POLITECNICO DI TORINO

Corso di Laurea Magistrale in

*N*anotechnologies for ICTs

Tesi di Laurea Magistrale

MRAM based neuromorphic cell for Artificial Intelligence

Relatori
Prof. Carlo Ricciardi
Dott. Ricardo Sousa

Candidato
David Salomoni

Anno Accademico 2019/2020

Contents

1 Summary

This thesis project investigates how to exploit MRAM Technologies for applications in Artificial Intelligence (AI). Magnetic Random-Access-Memory technology uses magnetic properties and electron spin to store information. The fundamental component of a MRAM is the Magnetic Tunnel Junction (MTJ), which basic structure is formed by two ferromagnetic layers separated by a thin oxide layer. The thickness of the insulator is small enough that allows electrons to tunnel through it. This quantum mechanical phenomenon is influenced by the magnetization direction of the ferromagnetic layers. If both the layers have the same magnetization directions (parallel orientation), it is more likely that the electrons will tunnel through the barrier. Instead, if they are in opposing directions (anti-parallel orientation), the probability is lower, resulting in two distinct resistance levels (or states). Applications in artificial intelligence architectures require cell elements that can take multiple states based on different cell inputs (e.g. in Artificial Neural Networks) and the use of random number generators (e.g. in Stochastic computing). The goals of the thesis are:

- to identify the possibility of using MTJ pillars connected together, in order to create a multilevel output MRAM cell, and compare parallel and series configurations. The global measured resistance state will depend on the actual states of the individual pillars;
- to study the feasibility of using an MTJ as a Random Bit Generator (RBG).

The thesis describes the work done and the results obtained, during a six months internship at Spintec, CEA Grenoble. The first part of the work was done through simulations using python language, modelling the Multilevel output MRAM cell. The second part was done in the laboratory, developing a measurement setup, for switching probabilities and writing error rates, to validate a possible way to use an MTJ as a random bit generator.

To exploit all the resulting states combinations in a multilevel output MRAM cell (2^N states, "N" number of MTJs connected), a difference in resistance values between the individual connected MTJ has to be present. To achieve this, the diameter size can be varied. The analysis was conducted to determine what diameter difference is best to have a good trade-off between distinguishable resistance states and the switching control of the individual pillars. The initial conditions and stack were chosen from the literature (TMR of 140% and writing pulse width of 10 ns), the best result obtained was at 7 nm difference (23 and 30 nm MTJs). The second step was to compare

parallel and series connected cells. No substantial difference was found between the resistance states separation. Instead, the resistance values are different. As one would expect, for the parallel configuration low resistance will appear, of the order of $\sim 20k\Omega$ and for series high resistance, of the order of $\sim 80k\Omega$. Thus, translating in high current and low voltage for the parallel configuration and the opposite for series. Furthermore, two other interesting aspects have been highlighted. The first is that the parallel configuration needs, in average, less power (around 20%) to switch between states respect to the series configuration. The second aspect is that when dealing with the series configuration, the best way is to control the switching between states through fixed current, and for the parallel configuration through fixed voltage across the whole multilevel MRAM cell. This to avoid dealing with some issues in transitions that may cause the switching also of the other MTJ(s). Finally, a different way to approach the task was studied: Probabilistic Switching. Where the multiple output state cell is composed by MTJs with the same size. This will lead to having the same switching voltage. By decreasing the amplitude (and/or the width) of the writing pulses, the switching probability will decrease, and this can be exploited. The overall behaviour of the cell is to change its resistance state after a certain number of pulses. The transitions will be probabilistic, so the number of pulses needed can vary. What is important is that the resistance will increase or decrease when needed. The drawback is that, because the junctions have the same resistance, the total number of arising states will reduce to $N + 1$ instead of 2^N.

The experimental part of the work was focused on the development of a real time measurement code, that allowed to study the switching probability of MTJs at state-of-the-art level. This was fundamental for the study on an MTJ based Random Bit Generator. To generate random bits, the mechanism used is similar to the writing of data bit using Spin Transfer Torque (STT-MRAM). The difference is that one writing voltage is set to a lower amplitude, so that the switching probability is exactly 50%. In this way, it is possible to exploit the stochastic nature of STT switching, to generate Random Bits. To evaluate the "quality" of the randomness of the bit sequence, a cumulative sum was compared to the one of MATLAB's random bit generator. When the sequence has a bias towards one state, the resulting cumulative sum will diverge. Further data analysis of the quality of randomness, such as the statistical test suite NIST SP-800, were not carried out, because the experiment done was not reproducible in a satisfactory manner. In fact, the results coming from different measurements, for the same applied voltage, converge to different probability values giving rise to a substantially different cumulative sums. Nevertheless, the results showed true potential for the application, as already mentioned in the literature. The lack of control of the 50% probability switching voltage, is driven by

electrical noise and by thermal fluctuations that make the switching voltage to have a distribution. Possible solutions can be studied including decreasing the writing pulses width for thermal purpose, or by implementing an algorithm that adjusts the switching pulses, similarly to a PID controller. Although the thesis work has given lots of new knowledge on the topic of MRAM for AI, it deserves a direct experimental approach to continue and complete the work done.

2 Introduction

This first chapter gives an introduction to the main fields necessary to understand the thesis objective. A first section will give an overview on Artificial Intelligence focusing on Artificial Neural Networks and Deep Learning. A second section introduces MRAM technology and its working principles.

2.1 Introduction to Artificial Intelligence and Neuromorphic computing

Artificial Intelligence (AI) gained lots of attention in the last decades. The meaning of AI though may differ from person to person, some people associate it to artificial lifeforms that can surpass human intelligence, other to any data processing technology. This section will serve as an introduction to the main principles of AI and Machine learning.

2.1.1 Artificial Intelligence

The popularity of AI is partly due to the fact that the term is now used for topics and fields that used to be called by other names. This happens because, also among the AI researchers, there is no exact definition of AI. For example, the English Oxford Living Dictionary gives the definition: "The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages". And the Encyclopedia Britannica states: "Artificial intelligence (AI), the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings" [1]. There is in fact a continuous redefinition of the field when some topics are classified as non-AI, and new topics emerge. For instance, fifty years ago automatic methods for search and planning were considered to belong to the domain of AI, while nowadays such methods are taught to every computer science student. One different, more useful, way to define AI is through a list of required characteristic properties: the first one being *autonomy* (the ability to perform tasks in complex environments without constant guidance by user); the second one being *adaptivity* (the ability to improve performance by learning from experience) [2][3].

2.1.2 Machine learning

One sub-field of AI is machine learning (ML), that can be defined as "the study of computer algorithms that improve automatically through experience" [3]. The key ability of an AI system is the capability to acquire its own knowledge through patterns from raw data. At the heart of machine learning there is statistics, especially methods such as linear regression and Bayesian statistics. Machine learning can be divided in subareas depending on the problems and goals to achieve. We can categorize them as:

- Supervised learning: The task is to predict the correct output or label related to a given input.
- Unsupervised learning: The goal is to find the structure of the data (for example "clusters", "dimensions"). There are no labels or correct outputs.
- Reinforcement learning: The AI agent (like a self-driving car) must operate in an environment, feedback about good or bad choices is available with some delay.

Taking as an example a classification problem, where images have to be correctly labelled. In supervised learning the idea is to take a number of examples and label each one by the correct label, and then use them to "train" an AI method to automatically recognize the correct label for the training examples as well as any other image. For complex classifications this is far more convenient than writing down all the exact rules as done in conventional hard-coded programs.

2.1.3 Artificial Neural Networks and deep learning

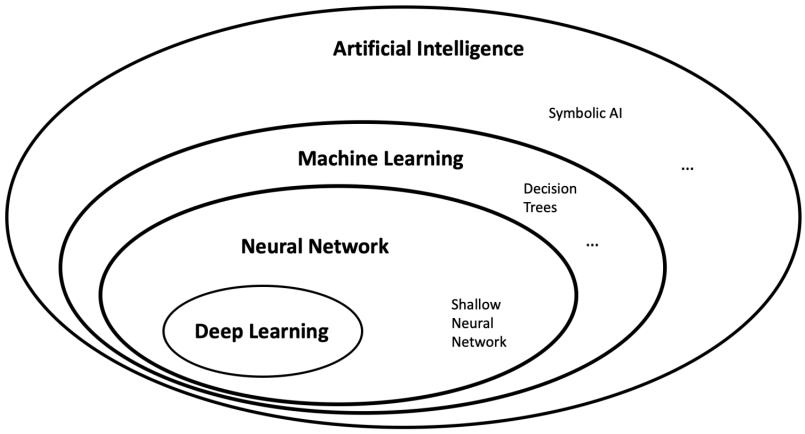

Figure 2.1.1: Venn diagram showing the relations between Artificial Intelligence, machine learning, neural network and deep learning. [4]

Artificial Neural Networks (ANN) are brain inspired structures. They achieved great progress in the recent years, it was possible thanks to the increased computing power that gave the ability to increase the complexity of the networks. Furthermore, the progress is also due to massive data sets and deep learning techniques. ANN is a sub-field of machine learning. There are also different non-neural network machine learning techniques, for example decision trees. Deep learning is a sub-field of ANN, the "depth" it refers to, is the complexity of a mathematical model or network. The structure is formed by "layers" of simple processing units (neurons) that are connected together; this enable the system to pass the information from the input through every single unit. The number of layers, and so the depth of the network, enables it to learn complex concepts build out of simpler concepts. Figure 2.1.2 gives an example of how a deep learning system can identify the concept of an image.

The visible layer, (the input pixels) is connected to a second layer called the first hidden layer, because it contains abstract features from the image, contrary from the visible layer (the input) that contains the observable variables. The network starts from a simple concept such as edges, that are easy to find by comparing the brightness of neighbouring pixels. Once the first hidden layer has identified the edges, the second hidden layer is able to find contours and corners, that can be seen as collection of them. From the contours and corners, the third hidden layer can

recognize different objects parts. In the fourth layer (output) the object identity can be found [3].

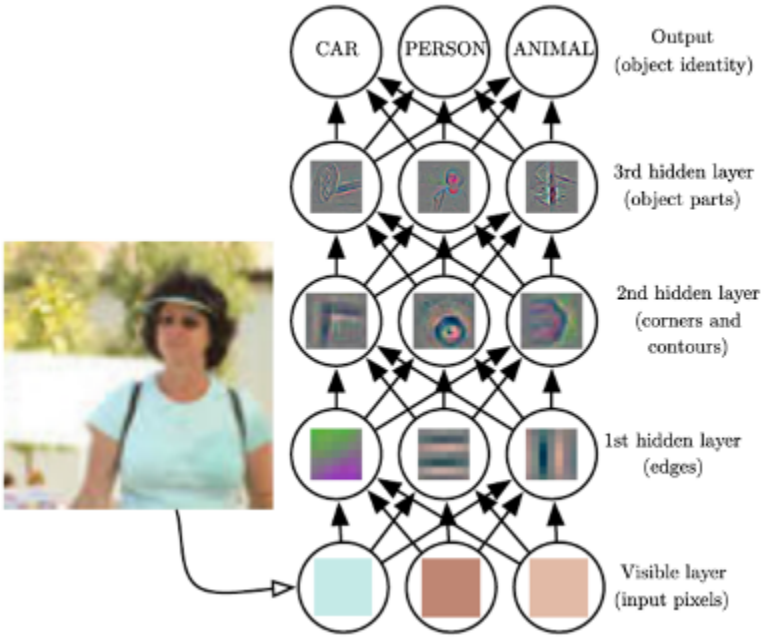

Figure 2.1.2: Deep learning model illustrated performing an image recognition [3].

Artificial Neural Networks, as said before, are brain inspired structure, it is then interesting to see what the key components of a Neural Network are.

A neural network, either biological or artificial, is composed by a large number of simple units called neurons. These neurons are able to receive and transmit signals to each other through dendrites, axons and synapses. The dendrites are the "wires" that provide the inputs to the neurons, every neuron has one axon, that is used to send the output. Each axon is connected to one or more dendrites. The connection between an axon and one dendrite is the synapse that has the role of weight, it can make the connection "strong" or "weak" depending on the needs.

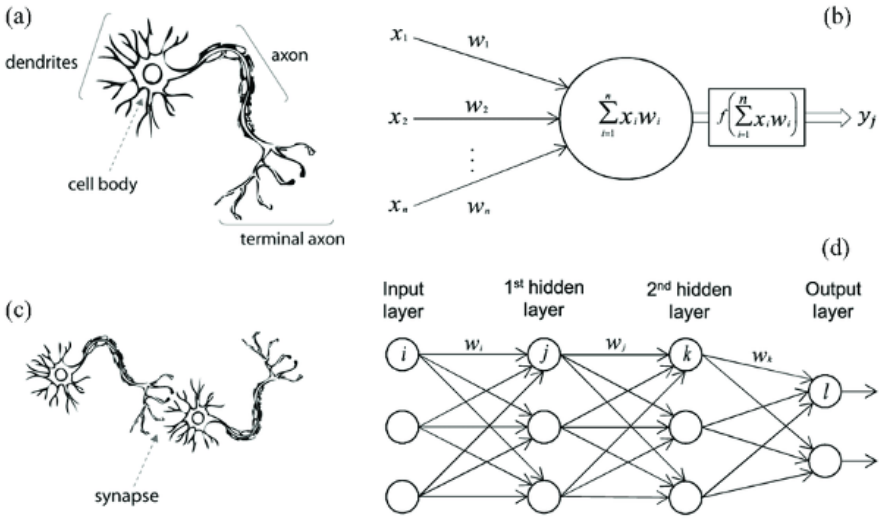

Figure 2.1.3: A biological neuron in comparison to an artificial neural network: (a) human neuron; (b) artificial neuron; (c) biological synapse; and (d) Artificial Neural Network [5][6].

Keep in mind that the goal of ANNs is not to simulate and model how the brain actually works, this concerns computational neuroscience, the actual aim is to build computer systems that can successfully solve tasks requiring intelligence. This means that like airplanes don't flap their wings to fly, likewise, in artificial neural networks the internal mechanism of the neurons is usually ignored, and the artificial neurons are often much simpler than their natural counterparts.

Artificial Neurons are simple processors of information, the input coming from other neurons is weighted by the synapse. What a neuron does is similar to a linear combination of the inputs ($linear\,combination = intercept + weight_1 \times input_1 + ... + weight_n \times input_n$), after the linear combination has been computed, the neuron puts it through a so-called activation function. Usually an activation function includes:

- identity function: do nothing and just output the linear combination
- step function: send a pulse (ON) if the value of the linear combination is greater than a threshold value, otherwise do nothing (OFF)
- sigmoid function: a "soft" version of the step function

A neuron on its own is not really capable of much, but if there are lots of them connected, the system can get extremely complex, as shown previously in fig 2.1.2 .

Every neuron and connection reacts to the incoming signals and adapts over time, this adaptation is known to be the key to functions such as memory and learning. In particular, these functions occur when the weights are adjusted to make the network produce the correct outputs. Generally, the weight adjustments are done using backpropagation and the same ideas as in linear or logistic regression. The network is fed with training data, one example at the time, each mistake or misclassification will result in an update in the weights. Many neural networks are extremely large, and the largest contain hundreds of billions of weights. Optimizing them all can be a challenging task that requires massive amounts of computing power. Thus, the key requirements for a hardware realization of such element is the ability to take multiple states based on different inputs, in the more efficient way possible, together with a non-volatile behaviour, in order to "remember" everything learned.

Comparing neural networks to traditional computer, two key features are distinguishable. For one, in a traditional computer, CPUs focus only on one thing at the time, instead neural networks are able to process vast amount of information simultaneously. Secondly the data storage is not separated as in traditional computers, but it can be stored short term in the neurons or long term in the weights, (synapses). These two differences make the models suited for different tasks.

2.1.4 Random bit generator

Many emerging computing schemes, as the ones described and other non von-Neumann architectures, have a critical relationship with random number generation. In fact, randomness can be used as a tool or a feature in preparing data for learning algorithms able to perform predictions from mapping input data, and more in general it helps the learning algorithms to be more robust and ultimately more accurate. Usually algorithms that exploit randomness are referred as stochastic algorithms or stochastic computing. Some examples of exploiting randomness are:

- Shuffling of training data
- Random subset of input features used in Random Forest Algorithms
- Random initial weights in ANN

Random number generators can be classified into two groups: Pseudo-Random Number Generator (PRNG) and Truly Random Number Generators (TRNG). PRNGs are performed by software algorithms that generate a sequence of RNs. They require seeds of RNs and the sequence of RNs will be always the same if coming from the same seed. TRNGs are implemented in hardware and generate a sequence of RNs

using a nondeterministic physical event.

2.2 Introduction to MRAM technology

This second section serves as an introduction to the technology used during this internship thesis. A first presentation to the MRAM technology and its place in the market will be given, followed by an overview on its working principles. Finally, the aim of this work is explained, that is, how to exploit MRAM technology as a hardware component for Artificial Intelligence.

2.2.1 Context:

Since the 20th century, when computers were created, the need for storing data has grown and kept evolving. Nowadays the main recording storage is the digital one, it has reached 50% of the market in the year 2002, marking the start of the digital age.

Memory devices can be divided into two main categories: Stand-alone and Embedded. In the first, the memory acts as an external component to the main unit and requires the highest storage capacity. Applications are in the industry market, enterprise and consumer storage (serves, HDD) and mass storage such as USB sticks, SD card or SDD. In the embedded category, instead, the memory is merged together with the main unit. The storage capacity is smaller but usually requires higher speed. They are mainly used for mobile devices, cache memories, micro-Controller units (MCU) and System On Chips (SOC). At the moment, the mass production relies on three main types of memories. For stand-alone market, Flash (NAND) technology is dominant, due to its high density, non-volatility and low cost. For the embedded market, Static RAM and Dynamic RAM are used. The two have common history and, even if they are volatile memories, they are used mainly due to their high speed, close to the nanosecond range. In figure 2.2.1 it is shown the tree view of the different memory categories between volatility and non-volatility, that is the ability of keeping or not the information when plugged off.

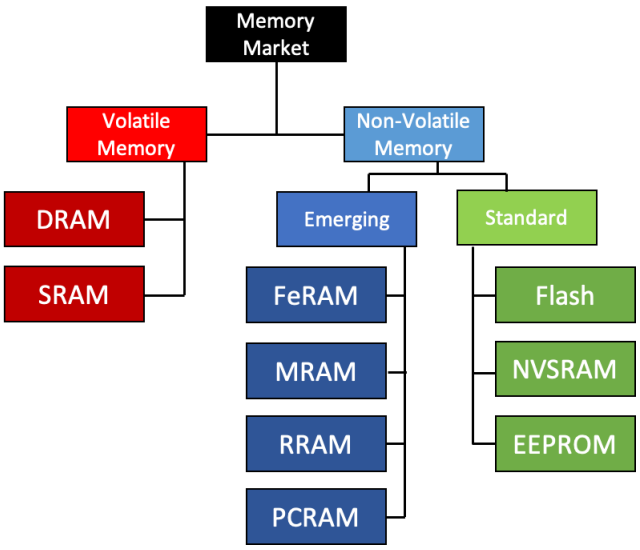

Figure 2.2.1: Tree view of memory market.

As shown in figure 2.2.1 the MRAM technology is an emerging non-volatile memory (e-NVM). MRAM technology is still far from realizing its potential, but as of early 2020, there are MRAM chips on the market ranging from very small ones to 1Gb chips, and companies are adopting this technology for many applications [7]. Analyst expect MRAM shipments and revenues to grow considerably in the next few years as many companies and research groups are developing next-generation MRAM technologies. Object Associates for example expects that the stand-alone MRAM and STT-MRAM revenues will grow 170X from 2018 to 2029 [8]. For the objective of the thesis it is not important to discuss about MRAM properties as a memory device but instead in the following we will discuss more in detail about its working principles.

2.2.2 MRAM working principle

Magnetic Random-Access-Memory is a Spintronic device, this technology uses magnetic properties and electron spin to store information. The fundamental component of a MRAM is the Magnetic Tunnel Junction (MTJ), which basic structure is formed by two ferromagnetic layers separated by a thin oxide layer. The thickness of the insulator is small enough (typically few nm), that allows electrons to tunnel through it. This quantum mechanical phenomenon is influenced by the magnetization direction of the ferromagnetic layers. If both the layers have the same magnetization direction (parallel orientation), it is more likely that the electrons will tunnel through the barrier, instead, if they are in opposing direction anti-parallel orientation), the probability is lower. One of the layers is called pinned-layer and it is kept always in the same direction of magnetization, the other, the free-layer, it can be instead switched from one state to the other. This leads to two different electrical resistances, one high and one low, that are useful for many applications. The resistance difference between the two stable resistance states of the MTJ can be characterized by the tunnel magneto-resistance ratio ($TMR = (R_{AP} - R_P)/R_P$). The different way of writing the data bit into the MTJ are presented later on. In fig 2.2.2 it is shown an example of a P-STT-MRAM typical bit-cell structure.

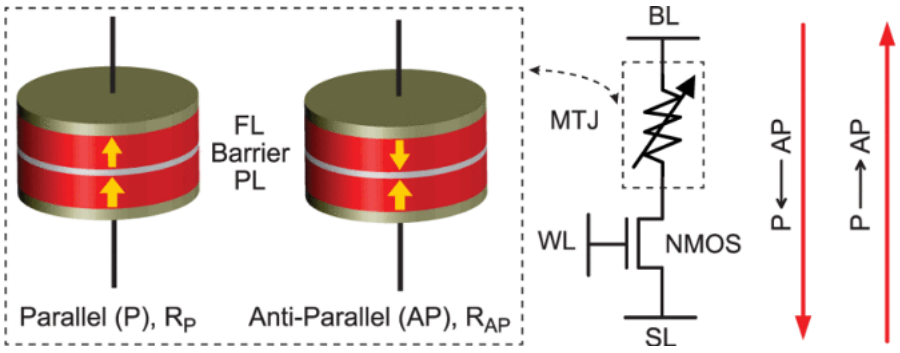

Figure 2.2.2: P-STT-MRAM typical 1T1MTJ vit-cell structure [9].

2.3 Thesis objective

The objective of the thesis is to study how to exploit MRAM technology for Artificial Intelligence applications. MRAM magnetic memories combine non-volatility with writing speeds of tens of nanoseconds. In conventional binary memory applications, every cell is characterized by only two states. However, synapses have the requirement of reaching multiple states, ideally in an analog manner. It is possible to obtain a multiple state cell based on individual MRAM pillars connected as a parallel or series resistor. The resistance measured across the whole cell will depend on the actual states of the individual pillars.

A second way to exploit the MRAM cell for AI, is to use it as a true random generator. This is possible thanks to the intrinsic probabilistic way of switching through spin transfer torque. If a 50% switching probability voltage is reachable this can translate to a random generator.

The work was carried out in two parts. At first, studying through simulations the possible arrangements of a multiple state MRAM cell. The possibility of switching each pillar independently relies on the natural dispersion of the switching voltages. This has to be taken into account together with how to have distinguishable resistance states. Varying the diameter size of the connected junctions is the method that was studied. Also a comparison between series and parallel configuration was done, including power consumption and the possibility to have deterministic or probabilistic switching. Secondly an experimental work was done. Developing a real time measurement set up and analysis code able to reach tests at state-of-the-art level. This was used to find a possible way to exploit MRAM cells as random bit generator, by varying the writing voltages across the junctions.

3 P-STT-MRAM

Magnetic Random-Access-Memories are a family of technologies that rely on MTJ and have the same reading principle. In this chapter, a more detailed explanation is given of the physical principles governing the perpendicular spin transfer torque MRAM technology (P-STT-MRAM). A fist part gives the basics of magnetism, it is followed by the definitions of Tunnelling Magnetoresistance Ratio and Perpendicular Anisotropy that are useful to understand the description of spin transfer torque.

3.1 Magnetism

To understand P-STT-MRAM some knowledge of magnetism is required. In this section the basics of magnetism are presented in order to help appreciate the subsequent explanations on Spin Transfer Torque.

The elementary quantity in solid-state magnetism is the magnetic moment $\overrightarrow{m}$. The intrinsic magnetic moments, inside a material, are associated with the spin of each electron and its orbital motion around the nucleus. The magnetization $\overrightarrow{M}(r)$ is a mesoscopic volume average of the dipole moment $\overrightarrow{m}$. The primary magnetic field $\overrightarrow{B}$ is related to the auxiliary magnetic field $\overrightarrow{H}$ and the magnetization by $\overrightarrow{B} = \mu_0(\overrightarrow{H} + \overrightarrow{M})$. Sources of magnetic field are magnetized material, as already said, and electric currents (Bios-Savart's law). The field produced by a given distribution of magnetization can be calculated by integrating the dipole field due to each volume element $\overrightarrow{M}(r)dV$, or using the equivalent distributions of electric currents or magnetic charges [10]. There are different types of magnetic materials. Inside a substance, electrons, following Pauli's exclusion principle, can combine into pairs with opposite magnetic moments in order to cancel each other out, resulting in a global null magnetic moment. In presence of an external magnetic field, magnetic moments can appear as a counteraction, and align themselves in the opposite direction of the external field, these are known as diamagnetic materials. Pure diamagnetic materials will tend to be repelled be magnetic fields. In some cases, though, electrons cannot pair entirely resulting in atoms having non-zero magnetic moment. Depending on how the magnetic moments are ordered, there will be a global magnetization, or not, of the material. The global magnetization of the material, and the different response to an external magnetic field, will determine what kind of magnetic material it is. If all the magnetic moments align along the same direction of the external field the material is called ferromagnetic. It will result in a global non-zero magnetic moment. Instead, in anti-ferromagnetic materials, the magnetic moments remain ordered but

anti-parallel to each-other, leading to cancel out the global magnetic moment. Other types of materials exist like, for example, superparamagnetic or ferrimagnetic.

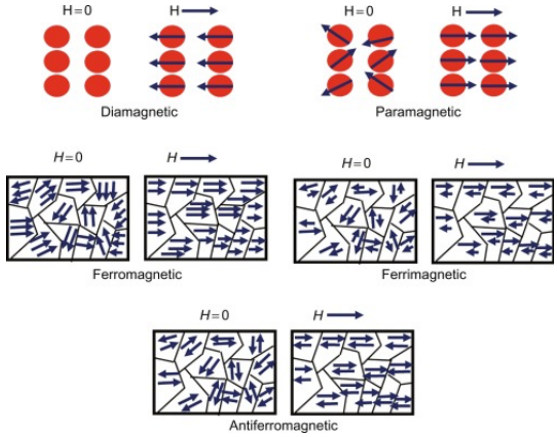

Figure 3.1.1: A schematic diagram depicting the ordering of spins in ferromagnetism, antiferromagnetism, ferrimagnetism, and paramagnetism with associated applied magnetic fields H [11].

"The essential practical characteristic of any ferromagnetic material is the irreversible nonlinear response of magnetization $\overrightarrow{M}$ to an imposed magnetic field $\overrightarrow{H}$. This response is epitomized by the hysteresis loop" [10]. The magnetization saturates at high value of magnetic field, this value is called saturation magnetization (M_s). The value of magnetic field at which the magnetization changes polarity is called the coercive field (H_c).

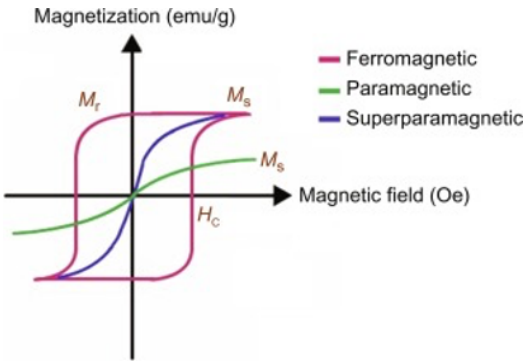

Figure 3.1.2: Hysteresis loops characteristic of ferromagnetic, paramagnetic and superparamagnetic. M_S is the saturation magnetization, M_r the remanence and H_C the coercive field [11].

3.2 Tunnelling Magnetoresistance Ratio

As said previously, the fundamental component of a MRAM is the magnetic tunnel junction. The key phenomenon that the junction exploits is quantum tunnelling, that allows a current to pass through it. Charged carriers, like electrons, have an intrinsic property called spin, it can be up or down. Usually currents are unpolarized, meaning that the number of electrons with spin up is the same as the one with spin down. A ferromagnet acts like a spin filter, when a current passes through it, the ratio between spin up and spin down electrons will change depending on the magnetization orientation, thus it will lead to a spin polarized current. Equivalently if you consider the density of states (DOS) of a magnetize material, as shown in fig 3.2.1, at a given Fermi Level there is an imbalance between the DOS of electrons with spin up and the DOS of electrons with spin down depending on the magnetization, leading to a polarized current. The oxide between the metal layers, usually MgO, creates an energy barrier, the current that passes through the pinned layer will be polarized along its magnetization direction and it can tunnel across the barrier. The probability of tunnelling through it is given by the Fermi Golden rule:

$$P^{\sigma} \propto \langle i|W|f\rangle^2 D_2(E_F) \tag{3.1}$$

The current will be proportional to:

$$J^{\sigma} \propto D_1^{\sigma}(E_F) \times D_2^{\sigma}(E_F) \tag{3.2}$$

Where σ is the spin, D_1 and D_2 are respectively the density of states in the pinned layer and free layer. The free layer can be either parallel or anti-parallel to the pinned layer. When the layers are parallel (P) the DOS will match, thus the electrons will be transmitted efficiently, the conductance factor in this condition will be named G_P. When the free layer is anti-parallel (AP) the DOS are mismatched, in this case not all the electrons will find states in the free layer leading to a lower probability of tunnelling and so lower current, the conductance is named G_{AP}.

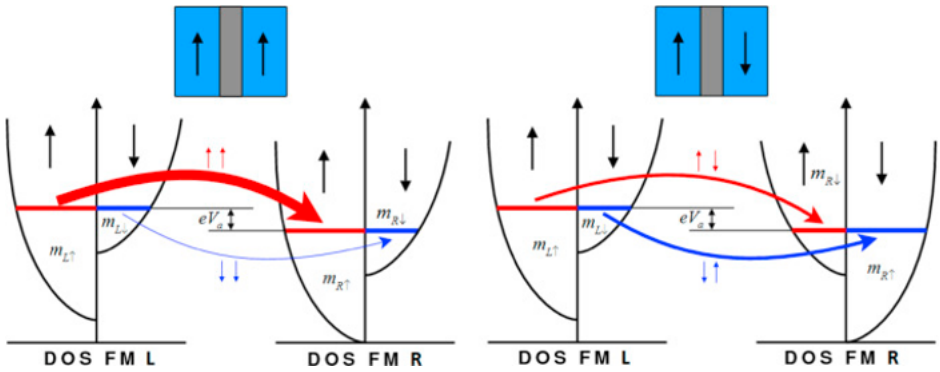

Figure 3.2.1: Schematic of the TMR effect in an MTJ, two-current model for parallel and anti-parallel alignment of the magnetizations [12].

The relation between conductance and DOS of the junction layers can be expressed by:

$$G_P \propto D_1^{\uparrow}(E_F)D_2^{\uparrow}(E_F) + D_1^{\downarrow}(E_F)D_2^{\downarrow}(E_F), \tag{3.3}$$

$$G_{AP} \propto D_1^{\uparrow}(E_F)D_2^{\downarrow}(E_F) + D_1^{\uparrow}(E_F)D_2^{\downarrow}(E_F) \tag{3.4}$$

To characterize the difference between the conductance in P or AP states the tunnelling magnetoresistance ratio is used (TMR) defined as:

$$TMR = \frac{G_P - G_{AP}}{G_{AP}} \tag{3.5}$$

Or with resistance values:

$$TMR = \frac{R_{AP} - R_P}{R_P} \tag{3.6}$$

The polarization factor P is given by the ration of the DOS:

$$P^i = \frac{D_i^{\uparrow}(E_F) - D_i^{\downarrow}(E_F)}{D_i^{\uparrow}(E_F)D_i^{\downarrow}(E_F)} \tag{3.7}$$

Where i can be 1 or 2 depending on the layer. The polarization factor P tells how many electrons are effectively polarized along the layer's direction; it is equal to 1

if all of them are and it depends on the material. The TMR can be written also in terms of polarization factors.

$$TMR = \frac{2P_1P_2}{1 - P_1P_2} \tag{3.8}$$

If the layers are both of the same material, and of similar thicknesses, their polarization factors are considered to be equal ($P_1 \approx P_2$), so:

$$TMR = \frac{2P^2}{1 - P^2} \tag{3.9}$$

So high polarization factors translate in high TMR.

3.3 MRAM generations

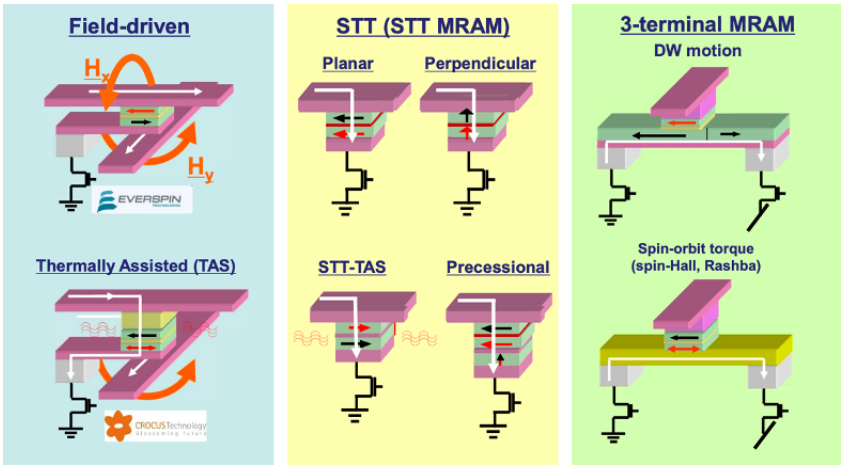

Figure 3.3.1: MRAM generations: on the left Field Driven, in the center STT-MRAM and on the right the new generation of three terminal devices [13].

Over time, different generation of MRAM have been developed, in fig 3.3.1 the main types are illustrated. They all rely on magnetic tunnel junctions and the read and storage principles are always the same. To read the state, the resistance value of the MTJ is measured by a current flowing through the junction, and the storage relies on the magnetization orientation of the free-layer. The difference between the categories of MRAM stands in the writing principle, meaning how the free layer of

the MTJ is switched. The first generation, field-driven MRAM, also known as toggle-MRAM uses external magnetic field coming from metal lines over and underneath the junction to switch the storage layer. A further step was introduced with Thermally Assist MRAM. The switching is helped by heating the junction and so reducing the thermal stability of the junction. The second generation, the one the thesis work will focus on, exploits spin transfer torque (STT-MRAM). A current passing through the device is able to switch the free layer. They are interesting thanks to their downsize scalability and lower power consumption. The STT working principles are seen in more detail later on. The 3 terminal MRAM are the third generation of MRAM, either a domain wall motion (DW motion) or Spin Orbit Torque (SOT) are exploited. In DW motion MRAM the free layer is replaced by a magnetic line; a horizontal current flowing through it controls the position of the domain wall and so the junction configuration. SOT-MRAM exploits a heavy metal line underneath the junction, that has a high spin-orbit coupling. When a charge current flows, a spin current is injected in the free layer, causing it to switch.

3.4 Perpendicular Anisotropy

The orientation of the spontaneous magnetization, that is the magnetic moment of the material when no external magnetic field is applied, will be determined by the magnetic anisotropy, depending on the magnetic object. There are three main contributions to the magnetic anisotropy.

The **Magneto-Crysitalline anisotropy** K_u, that is due to preferential crystallographic directions of the local magnetic moments, dictated mostly be spin-orbit interactions. If an external magnetic field is applied, the magneto-crystalline energy is:

$$E_{mc} = K_u \sin^2 \theta \tag{3.10}$$

where θ is the angle between the external field and the magnetization preferential axis.

The **shape anisotropy**, that originates from the dipole induced by the uniform magnetization of the ferromagnet. The dipole creates a demagnetizing field that forces the magnetization to align itself along the longest dimensions of the object. The demagnetizing field $\overrightarrow{H_d}$ is written as:

$$\overrightarrow{H_d} = [N]\overrightarrow{M} \tag{3.11}$$

where [N] is the demagnetizing tensor related to the shape of the magnetic object.

The **surface anisotropy** K_s instead is related to the interactions happening at the interfaces between to materials. At the nano-scale, where the surface to volume ratio increases remarkably, these contributions are significantly important and they can have an impact on the demagnetizing field and change the magnetization's orientation. P-STT-MRAM rely strongly on this parameter. In order to have perpendicular anisotropy the surface anisotropy has to overcome the shape anisotropy and force the magnetization out-of-plane. Thus, to achieve this condition, correct thicknesses and materials of the layers are adopted.

To pin-down the reference layer of the MTJ, it is used a Synthetic Anti-Ferromagnet (SAF). It is a three-level system that thanks to the so called RKKY interaction is able to fix the magnetization direction of the reference layer while keeping it magnetically neutral.

3.5 Spin Transfer Torque and LLGS equation

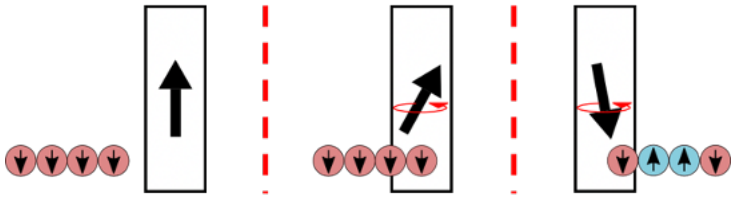

Figure 3.5.1: Spin Transfer Torque qualitative illustration. The red and blue circles represent the electrons with their spin. Moving from left to right the polarization of the current changes as well as the magnetization of the ferromagnet, due to spin transfer torque.

Magnetize materials, we have seen, can induce a current to be polarized in the same direction of their magnetization. This effect is reciprocal, in fact, also a polarized current can induce a ferromagnet to magnetize in one direction, it is called spin transfer torque (STT). Like the name suggests, the electron intrinsic magnetic moment is transferred to the material causing it to switch from one magnetization state to the other.

The behaviour of the magnetization of a ferromagnet can be described by the known Landau–Lifshitz–Gilbert (LLG) equation, taking into account the additional term

due to STT effect, introduced for the first time, in 1996 by Slonczewski[14]:

$$\frac{d\overrightarrow{m}}{dt} = -\gamma_0(\overrightarrow{m} \times \overrightarrow{H}_{eff}) + \alpha\left(\overrightarrow{m} \times \frac{d\overrightarrow{m}}{dt}\right) - P\frac{Jg\mu_B}{2eM_St}\left(\overrightarrow{m} \times (\overrightarrow{m} \times \overrightarrow{p})\right) \tag{3.12}$$

where $\overrightarrow{m}$ represents the free magnetic moment, α the damping factor, γ_0 the gyromagnetic ratio, P the spin polarization factor, J the current density, e the electron's charge and $\overrightarrow{p}$ the easy-axis unity vector, H_{eff} the effective magnetic field that takes into account the external field, the demagnetizing field and anisotropy. The first term induces a precession on the magnetic moment around the effective magnetic field due to the external magnetic field, the term is known as "Field torque". The second term, the "Damping torque" tends to bring the magnetization closer to $\overrightarrow{H}_{eff}$, in other words, it attenuates the amplitude of the precession, the efficiency depends on α. The third term is the "STT effect" which tends to align the magnetic moment towards the direction of the polarized current $\overrightarrow{p}$. Represented in red in figure 3.5.2, this contribution will help or contrast the damping torque depending on the direction of J.

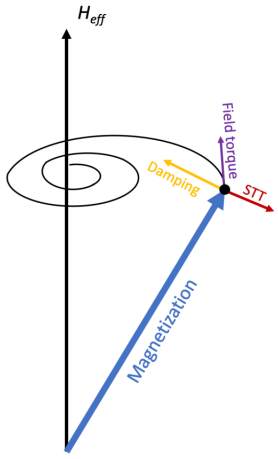

Figure 3.5.2: Magnetization dynamics described by LLG equation, the STT contribution in this illustration is not sufficient to switch the magnetization direction. Note that the direction of the STT contribution could also be helping the Damping torque, depending on the direction of the current.

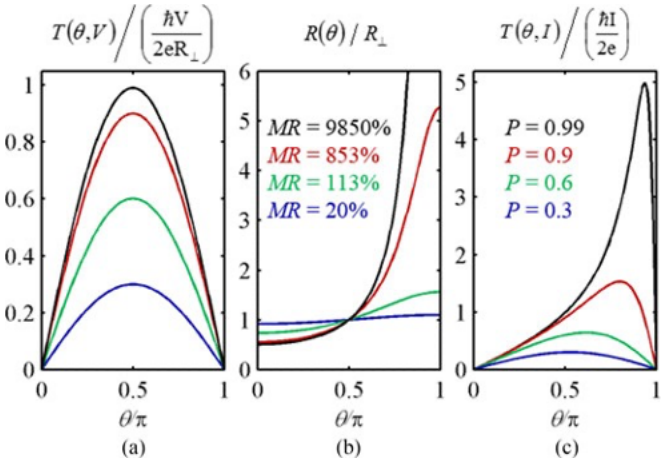

Figure 3.5.3: For four different values of P and considering $P_1 = P_2 \equiv P$. (a) Normalized spin torque for constant voltage applied across the junction. (b) Normalized resistance. (c) Normalized spin torque for constant current applied through the junction [15].

From figure 3.5.3 (results from the literature) it is possible to notice that the transfer torque is not the same if the junction is held at fixed voltage respect to when a constant current is applied. In fact, in the first scenario, the spin torque is symmetric about $\theta = \pi/2$, where θ is the angle between the free layer and the reference layer magnetization's direction:

$$T(\theta, V) = \frac{\hbar P_R}{2eR_\perp} V \sin\theta, \tag{3.13}$$

where, $R_\perp$ is the resistance when the free-layer and fixed-layer magnetization are perpendicular, and P_R is the spin polarization of the reference layer. This translates to having the same switching voltage for switching in both the polarities $AP \rightarrow P$and $P \rightarrow AP$. Instead, at constant current, the spin torque is larger near $\theta = \pi$, (P_F is the spin polarization of the free-layer),

$$T(\theta, I) = \frac{\hbar P_R}{2e} \frac{I}{1 + P_F P_R \cos\theta} \sin\theta, \tag{3.14}$$

leading to a switching current lower for $AP \rightarrow P$ than for $P \rightarrow AP$ [15][16]. These concepts are important for the power consumption analysis done later on. In fact, being the two switching voltages the same, it means that the power needed to switch the junction will be lower for one transition with respect to the other one ($P = \frac{V^2}{R}$, where R changes depending on the transition).

4 Multilevel MRAM cell Simulation

Artificial Intelligence applications, in particular Artificial Neural Networks require synapses able to reach multiple states. This because the synapses have the important role of memorizing and learning. Furthermore, the resistance value of every synapse has to be perfectly tuned in the whole neural network during the learning process. The combination of all the synapses and spiking neurons will give the ability to the network to achieve the task it was trained for. Thus, more states available for the synapse and more it can tune itself to the correct weight.

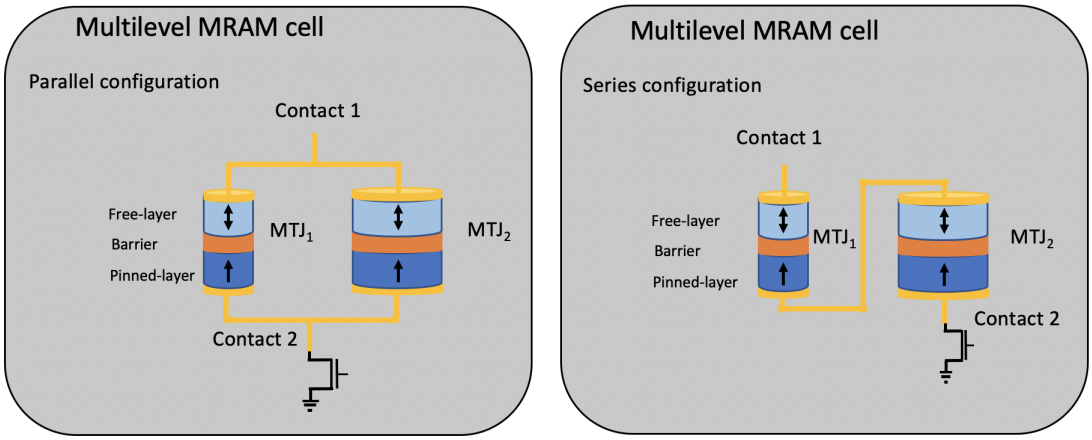

Figure 4.0.1: Illustration of a Multilevel Output MRAM cell with parallel and series connections between the two MTJ with different diameters.

One way of using MRAM cells for artificial synapses is to connect individual pillars (MTJ) together in order to have an overall cell with multiple states given by the combination of the individual ones. Every MTJ has two resistance states, one low when the magnetization directions are parallel (P) and one high when they are anti-parallel (AP). If two MTJs are connected they will give rise to a cell with ideally four possible resistance states:

1. P-P, lowest resistance
2. AP-P, intermediate resistance
3. P-AP, intermediate resistance
4. AP-AP, highest resistance

This is true if the resistance states of the two pillars are different, in fact, if they are not, state 2. and 3. will result in the same resistance value. To exploit all four the combination, the pillars must have different resistance, this situation is reachable having a difference in the size of the diameter of the MTJs. The larger MTJ will have a lower resistance with respect to the smaller one, translating in four overall states. In general, for a number N of MTJs connected, the possible combinations of states are 2^N, if the junctions are equal, the total number of distinguishable states reduces to $N + 1$.

MTJs	**2**	**3**	**4**	**5**	**6**
$\mathbf{2^N}$	4	8	16	32	64
$\mathbf{N+1}$	3	4	5	6	7

Table 1: Table highlighting the difference in the output states exploiting all 2^N combination respect to only $N + 1$, where N is the number of MTJs connected (first row). Of course these are ideal numbers, in reality the feasibility of connecting a large number of junction must be taken into account.

In the following sections the possibility of having a multiple state MRAM cell, with **two** MTJs connected, and its operating window were studied through simulations done by developing a python code.

4.1 Diameter dependence analysis

As said previously, in order to reach all the possible combination of states, a difference in resistance has to be present between the junctions. To achieve this one can vary the diameter size of the junctions. For this reason, the first part implemented in the code was to generate all the main parameters for the MTJ, that will be explained further on, studying their dependence on the diameter size of the pillars. The stack used during the simulation was taken from the literature [17].

The first figure of merit calculated was the thermal stability factor. The thermal stability factor is an important performance metric, defining its data retention capability, that is the time the junction is expected to remain in current state. This is expressed by the ratio between the barrier the junction has to overcome to change state over the thermal energy ($\Delta = \Delta E/k_BT$). The thermal stability dependency on diameter size has been already studied in literature, and it turns out that Δ is constant when D is larger than D_n, and D_n is of the order of the domain wall width $\delta_w = \pi(A_S/K_{eff}^{MTJ})^{0.5}$, where A_S is the exchange stiffness constant and A_S was

inferred to be 19 pJ/m [18][17]. The thermal stability is:

$$\Delta \sim \frac{\pi^3 A_s t}{4 k_B T} \quad for\, D > 30\, nm, \tag{4.1}$$

where k_B is the Boltzmann constant, T is the temperature and t is the free-layer thickness. Below this value instead the dependency is:

$$\Delta = \frac{K_{eff}^{MTJ} \pi (\frac{D}{2})^2 t}{k_B T} \quad for\, D < 30\, nm, \tag{4.2}$$

where D is the diameter and K_{eff}^{MTJ} is the effective perpendicular magnetic anisotropy energy density of the junction and it is equal to:

$$K_{eff}^{MTJ} = K - \frac{M_s^2}{2\mu_0}(N_z - N_x), \tag{4.3}$$

where M_S is the saturation magnetization, K is the perpendicular magnetic anisotropy energy density, N_Z is the out-of-plane and N_X the in-plane directions of the demagnetization factors. Assuming uniform magnetization along Z direction the N_Z value used is calculated for each MTJ diameter according to $E_d^z = \frac{N_z M_s^2}{2\mu_0}$, where μ_0 is the permeability in free space, N_z is dependent on the shape, or better on thickness and diameter, of the free-layer (that is approximated as a cylindrical shape). The resulting curve is shown in figure 4.1.1.

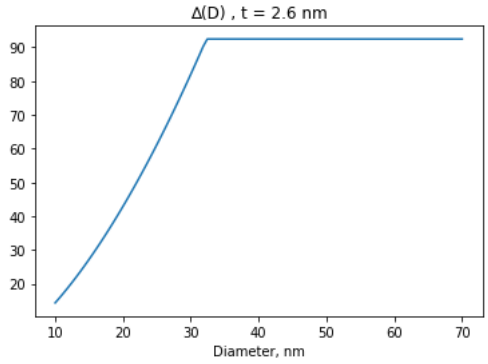

Figure 4.1.1: Thermal stability Δ in function of Diameter size D, t is the thickness of the free-layer (2.6 nm).

Secondly, the Ratio of thermal stability factor Δ to intrinsic critical current I_{C0} was calculated. The average absolute intrinsic critical current is defined as:

$$I_{C0} = \frac{|I_{C0}^{P-AP}| + |I_{C0}^{AP-P}|}{2} \tag{4.4}$$

where I_{C0}^{P-AP} and I_{C0}^{AP-P} are the critical currents at which the junction switches for parallel to antiparallel states and vice-versa. For diameters less than 30 nm it can be calculated as:

$$I_{C0} = 4\alpha \frac{e}{\hbar P} k_B T \Delta \tag{4.5}$$

Above the critical diameter the value it is found by multiplying the critical current density at $D = D_n$ to the actual area corresponding to the different diameters.

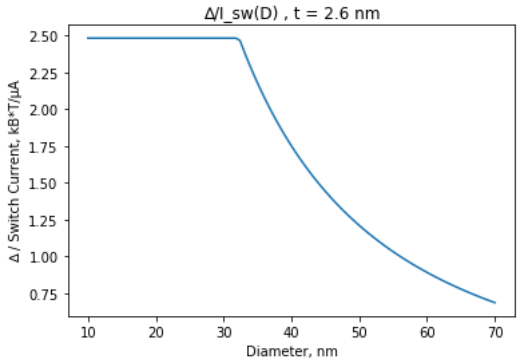

Figure 4.1.2: Ratio of thermal stability factor Δ to intrinsic critical current I_{C0} as a function of Diameter size D.

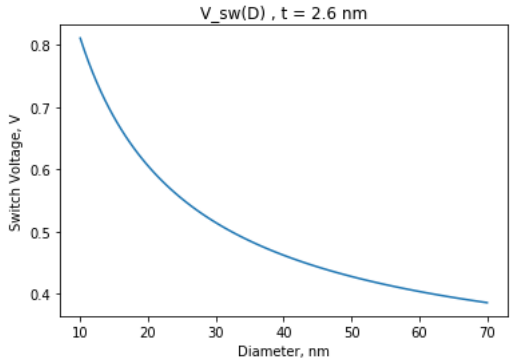

Figure 4.1.3: Switching Voltage as a function of Diameter size D, t is the thickness of the free-layer (2.6 nm).

The switching voltages for both transition can be found by the following expressions:

$$I_{C0}^{P-AP} = I_{C0} \cdot (1 + P^2) \tag{4.6}$$

$$I_{C0}^{AP-P} = -I_{C0} \cdot (1 - P^2) \tag{4.7}$$

$$V^{P-AP} = I_{C0}^{P-AP} \cdot R_P \tag{4.8}$$

$$V^{AP-P} = I_{C0}^{AP-P} \cdot R_{AP} \tag{4.9}$$

Where V^{P-AP} and V^{AP-P} are the switching voltages, and as said already in section 3.5, $V^{P-AP} \approx V^{AP-P}$. After studying the dependency of the main parameters on the diameter pillar size, the analysis on two connected MTJ can be carried out.

4.2 Multiple States distributions

The second step was to evaluate the expected states distributions appearing from the connection of two MTJs. In order to do this, a Gaussian shape dispersion of the resistance values was calculated according to fitting parameters taken from the literature. In particular, to evaluate the standard deviation σ of the resistance state the following procedure was used.

Starting from the RA product, and the TMR of the junctions it was possible to calculate the expected resistance in function of the diameter. In our case $RA = 11\Omega\mu m^2$ and $TMR = 140\%$, this is enough to find what are the resistance values of the individual MTJs, to extract the standard deviation, instead, the following empirical formula taken from the literature [19] was used.

$$FWHM_{(P)} = \frac{R_{AP} - R_P}{14.1} \tag{4.10}$$

$$FWHM_{(AP)} = 2 \cdot FWHM_{(P)} \tag{4.11}$$

From the $FWHM$ it was possible to find σ by dividing by $(2 \cdot \sqrt{2 \cdot \ln 2})$. Once found the resistance states with its corresponding distributions, the resistance of the two junctions in parallel connection was calculated and the standard deviation was propagated as follows, and shown in the topmost graphs of figures 4.2.1, 4.2.2 and 4.2.3.

$$R_{tot} = \frac{R_1 R_2}{R_1 + R_2} \tag{4.12}$$

$$\sigma_{R_{tot}} = \left[(\frac{1}{R_1^2}) \sigma_{R_1} + (\frac{1}{R_2^2}) \sigma_{R_2} \right] \cdot R_{tot}^2 \tag{4.13}$$

The resistance distribution, though, is not the only aspect to be considered when studying a multilevel Cell. Also the possibility to switch the individual pillars independently has to be studied. In order to do so, it is best to consider what is the switching probability of the MTJ in function of the writing pulse, instead of only the switching voltage/current. This because the switching probability gives a better description of the real device behaviour.

In p-STT-MRAM the data retention time τ is related to the thermal stability factor Δ by Arrhenius law [20][21]:

$$\tau = \tau_0 \exp \Delta \tag{4.14}$$

where $\tau_0 \sim 1\,ns$ is the inverse of the attempt frequency. The thermal stability Δ, as already said, is defined as the ratio between the energy barrier separating the two stable states and thermal energy, if a current I and an applied external field H is applied it becomes [22]:

$$\Delta = \frac{\Delta E(1 - \frac{I}{I_C})(1 \pm \frac{H}{H_k})^2}{k_B T} \tag{4.15}$$

where I_C is the extrapolated switching current at τ_0 and H_k the magnetic anisotropy on the switching layer.

And the switching probability P is[23]:

$$P(\tau) = 1 - \exp\left(-\frac{\tau}{\tau_0 \exp \Delta}\right) \tag{4.16}$$

And the Switching Current Density (SCD) is given by [23]:

$$SCD(I) = \frac{\Delta}{I_C \tau} \exp\left(-\frac{t_p}{\tau}\right) \tag{4.17}$$

where I_C is given by :

$$I_C(t_p) = I_{C0}\Big[1 - \frac{1}{\Delta} \ln \frac{t_p}{\tau_0}\Big] \tag{4.18}$$

In the following figures are shown the resistance distributions of a two MTJ connected in parallel with their individual switching voltage densities and switching probabilities curve in function of the switching pulse, (pulse width $10\,ns$).

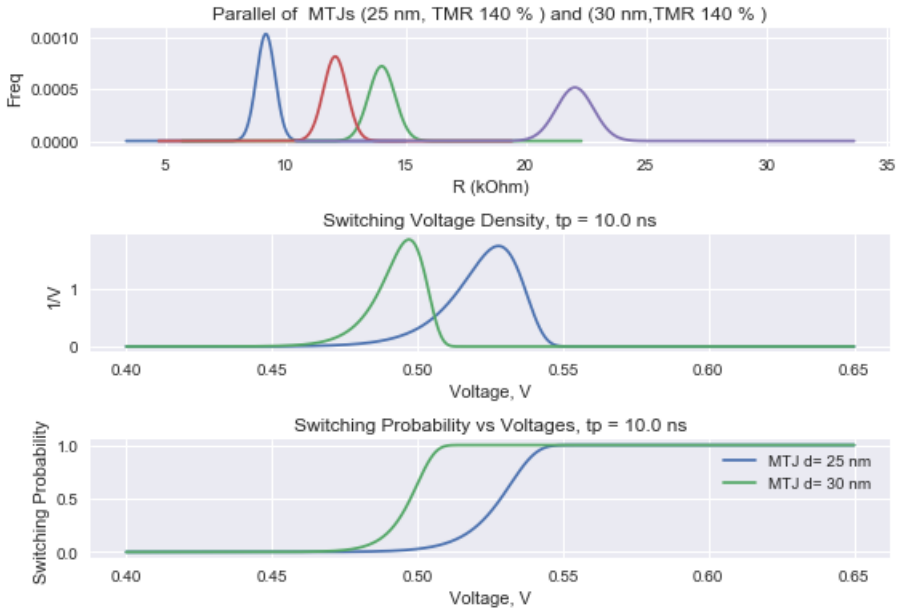

Figure 4.2.1: Top graph: Resistance states distributions of two MTJs connected in parallel; Middle graph: Switching Voltage Density; Bottom Graph: Switching Probability. The diameters of the considered MTJs are: 25 nm and 30 nm, the writing pulse width: 10 ns.

In figure 4.2.1 the two junctions connected have a diameter difference of 5 nm. The parallel resistance state distributions (the four Gaussian curves) appear reasonably separated. There is one problem. In fact, as it is shown in the second and third plot, the switching voltages have similar values, and moreover the switching probability density functions tend to overlap. Hence, even if the states are separated there will be no independent deterministic switching of the two junction, meaning that when one wants to switch only one junction also the other one will switch resulting in only two reachable states. To avoid this problem one can increase the difference in diameter of the two junctions, this will cause also the switching voltages difference of the two junction to increase and so to better control them. Figure 4.2.2 shows the same plots as figure 4.2.1 but with a difference in diameter size of $20\,nm$.

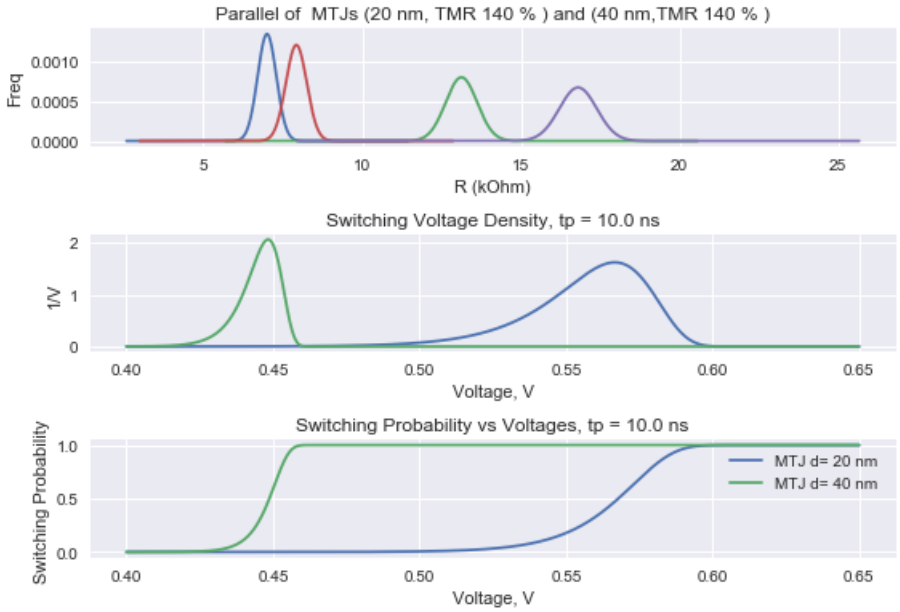

Figure 4.2.2: Top graph: Resistance states distributions of two MTJs connected in parallel; Middle graph: Switching Voltage Density; Bottom Graph: Switching Probability. The diameters of the considered MTJs are: 20 nm and 40 nm, the writing pulse width: 10 ns.

It seams that at $20\,nm$ difference the junction should be different enough to be switched independently one MTJ with respect to the other, (the details on how to achieve this will be explained further on). The issue now is that the two lowest resistance states are overlapping, causing the MRAM cell to have only three distinguishable states. There is, thus, a trade-off to be found between states and switching voltages separation.

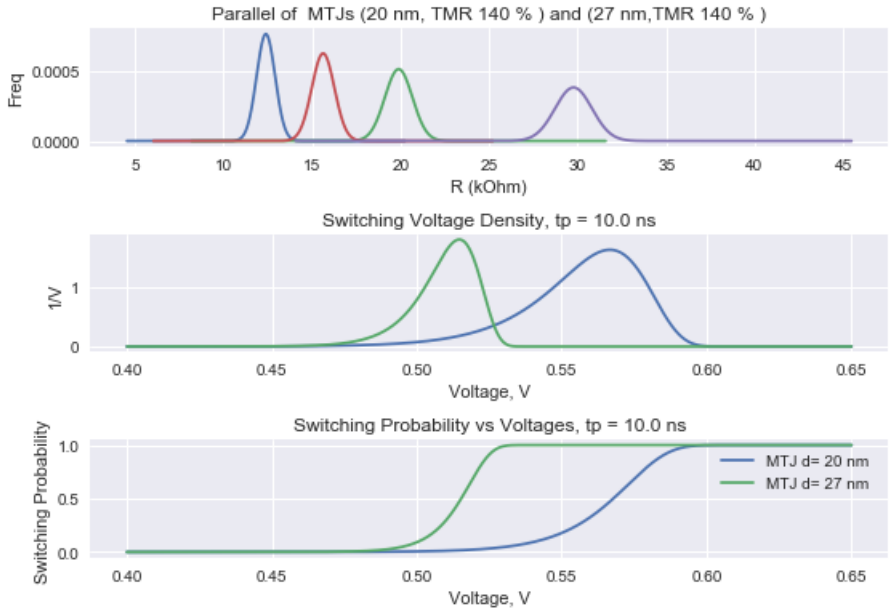

Figure 4.2.3: Top graph: Resistance states distributions of two MTJs connected in parallel; Middle graph: Switching Voltage Density; Bottom Graph: Switching Probability. The diameters of the considered MTJs are: 20 nm and 27 nm, the writing pulse width: 10 ns.

In figure 4.2.3 it is shown the best diameter difference in terms of switching differences and distinguishable states according to our conditions, and it is for a diameter difference of $7\,nm$.

4.3 Series and Parallel comparison

This section is divided in three parts, each of them compare the parallel configuration to the series configuration. At first, the distribution of the total resistance states are compared. Secondly, a power consumption analysis is done and at last, the way the switching has to be controlled for the two configuration is discussed.

4.3.1 Multiple States distributions

The following graphs represent the states distributions, as in the previous section, of a multilevel output MRAM cell created by parallel (in blue) connections between the junctions and the series configuration (in red).

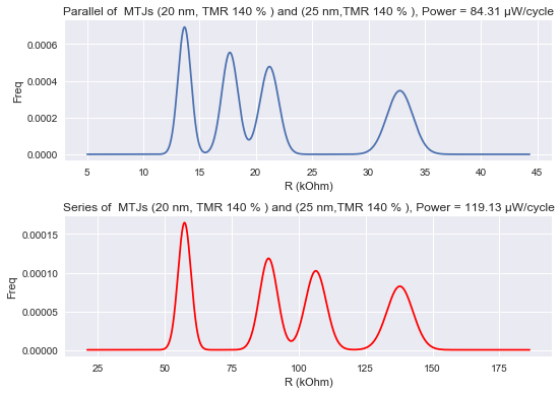

Figure 4.3.1: Resistance states distributions of two MTJs connected in parallel and Series, the diameters of the considered MTJs are: 20 nm and 25 nm.

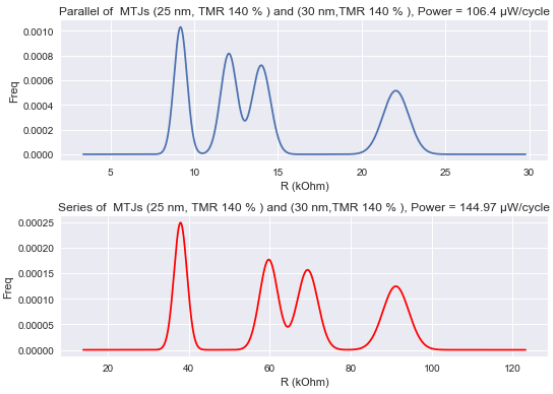

Figure 4.3.2: Resistance states distributions of two MTJs connected in parallel and Series, the diameters of the considered MTJs are: 25 nm and 30 nm.

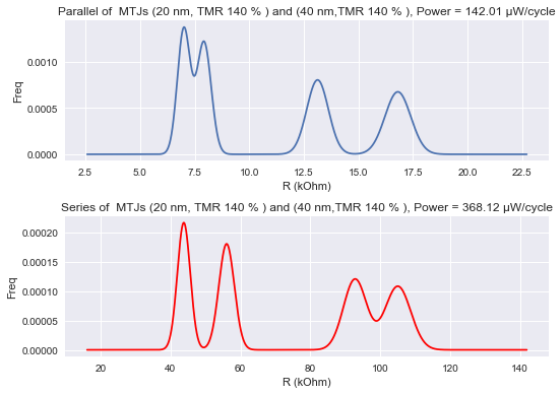

Figure 4.3.3: Resistance states distributions of two MTJs connected in parallel and Series, the diameters of the considered MTJs are: 20 nm and 40 nm.

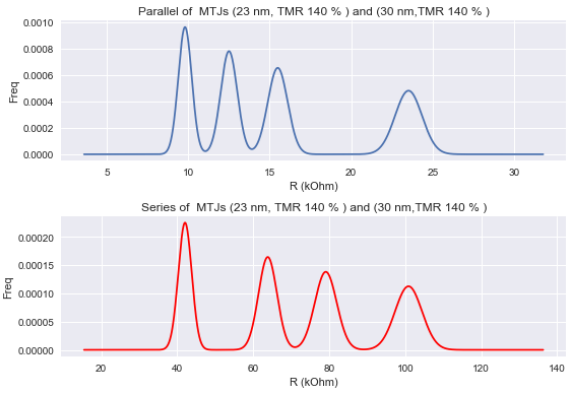

Figure 4.3.4: Resistance states distributions of two MTJs connected in parallel and Series, the diameters of the considered MTJs are: 23 nm and 30 nm.

The distributions show little differences, it is noticeable that the states overlapping when the diameter difference is not optimal are not the same for the two configurations but, nonetheless, it is present for both in a similar way for every corresponding graph. For both the configuration the best result, according to our initial conditions, is at a diameter difference of $7\,nm$. There is a significant difference instead in the values of the resistance, as one would expect, in the parallel configuration the overall resistance is much lower than the series configuration. This causes the two cells to have different characteristics in terms of current and voltage. The parallel configuration will have high current and low voltages and the opposite is true for the series configuration.

4.3.2 Switching

The transitions between the possible states of the whole cell must be taken under examination. In fact, it is not straightforward as one would expect. The individual junctions must switch independently one with respect to the other but this it is not always possible. The notation used is the following: considering two individual MTJs, MTJ_a and MTJ_b, to label the state they are in, the number 0 or 1 is used, 0 for low resistance (P) and 1 for high resistance (AP). With two MTJs connected the whole cell will have four total states:

1. $MTJ_a(P) + MTJ_b(P) = 00$
2. $MTJ_a(P) + MTJ_b(AP) = 01$
3. $MTJ_a(AP) + MTJ_b(P) = 10$
4. $MTJ_a(AP) + MTJ_b(AP) = 11$

figure 4.3.5 shows the corresponding states to the switching distribution, taking as an example one already seen before, with a diameter difference of 7nm.

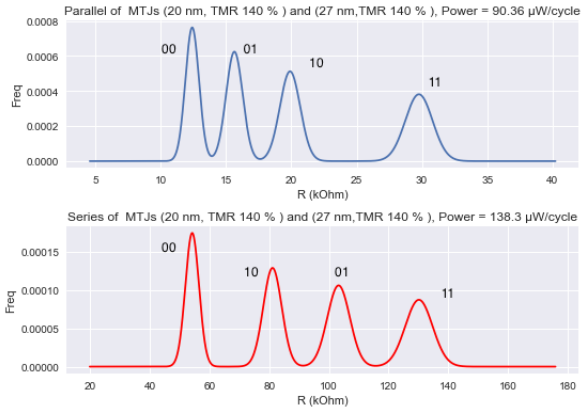

Figure 4.3.5: Resistance states distributions of two MTJs connected in parallel, the diameters of the considered MTJs are: 20 nm and 27 nm. With corresponding labelled states.

Now that every state is labelled, the possible transitions can be studied. In figure 4.3.6 a qualitative representation of the I-V characteristic for the parallel configuration Cell is illustrated. On the abscissa you find the voltage applied to the whole Cell and on the ordinates the total current. The four dashed lines are indicating the four possible resistance states.

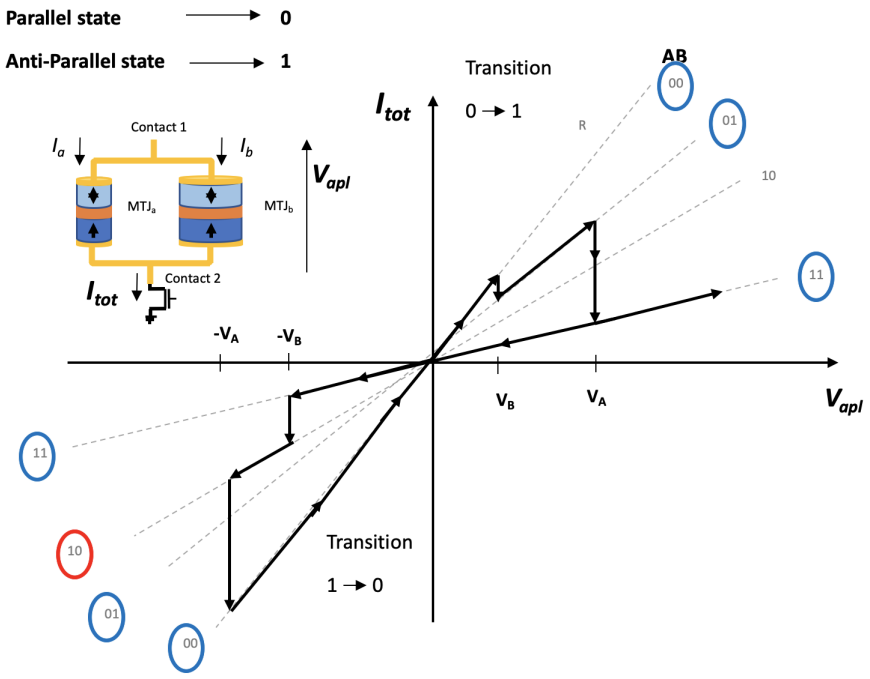

Figure 4.3.6: Current-Voltage characteristic of a Multilevel output cell, parallel configuration of two MTJ.

Starting, for example, from the lowest resistance state "00", increasing the voltage across the junction, will cause the current to increase proportionally to resistance state. This will happen until the voltage will be sufficiently high to be able to switch the MTJ with the lowest switching voltage (MTJ_b in the figure) going from P to AP. Thus, the resulting resistance will be higher,"01" state, and the corresponding dashed line will have shallower slope. If the voltage is increase even more, at some point, also the MTJ_a will switch to AP state, and because MTJ_b was already in the AP state the overall resistance will be the highest "11". So by increasing the voltage in one direction the transitions were only two, and the state "10" was never reached. To reach also the unfortunate state, the only possibility is, from the state "11", to change the polarity of the applied voltage in order to switch MTJ_b back to the parallel state. Finally to turn back to our initial state "00" it is sufficient to increase in modulus the voltage in the same direction switching back to parallel state also MTJ_a.

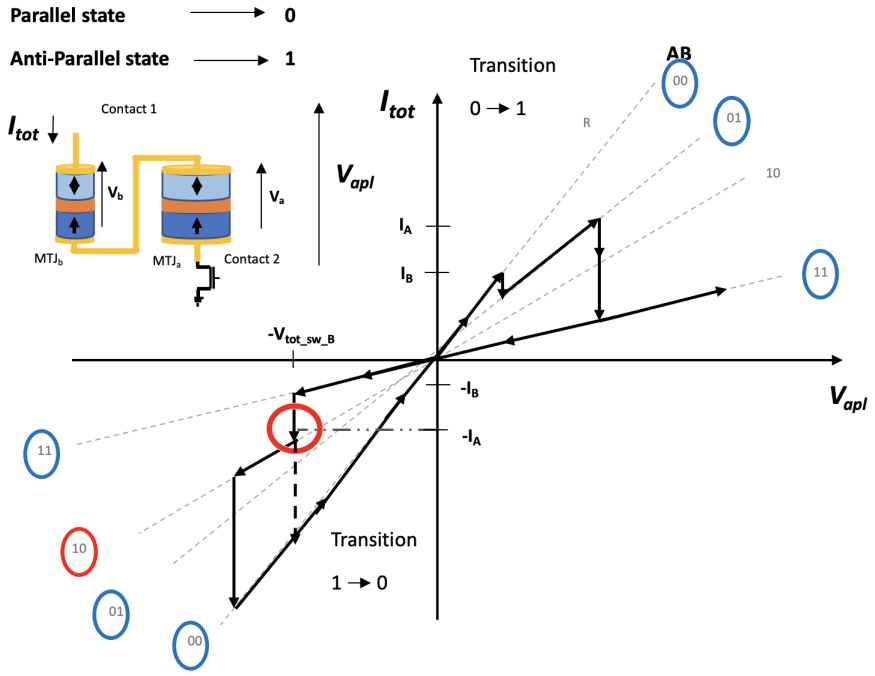

Figure 4.3.7: Current-Voltage characteristic of a Multilevel output cell, series configuration of two MTJ.

For the series configuration Cell, the behaviour is similar. There is only one difference. In this case the best parameter to identify when one MTJ is going to switch is the current. In fact, the voltage across one MTJ in function of the total applied voltage depends not only on its state, but also on the state of the other MTJ, that will increase or decrease the total current. Hence, if the switching control is done again through fixed voltage across the whole junction, the MTJ with lowest switching current will switch first. It seems only a matter of notation but actually the problem of controlling the whole junction through fixed voltage appears when going from a high resistance state to a lower one. When these transitions happen, the switching of one junction to the lower resistance state will increase the total current and, contrary to the parallel case, the total current is shared between both the junctions, hence, this variation may cause also the undesired switching of the second MTJ.

To avoid this problem, the second MTJ must have a switching current larger (in modulus) than the current passing through the whole junction during the transition, or instead, the problem will be avoided if the control of the switching was done fixing

the total current and not the total voltage across the junction.

A similar behaviour will appear in the parallel configuration if the switching was controlled through fixed current.

4.3.3 Power consumption

During this part the focus goes on the power consumption, in particular the power needed across the whole MRAM cell to achieve one or more transitions between the possible states. Considering two junctions connected in parallel or series, the total current or voltage across the cell are related as written in figure 4.3.8 and 4.3.9. Across a single MTJ the relationship between switching voltage and switching current can be expressed as follows:

$$I_{sw,a,j} = \frac{V_{sw,a,j}}{R_{a,j}} \tag{4.19}$$

where a is for the MTJ chosen (MTJ_a) and j can be "0" or "1" depending on the state of the MTJ. The power across the **single MTJ** is then:

$$P_{sw,a,j} = \frac{V_{sw,a}^2}{R_{a,j}} \tag{4.20}$$

The power needed across the whole Multilevel MRAM cell to switch MTJ_a can be calculated as:

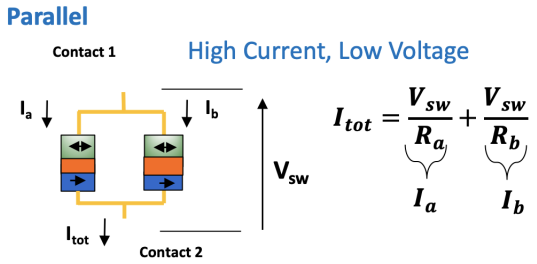

Figure 4.3.8: Illustration of the parallel configuration.

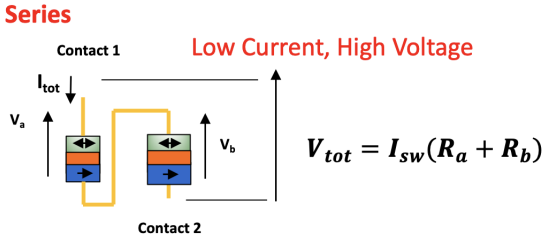

Figure 4.3.9: Illustration of the series configuration.

$$
\begin{aligned}
P_{\parallel} &= V_{sw,a} \cdot I_{tot} \qquad (4.21)\\
&= V_{sw,a} \cdot (I_{sw,a} + \frac{V_{sw,a}}{R_b})\\
&= V_{sw,a} \cdot I_{sw,a}(1 + \frac{R_a}{R_b})\\
P_{\parallel} &= P_{sw,a,0/1} \cdot (1 + \frac{R_{a,0/1}}{R_{b,0/1}}) \qquad (4.22)
\end{aligned}
$$

$$
\begin{aligned}
P_{\perp} &= V_{tot} \cdot I_{sw} \qquad (4.23)\\
&= I_{sw,a} \cdot (V_{sw,a} + I_{sw} \cdot R_b)\\
&= I_{sw,a} \cdot V_{sw,a}(1 + \frac{R_b}{R_a})\\
P_{\perp} &= P_{sw,a,0/1} \cdot (1 + \frac{R_{b,0/1}}{R_{a,0/1}}) \qquad (4.24)
\end{aligned}
$$

By manipulating mathematically the equations describing the power needed to switch MTJ_a, the result is that every transition has its on different consumption. Furthermore, one can see that for both eq. 4.22 and 4.24 there is the power needed to switch the single MTJ multiplied by a parenthesis containing the starting resistance states of the transition. The parenthesis are in a way symmetric one with respect to the other (parallel w.r.t. series).

To explain what is happening let's take one example: during the transition "00"→ "01" both MTJ start in a low resistance state. To switch the device, the whole Cell must be biased. In the parallel configuration, the total power will be the voltage across the Cell multiplied be the total current, that in this case is the maximum possible because both of the pillars are in low resistance. In the series configuration, instead, the switching will happen when the total current is equal to I_{sw}, multiplying this to the sum of the voltages across the two pillars will give the total power, but because they are both in low resistance state, the total voltage drop is the lowest possible and so the power needed is the lowest. The opposite will be true if we

would have taken as an example the transition "11"→ "10". Note that we are not yet comparing the parallel configuration to the series. The lowest and highest power values we are referring to, are with respect to the possible transitions of the same configuration.

Now comparing the two configurations instead, it seems that they have the same behaviour but for different initial conditions. When one consumes more the other consumes less and the other way round, but actually it is not the whole story. For two reasons, the first is that not every transition is possible. It happens that the "possible" transitions are mostly the ones where the series configuration dissipates more. Resulting in the possible transition favoring the parallel configuration. The second is that, the power needed to switch the single MTJ is not the same to go to AP or to P state. To understand better, a numerical example is shown in table 2. It is a cycle that reaches every state showing all the power needed for each transitions. It turns out that the total power is always less in the parallel configuration.

	"00" → "01"	"01" → "11"	"11" → "10"	"10" → "00"	**Total**
Parallel	13.60 μW	31.63 μW	32.66 μW	25.52 μW	103.43 μW
Series	23.15 μW	36.00 μW	55.56 μW	7.74 μW	122.47 μW

Table 2: Power consumption for MRAM cell in parallel and series configuration, MTJs diameters of 23 nm and 30 nm

The average switching power for the parallel configuration is 25.9 μW with respect to 30.6 μW of the series configuration, around $\sim$ 20% less. In the following table the power consumption for the same cycle is reported but for a different multi-state Cell, the difference is in the diameters of the junctions. In this case they are both equal to 20 nm. Junctions with same diameter have the same switching voltages. So it seems pointless to do such simulation but, as described better in the following section, there is a way that such Cell can be used exploiting probabilistic switching.

	"00" → "01"	"01" → "11"	"11" → "10"	"10" → "00"	**Total**
Parallel	10.5 μW	17.9 μW	25.3 μW	17.9 μW	71.79 μW
Series	10.5 μW	43.00 μW	25.3 μW	7.4 μW	86.45 μW

Table 3: Power consumption for MRAM cell in parallel and series configuration, MTJs diameters of 20 nm and 20 nm

Also in this case the total power is lower for the parallel configuration. For the parallel configuration is 17.9 μW with respect to 21.6 μW of the series configuration,

again around $\sim 20\%$ less. One can notice that for transitions that have as initial states both MTJs in the same state, the power is the same for the two configurations, in agreement with eq. 4.22 and 4.24. These conclusions were made considering the power needed for a transition, dependent only on the initial state of the cell, ignoring the power changes due to the switching of the magnetization direction during the switching itself. This condition is an approximation that seems reasonable, due to the fact that the time required for the magnetization to switch is usually much lower than the switching pulse width. Nevertheless, these results have to be proven experimentally.

4.4 Probabilistic switching

Artificial Synapses have the requirement of increasing or decreasing their resistance state during back-propagation (training of the ANN), this can also be called potentiation or depression. A different approach for reaching a multiple output state MRAM cell is then to abandon the idea of deterministic switching and to use a cell composed of MTJs all of the same size, and switch them in a probabilistic way. The same size will cause them to have the same switching voltage, as already said. If the voltage used to make the transition is lower than the switching voltage there will be a lower probability of switching for all the MTJs (eq. 4.16). For example, with a two MTJ multiple MRAM cell, ideally what will happen is that by applying different pulses, having a switching probability of around 50%, after the first pulse only one MTJ will switch and after the second pulse the other one will. Of course this is ideal, in reality the switching probability can be lower and number of pulses higher, but still the principle is the same. The drawback of this method is that, because the junctions are the same, they have the same resistance states, thus the resulting states of the Cell are going to be $N + 1$ instead of 2^N, as shown in figure 4.4.1 for two MTJ.

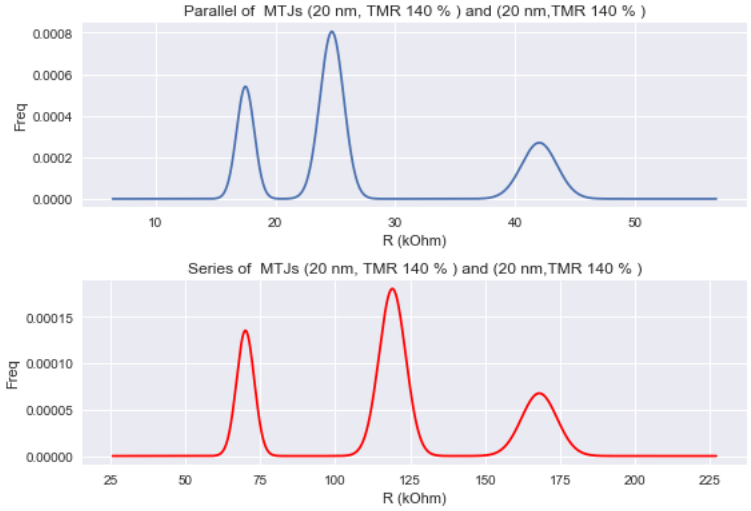

Figure 4.4.1: Resistance distribution for two MTJ connected in parallel and series with diameter of 20 nm.

Figure 4.4.2 and 4.4.3 show a simulation of probabilistic switching. The code used takes into account the switching probability as a function of voltage and all the figures-of-merit of the single MTJs connected.

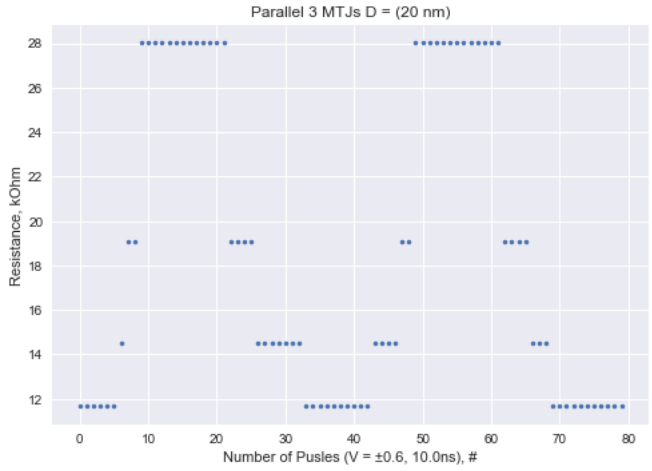

Figure 4.4.2: Simulation of probabilistic switching of 3 MTJs connected in parallel, every dot represents resistance the state after a writing pulse. The sequence of pulses are 20 positive and 20 negative.

Every dot represents the resistance state of the whole Multiple cell after a writing pulse. The sequence of the writing pulses is 20 positive pulse followed by 20 negative pulses. One can see that the transitions do not happen after the same number of pulses, and sometimes more than one pillar switches at the same time.

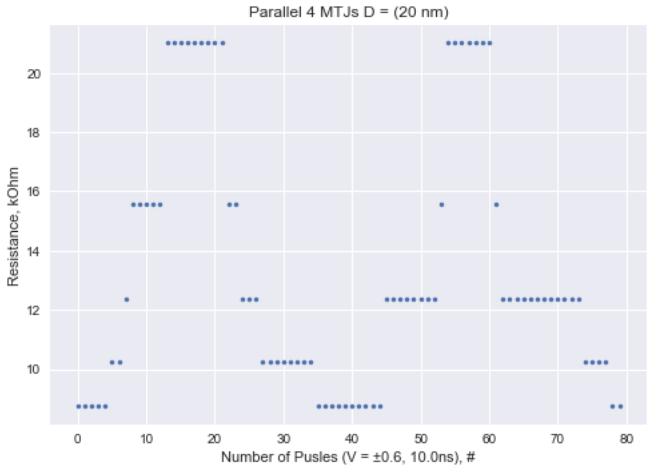

Figure 4.4.3: Simulation of probabilistic switching of 4 MTJs connected in parallel, every dot represents the resistance state after a writing pulse. The sequence of pulses are 20 positive and 20 negative.

For both the graphs the total number of states is $N+1$, as expected. This method is much simpler in terms of writing. Also, because the diameters of the pillars are the same, the design for the fabrication should be easier. The fact that to reach a large number of states you need a much larger number of MTJ with respect to deterministic switching though is a huge drawback. Furthermore, lots of MTJ connected will create an overall large resistance that has to be taken into account.

5 Experimental work

In this chapter the experimental work is described. It is divided into two parts. The first part is relative to the description of the measurement set up (real time measurement) and the developed analysis code. And the second part explains how the MRAM technology can be used as a random bit generator, and shows the results obtained.

5.1 Real time measurement

The most common measurement setup used is the one done in reflection, that is to say that the input signal is sent to one of the two electrodes of the pillar and the output signal is recovered directly to ground. More practically, only one cable is necessary to carry out these measurements. The experimental setup used during this work is different. The output signal is not connected to ground, it is transmitted to a second cable. We speak of assembly in transmission.

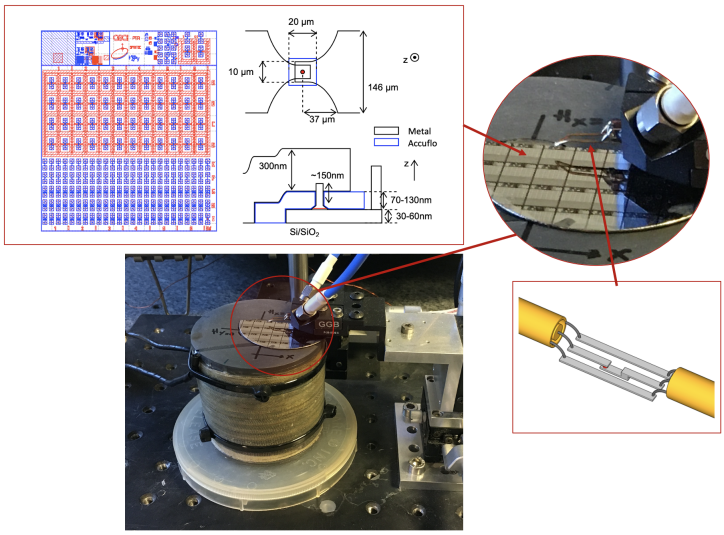

Figure 5.1.1: Picture of the wafer and two probes, used during the measurements. On the top left, the side view and top view of the contact socket are shown together with the complete reticle of one die with 48 devices for transmission testing and 120 single devices. On the bottom right, the illustration of the two probes connected to the junction (in red).

The setup for the measurement in transmission (Figure 5.1.2) includes a B1130A pulse generator, two outputs of which are used, added together by a power divider, in order to be able to generate two distinct voltage levels. This makes it possible to have a reading level (0.05÷0.2 V) while independently varying the write voltage (±0.2 V ÷ ±1.4 V).

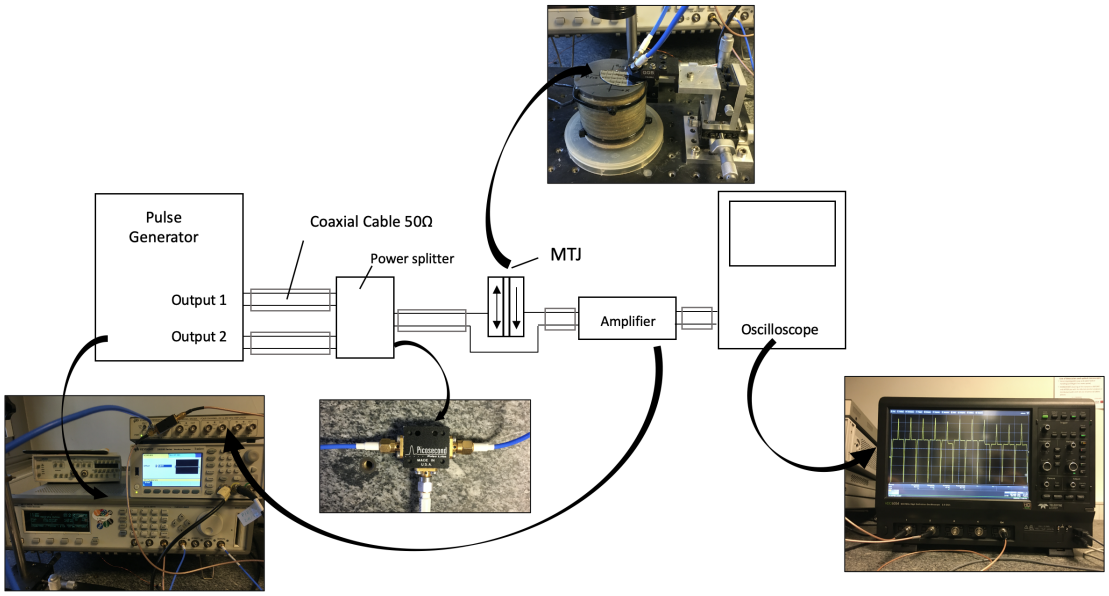

Figure 5.1.2: Real time Measurement setup illustration, including pictures of the LECROY HDO6054 digital oscilloscope, the power divider, the SRS SR445A amplifier and the B1130A pulse generator.

The voltage transmitted through the junction is measured with a LECROY HDO6054 digital oscilloscope after an amplification using an SRS SR445A amplifier. The ground line passes through the metal around the test device. Another type of arrangement is found in the literature where the pulse is split, and the difference between the pulse which has passed through the junction and the pulse that has not been distorted. Also to be mentioned is the possibility to apply a perpendicular magnetic field thanks to a coil on which the sample is placed on. In this 50 Ω transmission arrangement, the junction creates an impedance mismatch. The voltage seen on the

oscilloscope will depend on the resistance level of the junction such as:

$$V_{oscilloscope} = \frac{50\Omega}{R + 100\Omega} 2V_{applied} \tag{5.1}$$

Notice that when the MTJ is in the parallel state the voltage seen is higher than when the MTJ is in the anti-parallel state. The sample used during the work was sample V5529, its stack is: (3)W / (3)Ru / (0.7) Ta / (1.5) Pt / 6x [(0.5) Co / (0.2) Pt] / (0.6) Co / (0.8) Ru / (0.6) Co / 2 [(0.2) Pt / (0.5) Co] / (0.2) Pt / (0.15) Ta / (0.9) Co / (0.25) W / (1.0) $Fe_{53}Co_{17}B_{30}$ / MgO / (1.3) FeCoB / 0.3 W / 0.5 FeCoB / MgO cap / 0.4 Pt / 3 Ta / 7 Ru. Where the numbers in parentheses are nominal thicknesses in nm.

5.2 Data analysis

In this section it is explained how the acquired data is analyzed. The pulse generator sends the following sequence to the MTJ:

1. Positive writing pulse
2. Reading Voltage (usually 0.1V)
3. Negative writing pulse
4. Reading Voltage

This allows the analysis to appreciate what state the MTJ is after each writing pulse. The total number of sequences that can be measured can arrive to 10^9 repetition, for a total of 10^9 positive and negative writing pulses. A qualitative representation of the data acquired by the oscilloscope is shown in figure 5.2.1. The reading voltage is sent at $0.1\,V$, depending if the MTJ is in parallel or in anti-parallel configuration the oscilloscope will measure two different levels.

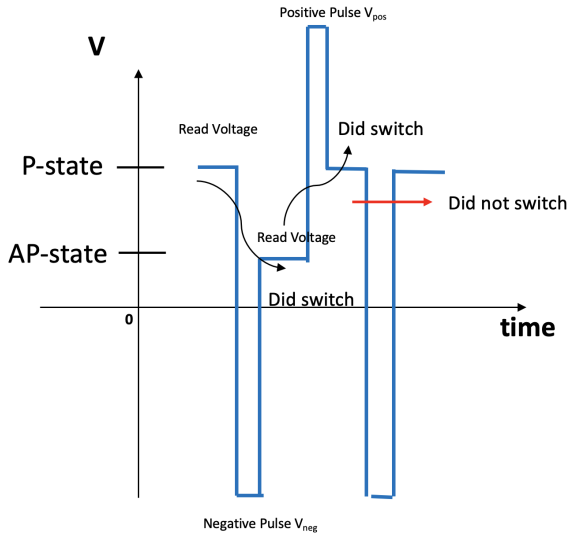

Figure 5.2.1: Qualitative visualization of the writing and reading pulses data acquired by the oscilloscope.

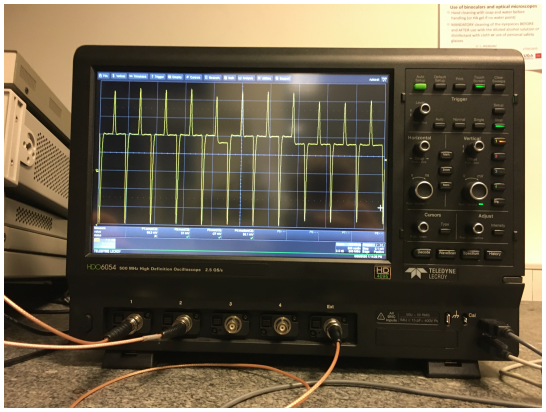

Figure 5.2.2: Picture of the oscilloscope during a trial pulse sequence.

All the instruments used during this thesis have been controlled through GPIB instructions (General Purpose Interface Bus). It is a digital communication standard between devices at short distance (wired link), its official name is IEEE-488. The first part of the internship work, in the laboratory, was the development of a MATLAB code for the real time measurement of the Switching Probability and Writing

Error Rate. The code gives the instructions to the instruments and then analyzes the data measured from the oscilloscope. Because the goal was to analyze a large number of events (or pulse), the MATLAB code had to be extremely efficient in order to perform all the calculations rapidly.

5.2.1 MATLAB Code structure

The code is divided in multiple parts. This section will serve to understand the main principle used to extrapolate the results from the incoming data.

Code structure for the Writing Error Rate and Switching Probability:

- **Initialization**: All the instruments are initialized and the corresponding MATLAB objects are created to allow communication.
- **Oscilloscope Configuration**: A fist small sequence of pulses is sent to the device and the data acquired is used to configure the oscilloscope optimally to the resistance and TMR of the device under test.
- **States detection**: The MTJ is forced, subsequently, in its two states thanks to a strong external magnetic field. The levels of voltages, during the reading part, are saved in order to have a reference, of the two states, for the actual measurement analysis.
- **Measurement**: The pulse generator sends the pulses wanted. The parameters to be chosen are:
  - Number of Pulses, from 10^2 to 10^9;
  - Positive writing pulse voltage, from $300\,mV$ to $1\,V$;
  - Negative writing pulse voltage, from $300\,mV$ to $1\,V$;
  - Writing Pulse time width, above 10 ns;
  - Reading Voltage, from $50\,mV$ to $200\,mV$;
  - Reading time width, above 20 ns;
  - Magnetic field offset, (not yet calibrated);

  The data acquired from the oscilloscope are sent to the computer.
- **Data analysis**: The data is analyzed by extrapolating the average voltage during each reading time after each writing pulses. Every value is then com-

pared to the reference levels in order to understand in what state the junction is. Knowing all the states, before and after the pulses, the code can calculate the pulses that did make the MTJ switch and the pulses that did not. (If the MTJ, before the writing pulse, is already in the state it supposed to be after, the transition is not taken into account.) The Error Rate and Switching Probability can be calculated by dividing the number of times it did not switched (for the Error Rate) or did switched (for the Switching Probability) by the total number of pulses that had the chance of switching (were not already in the state wanted after the pulse). This is done for both the positive and the negative writing pulses.

5.3 Switching Probability and Error Rate maps

In this section some examples of possible measurements that the code can perform are shown. Figure 5.3.1 shows the Writing Error Rate (WER) in function of the amplitude of the writing pulses. The WER is the probability that a failure during writing occurs.

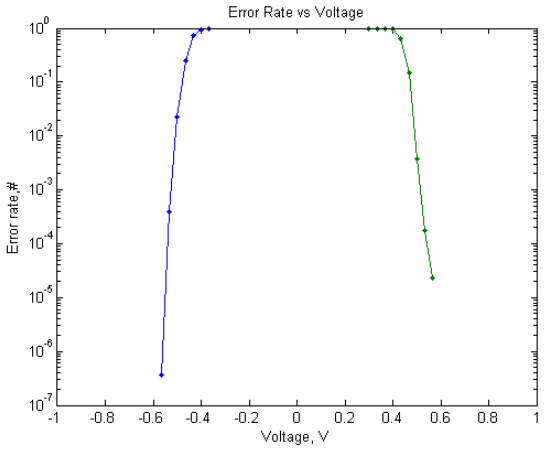

Figure 5.3.1: Error rate in function of writing voltage, the maximum number of pulses measured is 10^7, the negative writing had an Error (unswitched) around $7 \cdot 10^6$, instead the positive pulse never failed. More pulses had be studied to evaluate more precisely its WER.

The time required to perform such measurement depends on the number of points and specially the number of pulses needed. For each couple of positive and negative

writing pulses the measurement plus the analysis time increases with the number of pulses studied. The code stops after one writing pulse failed for each measured point. The timed required for each couple of points are:

- 10^5 pulses in 10 sec;
- 10^6 pulses in 60 sec;
- 10^7 pulses in 10 min;
- 10^8 pulses in 1h 45 min;
- 10^9 pulses in 16h;

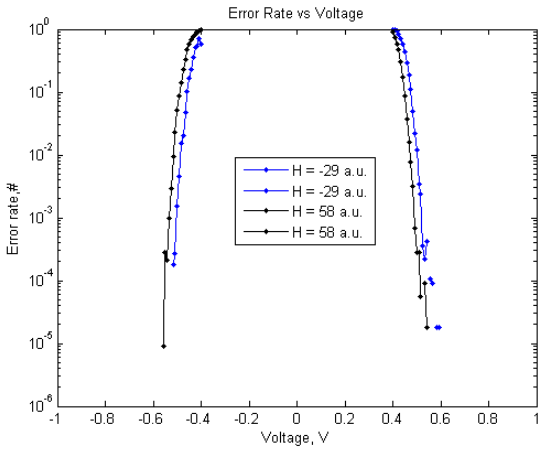

Figure 5.3.2: Error rate in function of writing voltage for two different offset magnetic fields applied

Figure 5.3.2 shows the WER in function of the switching voltage, as before, but this time two different applied external field were applied. This allows the characterization of the device also for different magnetic field offsets. The setup used for the magnetic filed offset was not yet calibrated so the arbitrary units from the MATLAB code were reported. If the number of pulses is lowered, in order to have a faster measurement, it is possible to extrapolate maps of both switching probability and WER in function of switching voltage and magnetic offset. The results are shown below.

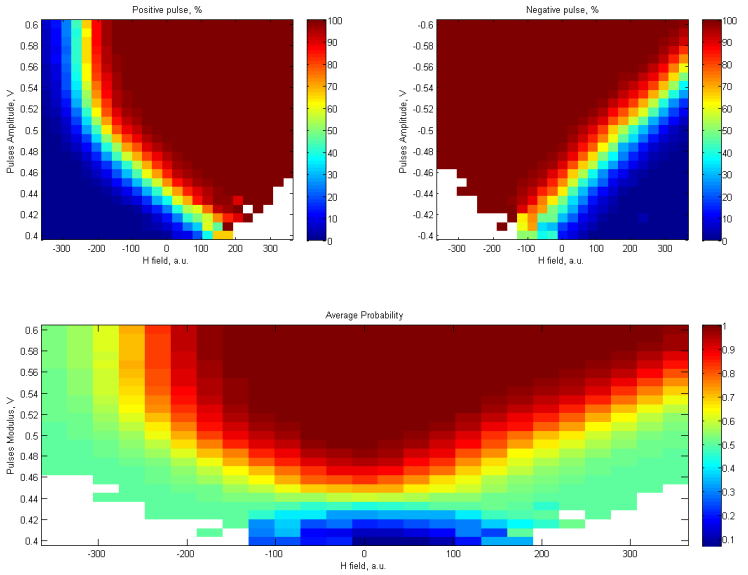

Figure 5.3.3: Switching probability map as a function of writing voltages and magnetic field offset. On the top left and right the results for the positive and negative writing pulses respectively, on the bottom the average of the two probabilities normalized to 1. The bottom graph for switching probabilities lower than 100% could be misleading. For example when there is a 50% probability it could be do to 50% for both positive and negative pulses but also for 60% for positive and 40% for negative pulses. The map is done to appreciate what is the window of 100% switching, that is achievable only if both probabilities are 100%.

The switching probability map shown in figure 5.3.3 was done in a measurement lasted one night (16h). Each probability point was calculated over a maximum of 10^5 pulses. The missing points (in white) are the ones where no switching ever occur, so one of the two transitions never had the chance to switch. This makes the probability given by 0 switching out of 0 transition, thus "Not a Number". This highlights the fact that the two transitions during this type of measurement are not completely independent, meaning that to have a large number of possible transitions both of the probabilities can not be to low, if not the MTJ will be "stock-still" in one state. This can be avoided by analysing the two transitions independently. Varying one while the other is kept at a 100% switching voltage. This will make the measurement much slower though. The goal of this map is actually to an overview

of the switching regions, to understand where correct switching start to happen. For this goal, large number of pulse are not needed. For example, if the negative pulse probability is 1%, starting from 10^5 total pulses, the positive pulse will be calculated over 10^3 pulses, that is sufficiently high for the purpose.

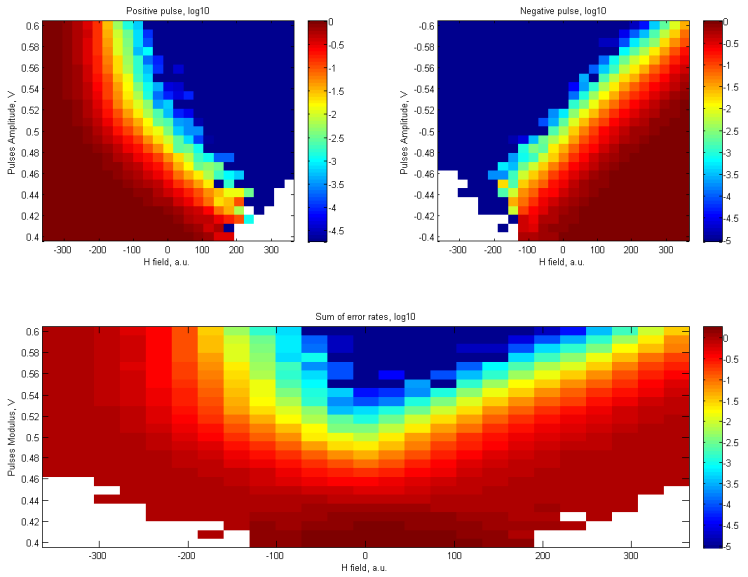

Figure 5.3.4: Logarithm of the Writing Error Rate map as a function of writing voltages and magnetic field offset. On the top left and right the results for the positive and negative writing pulses respectively. The bottom map shows the sum of the positive and negative WER.

The WER map gives a more precise understanding of highest probability region. If one pays attention the highest switching probability region (lowest WER) is now smaller than the 100% region of before, and the part where the transition from low probability and high probability has less resolution than the probability map.

5.4 Random bit generator

As already said in the introduction chapter, many emerging computing schemes, have a critical relationship with random number generation. Usually algorithms that exploit randomness are referred as stochastic algorithms or stochastic computing. Random generators have also a fundamental building block in cryptographic systems and other applications. Random number generators can be classified into two groups: Pseudo-Random Number Generator (PRNG) and Truly Random Number Generators (TRNG).

The real time measurement setup was used to validate the feasibility of using our devices as TRNG. Similar studies have already been done in the past, where a scalable truly RN generator, called "spin dice", was demonstrated by Akio Fukushima et al in 2014 [24]. The power of an MTJ based RNG is its scalability, the sample studied in our work has a nominal diameter size of 80 nm.

To generate random bits, the mechanism used is similar to the writing of data bit in STT-MRAM. The difference is that one writing voltage is set to a lower amplitude, so that the switching probability is exactly 50%. In this way, it is possible to exploit the stochastic nature of STT switching to generate RNs. A schematic of the pulses sent to the device is shown in figure 5.4.1.

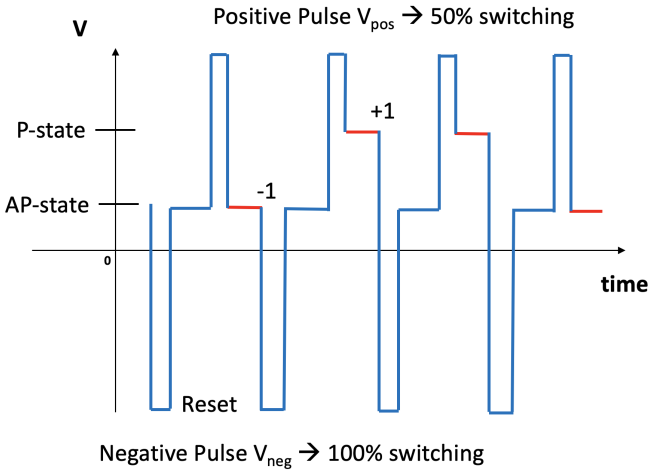

Figure 5.4.1: Schematic representation of the writing pulses used to generate Random bits from a MTJ. It consists in negative pulses with ideally no writing failure (Reset Pulse) and positive pulses with a 50% writing probability. A reading voltage of 100 mV is sent after every pulse. The two states are labelled with +1 or -1 in order to calculate the cumulative sum.

The results obtained are shown in figure 5.4.2 for two different writing voltages (0.491 and 0.492 V). The reset pulses voltage was set at 0.8V that assured no failure in bringing back the device in anti-parallel state. It must be mentioned that all of the experiments were done with no external magnetic field applied, so purely STT switching.

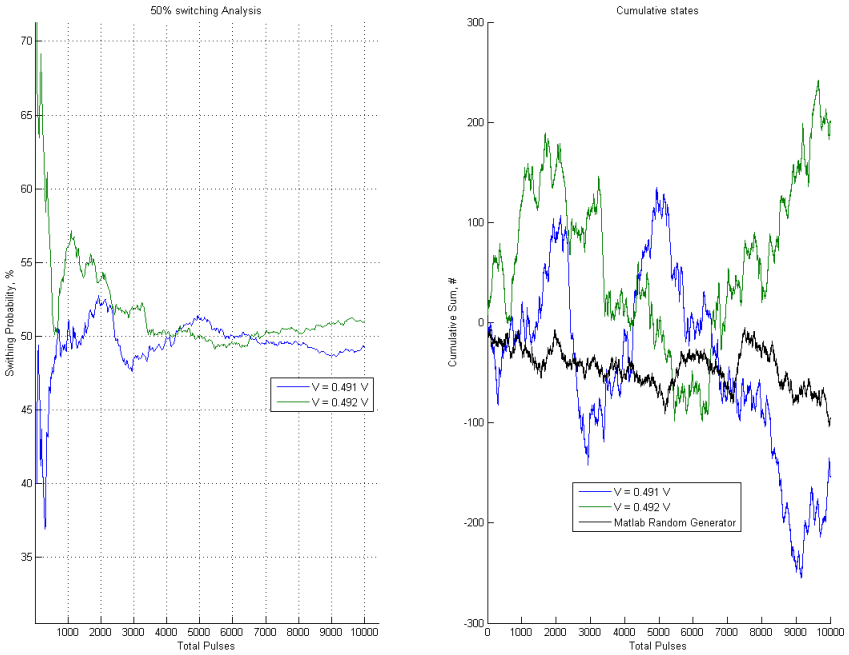

Figure 5.4.2: On the left, Switching Probability in function of the number of pulses for the positive writing pulse at 0.491 V (blue) and 0.492 V (green). On the right, the cumulative sum of the random states measured after the positive pulse, where the P state was labelled as +1 and AP state with -1. In black the cumulative sum calculated from MATLAB's random generator is plotted to compare the results. Ideally the sum should remain around the value 0 if no bias towards on state is present.

When the number of pulses is low, from the switching probability graph one can see a clear initial part where the probability varies considerably. Then it starts to saturate towards the expected value of $\sim 50\%$. It appears though, that the steps of the pulse generator were too large to find the ideal 50% probability value. In fact, at 10^5 pulses, the 0.491 V amplitude gave a probability just under 50% and the pulses at 0.491 V just above it. This could be also due to other mechanisms that will be discussed in the following. On the right graph, it is shown the cumulative sum of the random bits generated, to give a visual comparison with MATLAB's random generator. What is evident, is that the fluctuations of the sum are more marked with respect to the software counterpart. After an initial part, it starts to

be evident the bias towards one state. Indeed, above 8000 pulses the sum of the two measurements start to diverge towards positive (for 0.492 V) and negative (for 0.491 V) values. Further data analysis, of the quality of randomness, such as the statistical test suite NIST SP-800, were not carried out because the experiment done was not reproducible in a satisfactory manner. As shown in the following graphs, in fact, the results coming from different measurements but for the same applied voltages converge to different probability values, giving rise to substantially different cumulative sums.

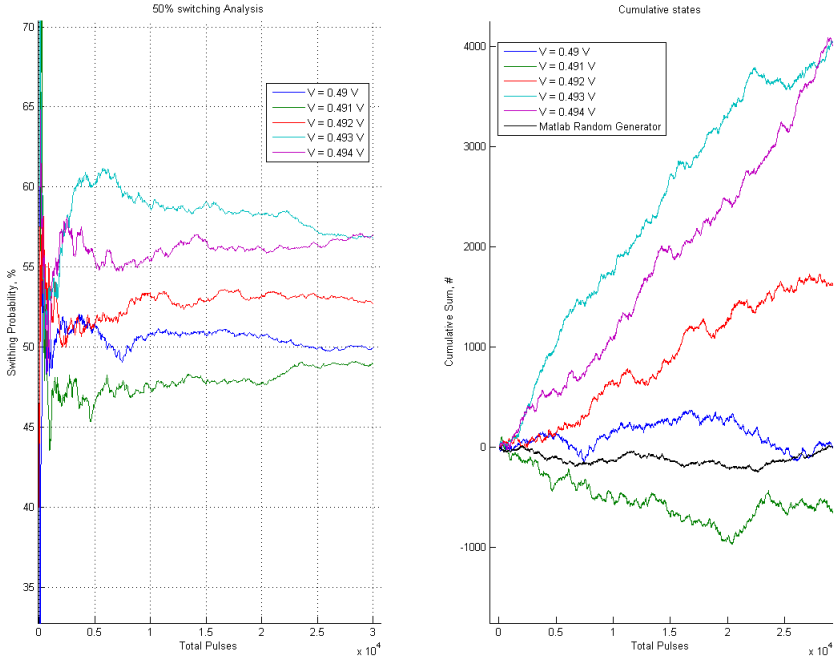

Figure 5.4.3: On the left Switching Probability in function of the number of pulses for the positive writing pulse at 0.490 V (blue), 0.491 V (green),0.492 V (red),0.493 V (magenta),0.494 V (light blue). On the right the cumulative sum of the random states measured after the positive pulse, where the P state was labelled as +1 and AP state with -1. In black the cumulative sum calculated from MATLAB's random generator is plotted to compare the results. Ideally the sum should remain around the value 0 if no bias towards on state is present.

The switching probabilities found for 0.491 V and 0.492 V in figure 5.4.3 are signifi-

cantly different to the ones found in figure 5.4.2. Even though in the last figure the measurement was done on more pulses, it is clear that the probability, for example of 0.492 V, is far from the $\sim 50\%$ found before. It should be also noticed that the lowest probability is not found for the lowest voltage as one would expect. The reason for these differences between the measurements, could be due mainly to two effects. The first is the thermal agitation. In fact, it is known to causes switching voltages to have a distribution. The dependency of the switching probability on the temperature is deeply embedded in the thermal stability factor, variation of temperature will translate in variation in switching voltages. The other effect could be due to electrical noise fluctuations. The MTJ switching, especially in the 50% probability region, is highly sensitive to voltage differences. A slight increase or decrease in applied voltage will drastically change the result.

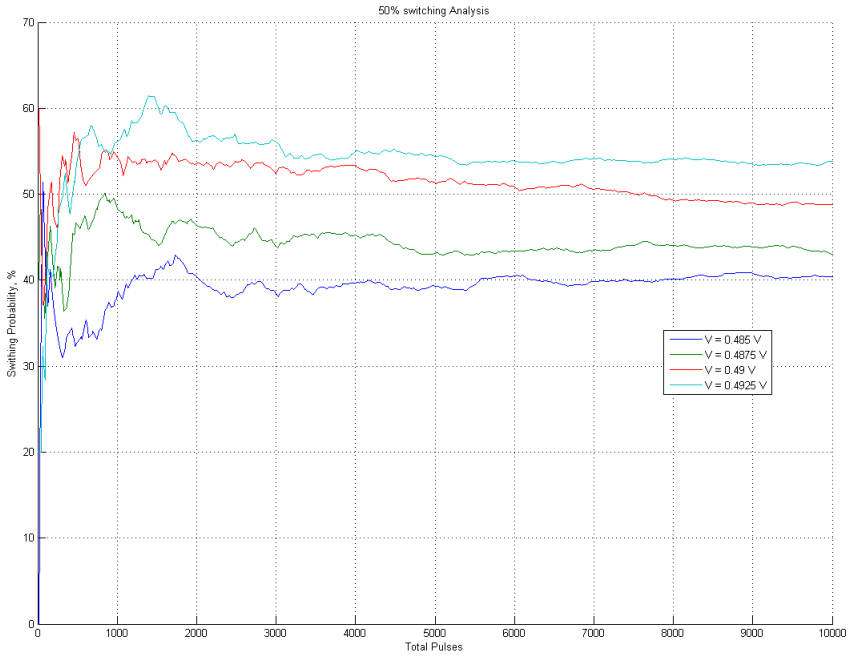

Figure 5.4.4: Switching Probability in function of the number of pulses for the positive writing pulse at 0.485 V (blue), 0.487 V (green),0.490 V (red),0.492 V (light blue).

Figure 5.4.5: Switching Probability in function of the number of pulses for the positive writing pulse at 0.485 V (blue), 0.487 V (green),0.490 V (red),0.492 V (light blue).

The last two graphs were the result of the switching probabilities coming from two different measurements for the same applied voltage pulses. Also here, the final converged probabilities are different. Underlining the impact of the uncontrolled parameters, like temperature or electrical noise, on the final result. Possible solutions for having a lower thermal impact can be found. For example, reduce the writing pulses width. Or a different approach, could be to implement an algorithm similar to a PID controller, that adjusts the switching pulses to have the 50% probability. This last method has to be studied to see if it is feasible.

6 Conclusions

This thesis work has brought new knowledge in the field of MRAM based devices for Artificial Intelligence applications. It has investigated possible multilevel MRAM cells for Artificial Intelligence. It has also presented the real time measurement setup used to evaluate the feasibility of a MTJ based True Random Number Generator, and the code developed for such task.

Considering a Multilevel Output MRAM cell, it has highlighted the importance of finding a correct diameter difference among the MTJ connected, in order to have a good trade-off between distinguishable states and deterministic switching. In the example studied during the simulations, for a two MTJ multilevel MRAM cell, the best result was found at a diameter difference of 7 nm. Furthermore, the work done gave interesting conclusions on the comparison between series and parallel connections in a Multilevel MRAM cell. It has stressed the relevance of the switching control of the cell. During switching, in fact, it is better to write the cell states through fixed Voltage, for the parallel configuration, and fixed Current, for the series configuration. This because when dealing with some transition between resistance states, if the junction is not controlled properly, the switching of one MTJ could lead to the undesired switching of other pillars. A second striking finding is the amount of power needed to switch between states for the two configurations. What appeared from the study, is that the parallel configuration needs, in average, less power to switch between states than the series configuration. The results suggested that the power consumption is around 20% less than the series configuration, for the numerical example considered of TMR = 140% and $RA = 11\Omega\mu m^2$. These conclusions were made considering the power needed for the transitions, dependent only on the initial state of the cell, ignoring the power changes due to the switching of the magnetization direction during the switching itself. This condition is an approximation that seems reasonable. In fact, the time required for the magnetization to switch is usually much lower than the switching pulse width. Nevertheless, this result has to be proven experimentally. Also a different approach consisting in probabilistic switching was studied. It was concluded that, according to the simulations, it is possible to switch stochastically MTJs with the same size connected together. The drawback of the method is the number of states rising, only $N + 1$ instead of 2^N.

Instead, considering the real time measurement setup developed during the work, it enable to have measurements of Switching probability and Writing Error Rates at state-of-the-art level. This allowed to study the feasibility of using MTJ of 80 nm in diameter as Random Bit Generator. The results showed true potential for

the application, as already mentioned in the literature. But for the conditions the experiment was carried out, it was not possible to achieve repeatedly satisfactory results. This was manly due to a lack of control of the 50% probability switching voltage needed for a TRNG. The problems encountered are driven by electrical noise and by thermal fluctuations that make the switching voltage to have a distribution. Possible solutions can be studied, including decreasing the writing pulses width for thermal purpose, or by implementing an algorithm that adjusts the switching pulses similarly to a PID controller.

The work accomplished during this thesis deserves to be continued and completed by a direct experimental approach. The presented results could be then compared more precisely with the experimental ones, in order to achieve performing devices useful for the desired applications.

List of Tables

List of Figures

References

[1] *The Key Definitions Of Artificial Intelligence (AI) That Explain Its Importance*. URL: https://www.forbes.com/sites/bernardmarr/2018/02/14/the-key-definitions-of-artificial-intelligence-ai-that-explain-its-importance/#4cefa9f84f5d.

[2] *Elements of AI*. URL: https://course.elementsofai.com.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[4] Eric D. Nielsen with Francois Chollet Shanqing Cai Stanley Bileschi. *Deep Learning with JavaScript: Neural networks in TensorFlow.js*. https://livebook.manning.com/book/deep-learning-with-javascript/chapter-1/v-1/13. Manning Pubns Co, 2020.

[5] Zhenzhu Meng, Yating Hu, and Christophe Ancey. "Using a Data Driven Approach to Predict Waves Generated by Gravity Driven Mass Flows". In: *Water* 12.2 (Feb. 2020), p. 600. DOI: 10.3390/w12020600. URL: https://doi.org/10.3390%2Fw12020600.

[6] Kenji Suzuki, ed. *Artificial Neural Networks - Architectures and Applications*. InTech, Jan. 2013. DOI: 10.5772/3409. URL: https://doi.org/10.5772%2F3409.

[7] *MRAM info*. URL: https://www.mram-info.com/introduction.

[8] *MRAM info*. URL: https://www.mram-info.com/analysts-expert-mram-revenues-grow-170x-2029-reach-4-billion.

[9] Wang Kang et al. "Reconfigurable Codesign of STT-MRAM Under Process Variations in Deeply Scaled Technology". In: *IEEE Transactions on Electron Devices* 62.6 (June 2015), pp. 1769–1777. DOI: 10.1109/ted.2015.2412960. URL: https://doi.org/10.1109%2Fted.2015.2412960.

[10] J. M. D. Coey. *Magnetism and Magnetic Materials*. Cambridge University Press, 2010. DOI: 10.1017/CBO9780511845000.

[11] Majid Montazer and Tina Harifi. "Nanofinishing: Fundamental principles". In: *Nanofinishing of Textile Materials*. Elsevier, 2018, pp. 19–34. DOI: 10.1016/b978-0-08-101214-7.00002-9. URL: https://doi.org/10.1016%2Fb978-0-08-101214-7.00002-9.

[12] D.A. Petukhov. "Spin-polarized current and tunnel magnetoresistance in heterogeneous single-barrier magnetic tunnel junctions". In: *Physica E: Low-dimensional Systems and Nanostructures* 80 (June 2016), pp. 31–35. DOI: 10.1016/j.physe.2016.01.009. URL: https://doi.org/10.1016%2Fj.physe.2016.01.009.

[13] Bernard Dieny and I. Lucian Prejbeanu. “Magnetic Random&#x02010Access Memory”. In: *Introduction to Magnetic Random&#x02010Access Memory*. John Wiley &#38 Sons, Inc., Nov. 2016, pp. 101–164. DOI: 10.1002/9781119079415.ch5. URL: https://doi.org/10.1002%2F9781119079415.ch5.

[14] T.L. Gilbert. “Classics in Magnetics A Phenomenological Theory of Damping in Ferromagnetic Materials”. In: *IEEE Transactions on Magnetics* 40.6 (Nov. 2004), pp. 3443–3449. DOI: 10.1109/tmag.2004.836740. URL: https://doi.org/10.1109%2Ftmag.2004.836740.

[15] Daniel C. Worledge. “Theory of Spin Torque Switching Current for the Double Magnetic Tunnel Junction”. In: *IEEE Magnetics Letters* 8 (2017), pp. 1–5. DOI: 10.1109/lmag.2017.2707331. URL: https://doi.org/10.1109%2Flmag.2017.2707331.

[16] John Slonczewski. “Spin-Polarized Current and Spin-Transfer Torque in Magnetic Multilayers”. In: *Magnetic Nanostructures in Modern Technology*. Springer Netherlands, pp. 1–35. DOI: 10.1007/978-1-4020-6338-1_1. URL: https://doi.org/10.1007%2F978-1-4020-6338-1_1.

[17] H. Sato et al. “Properties of magnetic tunnel junctions with a MgO/CoFeB/-Ta/CoFeB/MgO recording structure down to junction diameter of 11 nm”. In: *Applied Physics Letters* 105.6 (Aug. 2014), p. 062403. DOI: 10.1063/1.4892924. URL: https://doi.org/10.1063%2F1.4892924.

[18] H. Sato et al. “Junction size effect on switching current and thermal stability in CoFeB/MgO perpendicular magnetic tunnel junctions”. In: *Applied Physics Letters* 99.4 (July 2011), p. 042501. DOI: 10.1063/1.3617429. URL: https://doi.org/10.1063%2F1.3617429.

[19] R.C. Sousa et al. “Magnetic Random Access Memories (MRAM) Beyond Information Storage”. In: *IEEE Symposium on VLSI Circuits (session TMFS.2)* (June 2020).

[20] Xuebing Feng and P. B. Visscher. “Sweep-rate-dependent coercivity simulation of FePt particle arrays”. In: *Journal of Applied Physics* 95.11 (June 2004), pp. 7043–7045. DOI: 10.1063/1.1667808. URL: https://doi.org/10.1063%2F1.1667808.

[21] L. Tillie et al. “Data retention extraction methodology for perpendicular STT-MRAM”. In: *2016 IEEE International Electron Devices Meeting (IEDM)*. IEEE, Dec. 2016. DOI: 10.1109/iedm.2016.7838492. URL: https://doi.org/10.1109%2Fiedm.2016.7838492.

[22] M.P. Sharrock. “Measurement and interpretation of magnetic time effects in recording media”. In: *IEEE Transactions on Magnetics* 35.6 (1999), pp. 4414–

4422. DOI: 10.1109/20.809133. URL: https://doi.org/10.1109%2F20.809133.

[23] Mahendra Pakala et al. "Critical current distribution in spin-transfer-switched magnetic tunnel junctions". In: *Journal of Applied Physics* 98.5 (Sept. 2005), p. 056107. DOI: 10.1063/1.2039997. URL: https://doi.org/10.1063%2F1.2039997.

[24] Akio Fukushima et al. "Spin dice: A scalable truly random number generator based on spintronics". In: *Applied Physics Express* 7.8 (July 2014), p. 083001. DOI: 10.7567/apex.7.083001. URL: https://doi.org/10.7567%2Fapex.7.083001.