pears = pearsrnd(mean_ts,std_ts,skew_ts,kurt_ts,num,1);
r_pears = sort(r_pears);
figure(7)
hist(r_pears,100)
title('Histogram of the switching time (Pearson)')
xlabel('Time [ns]');
mean_pears = sum(r_pears)/length(r_pears)
sigma_pears = std(r_pears)
skew_pears = skewness(r_pears)
kurt_pears = kurtosis(r_pears)
[f_pears,y_pears]=ecdf(r_pears);
```

```
figure(8)
plot(tsunnorm*1e9,tsCDF_daq,'r');
hold on
plot(y_pears*1e9,f_pears,'b')
xlabel('Time [ns]');
legend('DAQUINO CDF','PEARSON')
WER = 1-f_pears;
figure(9)
loglog(y_pears*1e9,WER)
xlabel('Time [ns]');

ts_05 = interp1(tsCDF_daq,tsunnorm*1e9,0.5,'linear',0)
WER_target = 1e-6;
ts_WER = interp1(WER,y_pears*1e9,WER_target,'linear',0)
```

ANNEX B: ABSTRACT OF CONFERENCE PAPER

The following abstract takes the basis and work of this thesis and it was extended to the following conference paper.

Assessment of Write and Read Operations in Nanoscaled STT-MRAM Technologies

Esteban Garzón[1], Raffaele De Rose[1], Felice Crupi[1], Lionel Trojman[2] and Marco Lanuzza[1]
[1]*DIMES, University of Calabria, Rende 87036, Italy*
[2]*Micro and Nanoelectronics Institute (IMNE), Universidad San Francisco de Quito (USFQ), Quito, Ecuador*
E-mail address of corresponding author: esteban.garzon@unical.it

1. Summary

This work explores the scalability of STT-MRAMs based on Perpendicular Magnetic Tunnel Junctions (P-MTJs) and a 0.8V FinFET technology through a variation-aware simulation framework. Scaling from the 28-nm down to the 20-nm node allows write energy saving of about 68% at the expense of slightly reduced reading margins.

2. Introduction

Spin-transfer torque magnetic RAMs (STT-MRAMs) are gaining popularity thanks to their promising features in terms of integration density, long data retention, almost zero standby power and full compatibility with CMOS process [1-3]. STT-MRAMs are among the best candidates to replace conventional on-chip memories at advanced technology nodes, especially for normally-off applications in the Internet of Things (IoT) scenario [3]. Despite the above favourable properties and the reduced switching current of perpendicular magnetic anisotropy (PMA) devices, the STT-MRAMs scalability still remains challenging [1, 4]. In this regard, the effect of technology scaling is here explored (considering 28-, 24- and 20-nm nodes) for a 128×128 STT-MRAM array based on circular PMA STT-MTJs and FinFETs. Our analysis exploits a hybrid MTJ/CMOS simulation framework relying on the use of a state-of-the-art MTJ compact model [5] encapsulated in the Cadence Virtuoso design tool. To assure better predictions, the MTJ compact model has been calibrated on experimental data provided in [6].

3. Simulation Framework

Fig. 1 shows the block diagram of our Verilog-A MTJ compact model [5] along with the sketch of the considered MTJ. The model computes the MTJ resistance in both parallel (P) and antiparallel (AP) states, and the switching time (t_S) taking into account its stochastic nature. Depending on the injected current (I_{MTJ}) with respect to the critical switching current (I_{c0}), the model adopts two different formulations for the t_S estimation: (i) the Néel-Brown model [6] for the thermal activation regime (*i.e.* $I_{MTJ} < I_{c0}$), and (ii) the analytical formula presented in [5] for the fast switching regime (*i.e.* $I_{MTJ} > I_{c0}$). Moreover, MTJ process variability related to the oxide thickness (t_{OX}), free layer thickness (t_{FL}), cross-section area, and tunnel magnetoresistance (TMR) is also modelled. Fig. 2(a) provides the architecture of the STT-MRAM array and the four bit-cells configurations considered in this work (Fig. 2(b)-(e)): (b) one NMOS/one MTJ in reverse connection (RC), *i.e.* the access transistor is connected to the MTJ free layer (1T1MTJ-RC), (c) 1T1MTJ in standard connection (SC), *i.e.* the access transistor is connected to the MTJ pinned layer (1T1MTJ-SC), and 2T1MTJ bitcells with NMOS/PMOS transistors in (d) RC and (e) SC.

4. Simulation Results

In the early stage, our analysis was aimed at identifying the optimal bitcell configuration at the 28-nm node for the write operation, which typically requires higher energy cost than the read operation. Having established that the 2T1MTJ-RC has the potential to reduce write delay and energy, the above bitcell configuration was taken as reference for the rest of this study. Figs. 3(a)-(b) show the ratio between the write current (I_{write}) and the I_{c0} for P→AP and AP→P transitions as a function of the bitcell area. Area is here expressed in terms of F^2, where F is the technology minimum feature size. As the MTJ scales, the I_{write}/I_{c0} ratio is enhanced, especially at smaller sizes. From Figs. 3(c)-(d), this translates into lower t_S and write energy (E_{write}) at smaller nodes, also leading the minimum energy point (MEP) moving towards smaller bitcell areas. The effect of process variations on t_S is shown in Figs. 4(a)-(c) for the bitcell sizes corresponding to the MEPs of Fig. 3. For the target write error rate (WER) of 10^{-7}, scaling from 28- to 20-nm node allows delay and energy to be lowered by 20% and 40%, respectively.

The impact of scaling on reading performance was evaluated referring to a conventional voltage sensing scheme [7], where a fixed current (I_{read}) is applied to the bitcell and then the corresponding bitline voltage (V_{BL}) is compared with a reference voltage (V_{REF}) by a sense amplifier. Fig. 5 shows the distributions of the V_{BL}s obtained for an I_{read} that ensures a read disturbance rate (RDR) of 10^{-9} [7]. It also illustrates the sensing margin (*i.e.* $V_{SM} = V_{BL(AP)} - V_{BL(P)}$) and how to set the optimal V_{REF}, *i.e.* the voltage value that makes the read error rate (RER) in the two states exactly the same (*i.e.* $RER_{(P)} = RER_{(AP)}$) [7].

Table I summarizes the main results of this work, suggesting that the technology scaling allows the E_{write} to be reduced (by about 68% from 28-nm down to 20-nm node), while also assuring faster write access and

higher integration density. This occurs at the cost of a slight degradation in terms of reading margins (less than 7% from 28-nm down to 20-nm node).

5. Conclusion

In this work, the impact of technology scaling on writing and reading performance of a 128×128 STT-MRAM array has been investigated. Our analysis has been done exploiting a Verilog-A MTJ compact model and a 0.8V FinFET technology, while considering realistic scaling and variation effects on both MTJ and FinFET devices. Simulation results show that the scaling potentially leads to considerable write energy savings at the cost of a slight decrease of the reading margins.

References

[1] K. C. Chun et al, IEEE JSSC. 48 (2013), 598-610
[2] K. Kwon et al, IEEE TNANO. 14 (2015), 1024-1034
[3] M. Alioto, Springer. 1 (2017)
[4] Z. Xu et al, Elsevier SEE. 102 (2014), 76-81
[5] R. De Rose et al, IEEE TED. 64 (2017), 4346–4353
[6] Y. Zhang, et al, IEEE TED. 62 (2015), 2048–2055
[7] K. T. Quang et al, IEEE ISCAS (2016, Montreal, Canada), 1238-1241

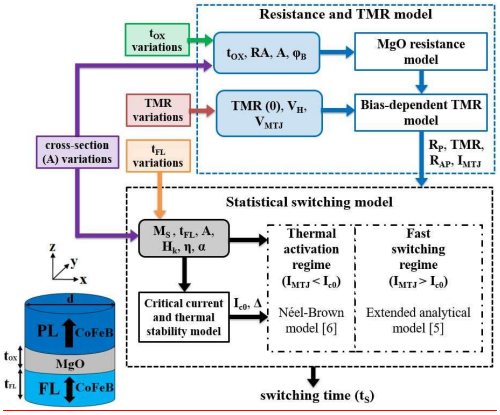

Fig. 1. Block diagram of the MTJ compact model with the sketch of the MTJ device (bottom left). PL: pinned layer, FL: free layer

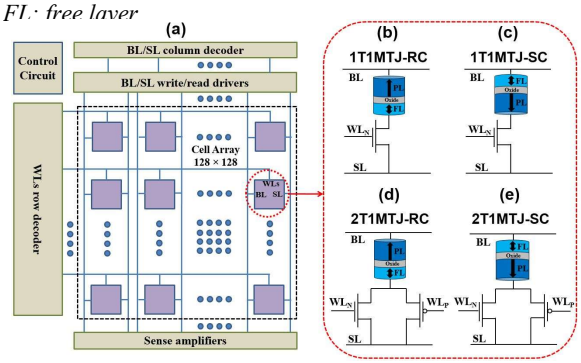

Fig. 2: (a) Reference architecture for the 128×128 STT-MRAM array with the considered bitcell configurations: (b) 1T1MTJ-RC, (c) 1T1MTJ-SC, (d) 2T1MTJ-RC, (e) 2T1MTJ-SC.

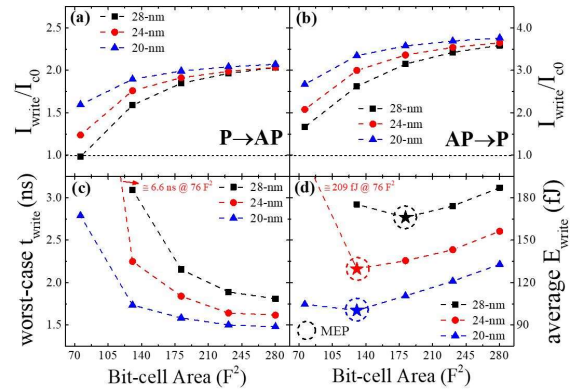

Fig.3: I_{write}/I_{c0} ratio for (a) P→AP and (b) AP→P transitions, (c) worst-case t_{write} for a WER of 10^{-7}, and (d) corresponding average E_{write}, as a function of the bit-cell area at different nodes for the 2T1MTJ-RC bitcell.

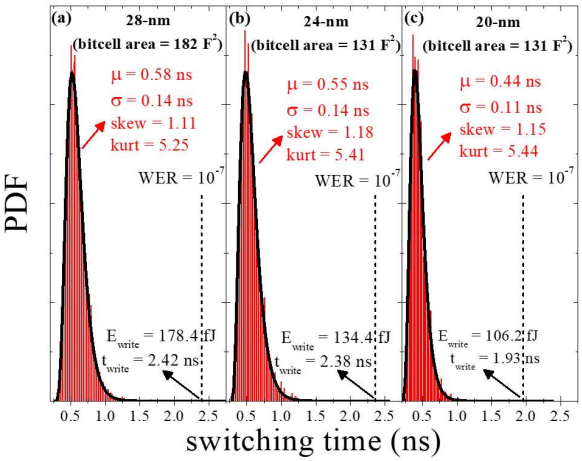

Fig.4: P→AP ts distribution for the 2T1MTJ-RC bitcell at the different technology nodes. The estimation of the t_{write} for the target WER of 10^{-7} and of the corresponding average E_{write} has been done by using a fitting Pearson PDF to account for the right-skewed shape of the ts distribution.

TABLE I
SUMMARY RESULTS

Description	Units	Techn. node (nm)		
		28	24	20
Bit-cell area	F^2	182	131	131
t_{write} (WER = 10^{-7})	ns	2.42	2.38	1.93
E_{write}	fJ	178.4	134.4	106.2
I_{read} (RDR = 10^{-9})	μA	24.56	17.25	10.76
E_{read} (t_{read} = 1 ns)	fJ	8.30	5.64	3.05
Nominal V_{SM}	mV	187	183.4	174
$V_{SM(P)}$	mV	75.5	73.6	68.1
$V_{SM(AP)}$	mV	111.5	109.8	105.9
Optimal RER	---	2.3×10^{-6}	1.8×10^{-6}	8×10^{-7}

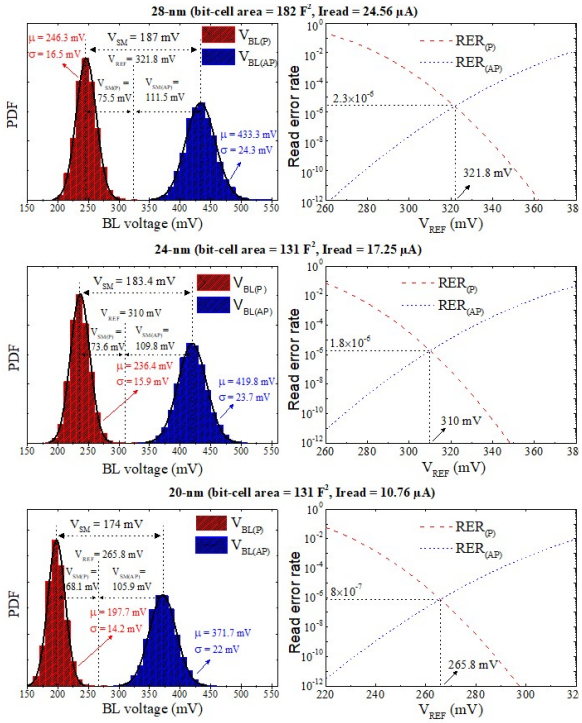

Fig. 5: Statistical distributions of the read V_{BLS} obtained for a fixed I_{read} that ensures a RDR of 10^{-9} and the corresponding estimation of the V_{REF} at different technology nodes.