

Resistive and spintronic RAMs: device, simulation, and applications

Elena Ioana Vatajelu, Lorena Anghel, Jean-Michel Portal, Marc Bocquet, Guillaume Prenat

▶ **To cite this version:**

Elena Ioana Vatajelu, Lorena Anghel, Jean-Michel Portal, Marc Bocquet, Guillaume Prenat. Resistive and spintronic RAMs: device, simulation, and applications. IOLTS 2018 - IEEE 24th International Symposium on On-Line Testing And Robust System Design, Jul 2018, Platja d'Aro, Spain. pp.109-114, 10.1109/IOLTS.2018.8474226 . hal-01976583

HAL Id: hal-01976583
https://hal.science/hal-01976583

Submitted on 8 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Resistive and Spintronic RAMs: Device, Simulation, and Applications

Elena Ioana Vatajelu, Lorena Anghel
TIMA Laboratory
Université Grenoble Alpes - CNRS
Grenoble, France

Jean Michel Portal, Marc Bocquet
Aix-Marseille Université
Marseille, France

Guillaume Prenat
SPINTEC
CEA-CNRS
Grenoble, France

***Abstract*— The emergence of non-volatile random access memory technologies, such as resistive and spintronic RAMs are triggering intense interdisciplinary activity. These technologies have the potential of providing many benefits, such as energy efficiency, high integration density, CMOS-compatibility, re-configurability, non-volatility and open the path towards novel computational structures and approaches, for the traditional Von-Neumann architectures and beyond. These promising characteristics, coupled with the ever-increasing limitations faced by traditional CMOS-based storage and computational structures, have driven the research community towards completely revisiting the existing computing and storage paradigms, now focusing on providing hardware solutions for in-memory and neuromorphic computing. This has resulted in an intensified research activity in the device physics, striving to achieve circuit-worth devices, reliable compact models and novel architectures. The purpose of this paper is to provide a comprehensive overview of the device physics, issues related to its use in electronic circuits, methodologies for their compact modelling and simulations, and their integration in storage and computational structures.**

Keywords*— *reliability, non-volatile RAM, OxRAM, STT-MRAM, in-memory computing, neuromorphic computing,

I. INTRODUCTION

With technology scaling, the existing memories become increasingly power hungry, less reliable and their fabrication becomes more expensive due to increased manufacturing complexity. Although current memory technologies boast smaller feature sizes, their performance is not proportionally improved from generation to generation. The crossing point, the so called "memory wall" – where technology scaling will lose its profitability – is going to be reached soon due to the growing disparity of speed between Central Processing Unit (CPU) and memory outside the CPU [1]. An important reason for this disparity is the limited communication bandwidth beyond chip boundaries. This opens the path and forces the search for brand new generations of memories and computing paradigms. There are several emergent memory technologies that attempt to address the aforementioned technical challenges and constraints. The major focus is today on novel non-volatile technologies.

The International Technology Roadmap of Semiconductors (ITRS) in its 2015 report identified the Spin Transfer Torque MRAM (STT-MRAM), and Redox RAM (ReRAM) as emerging memory technologies recommended for accelerated research and development leading to scaling and commercialization of non-volatile resistive memories [2]. These emerging devices have many advantages such as: CMOS process compatibility, low fabrication cost, zero static power, nanosecond switching speed, great scalability, and non-volatile nature. In addition, these emerging memories favor increasing system complexity and performance, opening the scientific community to new applications and computation paradigms which had been unfeasible a few years back due to technological limitations, such as in-memory computing and bio-inspired computing. The use of resistive and magneto-resistive technologies is not limited to memories; they can also be used for combinational logic design [3], to integrate non-volatility in latches and Flip-Flops [4], for ultra-low-power normally-off/instant-on computing [5-6].

Among the recent applications of these novel devices, there is a growing interest towards hardware implemented neural algorithms for pattern recognition and classification. Top projects in neuromorphic engineering have led to powerful brain-inspired chips able to simulate numerous neurons to investigate a new kind of computer architecture (SyNAPSE [7], TrueNorth [8]), or to help neuroscientists through the Human Brain Project (SpiNNaker [9]). Overall, considering the large number of neurons needed to perform efficient classification, designers face the same obstacles: storage of at least hundreds of thousands of parameters (synaptic weights) and access to these parameters in order to analyze the input flow of data, sometimes with real-time constraints. We can then distinguish some promising approaches to face the problem and explore them: at the memory level, use of nonvolatile technologies to get parallel access to the memory hierarchy, at circuit level, investigate non-Von Neumann processing that brings computation and memory together, and take advantage of the memristive-like devices.

In-memory computing is an emerging concept based on the tight integration of traditionally separated memory elements and combinational circuitry, that allows minimizing the time and the energy needed to move data across the processor. In this paradigm, the physical properties of novel memory devices are used for both storing and processing information [10-11]. This paradigm leads to an efficient implementation of different arithmetic logic functions, such as bit wise operations where two or more memory rows storing bit-vectors are activated simultaneously. Thanks to in-array calculation, the result does not need to travel the memory bus anymore. In-memory computation of simple OR and AND operations have been proposed already in STT-MRAM [12], others exploit analog characteristics of non-volatile memories to support addition and multiplication inside a crossbar memory, by paying the cost of the computation approximation.

Despite the promising nature of the in-memory computing based architectures and neuromorphic architectures build with emerging devices, many issues related to the devices themselves and to their double use (storage and computing unit) have still to be solved. From the device perspective, the modeling and characterization of fabrication defects,

variability and reliability issues, fault analysis and modeling are still lacking maturity, as consequence of the novel fabrication process and variety of design proposals [13-14]. These issues explain the limitations of the existing reliability enhancement strategies for in-memory computing and neuromorphic structures, thus motivating our research in this direction.

The remainder of this paper is organized as follows. Section II provides the fundamental physics, compact modeling and main variability sources affecting the OxRAM device. Section III provides the fundamental physics, compact modeling and main variability sources affecting the STT-MRAM device. In Section IV, storage and computing structures based on novel non-volatile RAMs are presented, together with issues related to their robustness. Section V concludes this paper.

II. OxRAM: from Physical Device to Circuit Simulation

OxRAM technology is seen as a promising candidate for Storage Class Memory (SCM) applications or for embedded applications, including disruptive architecture such as neural networks. This trend is mainly due to enhanced performances versus classical Flash and to a straightforward integration in the back-end of line (BEOL) of classical CMOS process. Indeed, the OxRAM stack is processed with materials already present in the CMOS technology, such as metal oxide. Among the large panel of possible stacks, the one-presented Fig. 1, based on a 5 nm thick HfO_2 resistive switching layer sandwiched between a TiN/Ti Top Electrode (TE) and a TiN Bottom Electrode (BE) has focused a large attention, since HfO_2 is already used in the MOS HKMG (High k Metal Gate) stack. The resistive switching layers are deposited by Atomic Layer Deposition (ALD), whereas the metallic electrodes are deposited by Physical Vapor Deposition (PVD) [15-17].

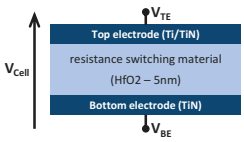


Fig.1 Schematic representation of RRAM stack based on HfO_2.

OxRAM modeling used in this study is based on the work presented in [18]. This approach relies on electric field-induced creation/destruction of oxygen vacancies within the switching layer. The model enables continuous accounting for SET and RESET operations into a single master equation (equation 1) in which the resistance is controlled by the radius of a conductive filament (namely r_{CF}):

$$\frac{dr_{CF}}{dt} = \frac{r_{CFmax} - r_{CF}}{\tau_{SET}} \cdot e^{\left(-\frac{E_{aSET} - \alpha_{SET} \cdot q \cdot V_{Cell}}{k_B T}\right)} - \frac{r_{CF}}{\tau_{RESET}} \cdot e^{\left(-\frac{E_{aRESET} + \alpha_{RESET} \cdot q \cdot V_{Cell}}{k_B T}\right)} \quad (1)$$

where V_{Cell} is the voltage applied between the top and the bottom electrodes, q is the elementary charge of electron, k_b is the Boltzmann constant, T is the temperature in the structure. The parameters controlling the SET (resp. RESET) operation are: τ_{SET} (τ_{RESET}) the nominal rate, Ea_{SET} (Ea_{RESET}) the activation energy, α_{SET} (α_{RESET}) barrier-lowering coefficient.

The FORMING operation is also perfectly taken into account thanks to a second state variable (r_{CFmax}) determined with equation 2:

$$\frac{dr_{CFmax}}{dt} = \frac{r_{work} - r_{CFmax}}{\tau_{FORM}} \cdot e^{\left(-\frac{E_{aFORM} - \alpha_{FORM} \cdot q \cdot V_{Cell}}{k_B T}\right)} \quad (2)$$

where Ea_{FORM} is the activation energy for Electroforming and τ_{FORM} the nominal forming rate.

After model card extraction, the model matches quasi-static and dynamic experimental data measured on HfO_2-based memory elements. To study reliability issues at circuit level, the device to device (D2D) spread and the cycle to cycle (C2C) variability of OxRAM cells are included in the model.

A. Device-to-device spread

To account for the D2D spread, one parameter (σ_{D2D}) is added to the model card according to equation 3. This parameter modifies the barrier-lowering coefficients, thus maintaining the physical consistency between FORMING, SET and RESET operations.

$$\begin{aligned} \alpha_{SET_D2D} &= \alpha_{SET} \cdot (1 + \sigma_{D2D}) \\ \alpha_{RESET_D2D} &= \alpha_{RESET} \cdot (1 - \sigma_{D2D}) \\ \alpha_{FORM_D2D} &= \alpha_{FORM} \cdot (1 + \sigma_{D2D}) \end{aligned} \quad (3)$$

Depending on the setting of the parameter σ_{D2D}, corner-cases, as well as Monte-Carlo simulations, are possible. Corner cases depict the two extreme OxRAM behaviors observed experimentally. The first corner promotes the SET mechanism and degrades the RESET, whereas the second one is the opposite [19]. Between these two extreme cases, Monte-Carlo simulations can be enabled with random sampling of the parameter σ_{D2D}. The Fig. 2 shows quasi-static I-V characteristics of the OxRAM device to depict the D2D variability. It is worth to note that our model is in good agreement with measurement for nominal case. Moreover, the corners definition ensures to fully capture the variability range for all modes of operations (FORMING, SET and RESET).

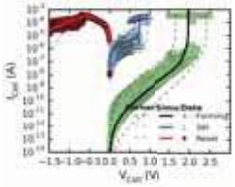


Fig.2: Experimental I(V) characteristics for Electroforming, Set, and Reset measured on a large number of memory elements reflecting the device-to-device variability presented in [15]

The Fig. 3.a describes the dependence of the switching time to the voltage amplitude of the programming pulse. Fast-programming operations (≈ns) can be performed using middle voltages (1.5 V) with respect to standard CMOS biasing for advanced node (1V for 28 nm). Furthermore, the Fig. 3.b. underlines the relation between the R_{HRS} value and the RESET voltage. Our model captures well both dependencies (programming time versus programming voltage and R_{HRS} value versus RESET voltage) for nominal and corner cases.

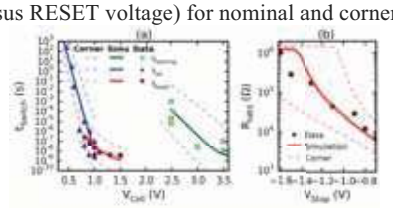


Fig. 3: Experimental (a) switching time for Forming, Set and Reset operations as a function of voltage and (b) R_{HRS} as a function of stop voltage during Reset operation presented in [15] and the simulation results

B. *Cycle-to-cycle variability*

To take into account C2C variability, detection of SET/RESET events during simulation has to be provided to the model. To achieve this task, the finite state machine described in Fig. 4 is included in the model.

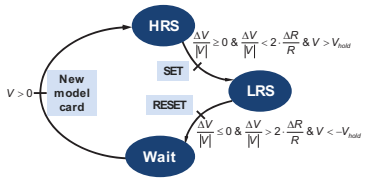


Fig. 4: Finite state machine for detection of SET/RESET cycles

At each new SET/RESET cycle, a random variable x is re-evaluated according to a normal law. This variable is used to process new model card parameters according to equation 4.

$$E_{aSET_C2C} = E_{aSET} \cdot (1 + x \cdot \sigma_{Ea})$$
$$E_{aRESET_C2C} = E_{aRESET} \cdot (1 - x \cdot \sigma_{Ea}) \quad (4)$$

σ_{Ea} determines the range of variation for SET/RESET activation energy. The Fig. 5 presents the cumulative distribution of SET/RESET voltages during 500 cycles long simulation. Depending on the parameter σ_{Ea}, C2C variability may face strong variation. This model feature is very useful to take into account aging effects. Indeed, increasing σ_{Ea} parameter during cycling simulation reflects nicely aging effect and thus reliability issues.

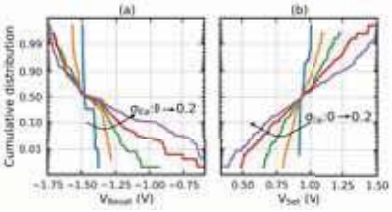


Fig. 5: Cumulative distribution of (a) RESET and (b) SET voltages during 500 cycles simulation with respect to the σ_{EA} parameter

III. STT-MRAM: FROM PHYSICAL DEVICE TO CIRCUIT SIMULATION

The Magnetic Random Access Memory combines an intrinsic non-volatility with fast access time (few ns), low-power consumption (100fJ at bit cell level), high density (~20F²), high endurance and a natural immunity to radiations (since the information is not held by a charge). Thanks to these characteristics, MRAM can be introduced at all the levels of the memory hierarchy, from main memory to latches or registers, or even mixed by the logic itself in the combinational parts of the circuits [20]. To evaluate the benefits that can be expected from these hybrid CMOS/magnetic circuits, it is necessary to develop a full design flow, from device to system level, compatible with the standard design flows of microelectronics. Here, we introduce compact modeling of the magnetic devices for electrical simulations.

A. *Magnetic Tunnel Junctions description*

MRAM technologies rely on Magnetic Tunnel Junctions (MTJs, Fig. 6), nanostructures basically composed of two ferromagnetic (FM) layers separated by a thin insulator. The magnetization of one of the FM layers is pinned and acts as a reference (RL), while the magnetization of the other layer (Storage Layer, SL) can be switched between two stable states, Parallel (P) or Anti-Parallel (AP) to the RL, with a hysteretic behavior. The logic information is coded by the resistance of the stack, which depends on the relative orientations of the two layers, smaller for P than for AP state. The TMR (Tunnel Magneto Resistance) ratio gives the relative variation of resistance between the two states. It is typically around 150% to 200%. Reading the information consists in measuring the resistance of the device.

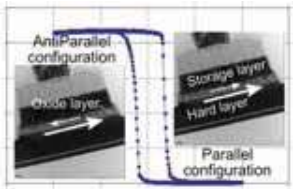


Fig 6: In-plane magnetized Magnetic Tunnel Junction. The resistance is either high or low depending on the relative orientation of the magnetizations

Writing the information is performed by switching the magnetization of the SL. Today, MRAM writing relies on the Spin Transfer Torque (STT) effect [21-22]. It consists in writing the magnetization by applying a current directly through the MTJ. The current gets spin polarized by the RL and transfers the magnetic moment to the SL, resulting in magnetization switching if the current is strong enough. This approach allows a good density (with a bit cell composed of one MTJ and one selection transistor). It is very scalable, since the writing current decreases with the surface of the device. In this technology, there is a strong correlation between the stability of the MTJ information (i.e. the retention time) and the writing current, from one side, and between the writing current and the writing speed, on the other side. This is a degree of freedom that can be used to adapt the device to the application. The STT technology and in particular its "perpendicular" implementation [23] (where the magnetization stands perpendicular to the plane of the magnetic layers) is currently seen as the most promising for logic applications and is studied by more and more academic and industrial actors of microelectronics.

B. *Compact modeling*

The operation of electrical simulator imposes a particular formalism of the devices models, which should be written under the form of an electrical equivalent model. Many compact models of MTJs have been proposed, with two main approaches for modeling the write operation: a more "physical" approach, which takes into account the dynamics of the magnetization driving the switching speed, and a more "behavioral" approach in which the typical switching duration is calculated as a function of the value of the writing current. The first approach is more accurate and predictive since it really provides the real-time state of the magnetization. However, it is relatively slow and not adapted to the simulation of complex circuits like memory arrays for instance. The second one is faster but needs to be calibrated to fit the actual behavior of a device, and does not give the "analog" behavior of the MTJ for accurate characterization of the circuit. Here, we present two examples of models developed at Spintec, following these two approaches. Concerning the read operation modeling (resistance of the MTJ), both models take into account the dependence of the resistance upon the magnetic state and the polarization voltage, the dependence of

the transport and magnetic parameters upon the temperature and the heating of the MTJ due to Joule effect. In both cases, the model is "BSIM-like", with a generic model of the MTJ associated with a corner file to provide the parameters and their variations for a given technology.

Physical compact model. In this model [24], the writing is modeled using the Landau–Lifshitz–Gilbert [21] equation giving the dynamic evolution of the magnetization of the SL due to STT currents applied to the MTJ, with a precessional behavior. Fig. 7 shows simulation results using the model. The first curve represents the pulse of voltage applied to the MTJ for writing. The second curve represents the resulting current flowing through it. We can see that for a constant pulse of voltage, we have a straight change of current, which is due to the resistance switching. The third curve represents the temperature of the MTJ. We see that when applying the writing current, the temperature increases. The dynamics of the temperature changes at the switching point because of the change of resistance resulting in a change in the power brought to the system by Joule effect. The fourth curve represents the resistance of the MTJ. We see that before and after the switching, the resistance is not constant because the tunnel resistance depends on the temperature. Moreover, the TMR depends on the polarization voltage, so the final AP resistance is only reached after the writing pulse is stopped and cooling of the MTJ. In the inset, we can see the typical damped precessions of the magnetization before and after the switching. They are in the GHz range of frequency and drives the speed of the writing operation.

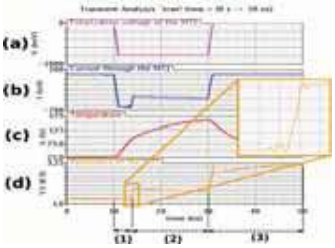


Fig 7: simulation results using the dynamic compact model of the MTJ

Behavioral compact model [25]. the magnetization dynamics is not modeled, resulting in much faster simulations. The typical duration of the switching is calculated from the parameters of the devices and the value of the writing current. Indeed, the switching speed is directly related to the value of the current, following the Sun's and Neel-Brown's models. Fig 8 shows simulation results obtained with both kinds of models, calibrated to give similar results. We see that in the behavioral model, the switching is "binary", without intermediate states of the magnetization (precessions in particular).

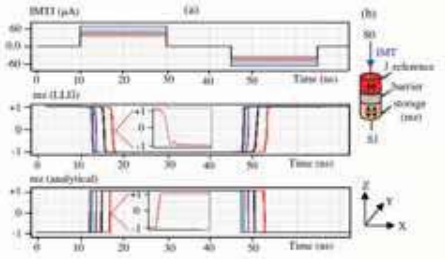


Fig 8: comparison between the compact models of the MTJ

C. *Reliability*

Although the STT writing scheme is widely investigated by major semiconductors companies (Qualcomm, TSMC, IBM, NEC, Everspin/Global Foundries, Toshiba, Samsung), it still suffers from limitations, especially in terms of reliability. Due to the relative height of the MTJ stack compared to its lateral dimensions, the etching process is relatively tricky and results in process variations in the size and shape of the device, affecting the resistance and the writing current. It can result in read and write failures. Moreover, since the writing and reading paths are the same, accidental writing can occur if the reading current is too large. It is generally referred as read disturb. When designing circuits embedding STT MTJs, it is necessary to keep a good ratio between the reading and writing voltages as well as between the writing and breakdown voltages (to avoid accelerated aging). These effects can be easily taken into account in a compact model, because they can all be linked to shape of the device, and so incorporated in the standard process variations in the corner file. A specificity of the STT technology is the effect of the thermal noise, which affects the magnetization and results in a stochastic switching occurrence over time. Since this stochasticity is related to the time, it cannot be directly taken into account in the form of process variations. In the "dynamic" compact model of Spintec, we have proposed a solution based on the use of a noise module affecting the magnetization. Using transient noise simulations in Cadence, we could tune the parameters of the noise and find a good agreement between the simulation and characterization results (Fig. 9).

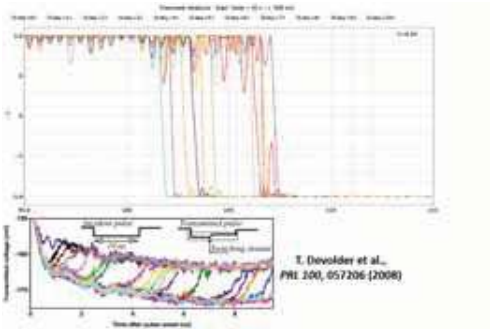


Fig 9: modeling of the stochasticity of the STT switching process and comparison with characterization results

IV. Storage and Computing based on Non-volatile RAMs

To illustrate the usage of novel non-volatile memories as computation and storage unit, compact models have been used to study the effect of variability on the circuit behavior. To this extent, the evaluations have been done on a specific test case of neuromorphic circuit. In the following we will address the topic of neuromorphic computing and will describe the research efforts directed to hardware implementation of spiking neural networks with resistance-change devices used to implement the synaptic and neuronal function. In addition, we will present an analysis of such network under the effect of device variations and analyze its robustness and learning efficiency.

Implementation of artificial neural networks with non-volatile devices have made great progress recently. An artificial neural network has two basic computing elements: the neuron and the

synapse. The neural computation consists in the weighted summation of neuron input signals subjected to the transfer function of the neuron. Research work related to emergent non-volatile technologies utilization (i.e. spin transfer torque magnetic random access memories and memristors) has been mostly focused on how to mimic the biological synapse functionality. Significant benefits can be gained by using bidirectional RRAMs with continuous conductance tuning capability for neuromorphic computing, since these devices behave as natural electrically-controlled synaptic device. The bidirectional analog switching property could be utilized in neuromorphic systems inspired by biology such as spiking neural networks, as well as in hardware accelerators for machine learning algorithms with artificial learning rules [26].

The availability of magnetic junction devices made them very attractive for spin-based neurons and spin-based synapses implementations, as they mimic biological functionality of the cognitive process of the brain at low energy with only one single device [27]. It has been proposed to use the intrinsic spintronic switching stochasticity to emulate the learning ability of neural synapses [28-30]. Binary spintronic devices have been explored for energy efficient neuromorphic computing [30]. A great advantage of the spintronic devices is that they can be used also to perform the neuron function when they are used in biology-inspired spiking neural networks, opening the path towards full-spintronic neural networks [31]. Indeed, several approaches have been proposed to mimic the neuron function by a spintronic device, such as domain wall motion (DWM) neuron [32] or magnetic tunnel junction (MTJ) neuron with spin orbit torque assist [30].

An example of a fully-stochastic spiking neural network was presented in [29]. This network is designed for pattern recognition. It is a two-layer (one input, one output), fully-connected (all neurons in the input layer are connected to all neurons in the output layer) neural network constructed with spiking neurons and stochastic synapses (as illustrated in Fig. 10a) [29]. The output layer has the added function of lateral inhibition on a winner-takes-all strategy to assure that neurons learn different patterns. In this design, the link between the input neuron and the output neuron is implemented with a compound magnetoresistive synapse (CMS) as shown in Fig. 10c. This synaptic device employs multiple (N) binary MTJ elements connected in parallel. They operate in stochastic regime and act as one single synapse. This CMS is expected to exhibit $CL = N+1$ discrete conductance levels obtained by summing up the parallel conductance, ranging from the minimum compound synaptic weight – achieved when all MTJs are in high resistive state (anti-parallel magnetization) – to maximum compound synaptic weight – achieved when all MTJs are in low resistive state (parallel magnetization). The spiking neuron is illustrated in Fig. 10b. It is designed using an MTJ device which can operate in stochastic writing mode (when V_{write} is enabled), read mode (when V_{Read} is enabled), or reset mode (when V_{Reset} is enabled). The MTJ device is initialized at low resistance state (parallel configuration). During the write operation the synaptic current, with the corresponding weight is applied to the MTJ device, through the inverting amplifier. If the MTJ device changes its state, the pulse generator is activated, therefore the neuron fires.

Fabrication- and environmental-induced process variations have been investigated and the impact of their behavior on neuron firing rate and the learning process has been presented in [23-24].

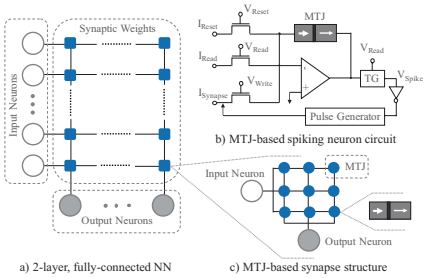


Fig. 10: Schematic representation of the neural network under study, [29]

Our results showed that while the read and reset operations of the spiking neuron are controlled in such a way to compensate for reliability issues, the stochastic write operation is affected by process parameter variations and variations in the environmental conditions. The switching probability of the neuron's MTJ, directly related with the probability that the neuron will spike, is dependent on the amplitude of the synaptic current and on the operation temperature. The combination of these effects can cause a decrease in the neuron firing rate, which in turn, slows the learning process and decreases its accuracy, and, in extremis, causes an evaluation error of the neural network. Or, on the contrary, it can cause an increase in the neuron firing rate which might lead to un-necessary potentiation or depression actions (see Fig. 11).

Moreover, the precision of the learning and recognition process is dependent on the number of conduction levels which can be achieved by the compound synapses. The speed and power consumption during the learning operation are dependent on the number of potentiation/depression operations required to switch between conductance levels. For a specific synaptic design, the number of conductance levels is dependent on the number of MTJ devices used for its design. This number defines the resolution of the synaptic weight.

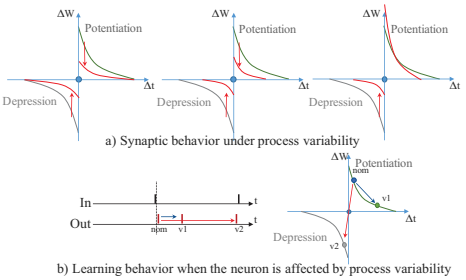


Fig. 11: Schematic representation of the MTJ process variability effect on the synaptic potentiation and depression

The fabrication-induced process variability affects the learning and recongnition process, which in turn translates in the necessity of a larger number of samples in the training set. Under certain variability scenarios we could also observe variations in learning efficiency for different patterns fed to the same network. This case is mostly observed when the process variability causes asymmetry between the potentiation/ depression curves. To investigate the behavior of the SNN under study, we have evaluated the learning precision of the network as a function of the network design and training set

size. To that end, we have designed 2 SNNs with different settings for the synaptic weight implementation, i.e., the synaptic weight is implemented as 1 stochastic MTJ and 16 stochastic MTJs in parallel, respectively. The size of the training set is also explored and increased at each simulation step. To evaluate the effect of device-to-device variations, we have performed the same set of simulations, but considered on 100 instances of SNN under random process variability and the results obtained are illustrated in Fig 12. The nominal (Ni), minimum (Ni_min) and maximum (Ni_max) values obtained for the recognition error in all scenarios are illustrated in the figure (with i=1, 16, the number of MTJ devices in a synapse). The simulation results show that the process variability has indeed an effect on the learning process of the SNN under study, but this effect can be mitigated by increasing the size of the SNN or/and the size of the training set.

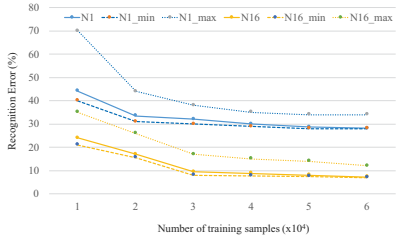


Fig. 12 – SNN Recognition Error (%) as a function of the size of the training data-set for different designs of the compound synapse

V. Conclusions

This paper presents the main strengths and weaknesses of the two most relevant novel non-volatile memories, i.e., the resistive RAM (especially the HfO2 based OxRAM) and the spin-transfer-torque RAM. The leading ongoing concern with RRAM is the variability of the switching parameters. It has now become evident that the low power switching of RRAM will impose variant filament formation due to movement of only countable number of atoms. Therefore, not only each device might show different resistance value but also one can have cycle to cycle variation. While variability is also important for the STT-MRAM devices, the main concern related to these devices is the write operation which requires high current densities.

These shortcomings are visible at circuit level, affecting the data storage robustness and the computation efficiency of architectures build on novel non-volatile memory devices. This brings forth the need for robustness and reliability analysis in all future computation architectures, even the ones intrinsically fault tolerant, as is the case of neural networks.

References

[1] C. Pancratov, J.M. Kurzer, K.A. Shaw, L. Matthew, "Why computer architecture matters: memory access," Computing in Science and Engineering, vol. 10, no. 4, pp. 71-75, 2008

[2] ITRS 2015 report. [Online]. Available: http://www.itrs.net/

[3] I. Vourkas, G. Sirakoulis, "Memristor-Based Nanoelectronic Computing Circuits and Architectures," Series: Emergence, Complexity and Computation, Vol. 19, Springer Publisher.

[4] W. C. Black and B. Das, "Programmable logic using giant-magnetoresistance and spin dependent tunneling devices," Journal of Applied Physics, pp. 6674-6679, 2000.

[5] S. Balatti, S. Ambrogio, D. Ielmini, "Normally-off Logic Based on Resistive Switches—Part I: Logic Gates," IEEE Transactions on Electron Devices, vol.62, no.6, pp.1831-1838, 2015

[6] S. Senni, et al., "Normally-Off Computing and Checkpoint/Rollback for Fast, Low-Power, and Reliable Devices," in *IEEE Magnetics Letters*, vol. 8, 2017 doi: 10.1109/LMAG.2017.2712780

[7] DARPA SyNAPSE, on-line at http://www.artificialbrains.com/darpa-synapse-program

[8] Introducing a Brain-inspired TrueNorth's neurons to revolutionize system architecture, http://www.research.ibm.com/articles/brain-chip.shtml

[9] The Human Brain Project, on-line at https://www.humanbrainproject.eu/

[10] S. Kvatinsky et al., "MAGIC—Memristor-Aided Logic," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 61, no. 11, pp. 895-899, Nov. 2014, doi: 10.1109/TCSII.2014.2357292

[11] Y. Wang *et al*., "An Energy-Efficient Nonvolatile In-Memory Computing Architecture for Extreme Learning Machine by Domain-Wall Nanowire Devices," in *IEEE Transactions on Nanotechnology*, vol. 14, no. 6, pp. 998-1012, Nov. 2015, doi: 10.1109/TNANO.2015.2447531

[12] Z. Pajouhi, et al., "Exploring Spin-Transfer-Torque Devices for Logic Applications," in *IEEE Transactions on CAD*, vol. 34, no. 9, pp. 1441-1454, Sept. 2015

[13] E. I. Vatajelu, H. Aziza, C. Zambelli, "Nonvolatile memories: Present and future challenges," in IDT, 2014.

[14] S. Kannan et al., "Modeling, detection, and diagnosis of faults in multilevel memristor memories," TCAD, 2015.

[15] E. Vianello, et al., "Resistive Memories for Ultra-Low-Power embedded computing design," *IEEE IEDM,* pp. 6.3.1-6.3.4, 2014

[16] T. Cabout, et al., "Temperature impact (up to 200 °C) on performance and reliability of HfO2-based RRAMs," *IEEE International Memory Workshop* p. 116–119, 2013.

[17] T. Diokh, et al, "Investigation of the impact of the oxide thickness and RESET conditions on disturb in HfO2-RRAM integrated in a 65nm CMOS technology," *IEEE IRPS,* pp. 3-6, 2013.

[18] M. Bocquet, D. Deleruyelle, H. Aziza, C. Muller, J.-M. Portal, T. Cabout and E. Jalaguier, "Robust compact model for bipolar oxide-based resistive switching memories," *IEEE Transactions on Electron Devices (TED),* vol. 61, pp. 674 - 681, 2014.

[19] F. Nardi, S. Larentis, S. Balatti, D. C. Gilmer et D. Ielmini, «Resistive switching by voltage-driven ion migration in bipolar RRAM -- part I : experimental study,» *IEEE Transactions on Electron Devices,* vol. 59, n° %19, pp. 2461-2467, 2012

[20] Deng E., et al., Non-Volatile Magnetic Decoder Based on Magnetic Tunnel Junctions, Electronics Letters, Ed. IEEE, Vol. , 2016

[21] J. Slonczewski, "Currents and torques in metallic magnetic multilayers," Journal of Magnetism and Magnetic Materials, vol. 159, 1996.

[22] Fuchs, N. et al., "Spin-transfer effects in nanoscale magnetic tunnel junctions," Applied Physics Letters, vol. 85, no. 7, pp. 1205–1207, 2004.

[23] K. C. Chun, et al., "A scaling roadmap and performance evaluation of in-plane and perpendicular mtj based stt-mrams for high-density cache memory," IEEE JSSC, vol. 48, no. 2, pp. 598–610, 2013.

[24] W.Guo et al., Journal of Physics D : Applied Physics, vol. 43, no. 21, p. 215001, May 2010

[25] Jabeur, K., et al. (2014). Comparison of Verilog-A compact modelling strategies for spintronic devices. Electronics Letters, 50(19), 1353-1355.

[26] H. Wu et al. Device and circuit optimization of RRAM for neuromorphic computing. In 2017 IEEE IEDM, pages 11.5.1_11.5.4, Dec 2017.

[27] A. Basu et al. Low-power, adaptive neuromorphic systems: Recent progress and future directions. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2018.

[28] D. Zhang et al. All spin artificial neural networks based on compound spintronic synapse and neuron. IEEE Transactions on Biomedical Circuits and Systems, 10(4):828-836, Aug 2016.

[29] D. Zhang et al. Stochastic spintronic device based synapses and spiking neurons for neuromorphic computation. NANOARCH, 2016.

[30] A. Sengupta et al. Probabilistic deep spiking neural systems enabled by magnetic tunnel junction. IEEE Transactions on Electron Devices, 63(7):2963-2970, 2016.

[31] A. Sengupta et al. Performance analysis and benchmarking of all-spin spiking neural networks (special session paper). IJCNN, 2017.

[32] M. Sharad et al. Spin-neurons: A possible path to energy-efficient neuromorphic computers. J App. Phys., 114(23), 2013.

[33] E. I. Vatajelu and L. Anghel, "Reliability analysis of MTJ-based functional module for neuromorphic computing," *2017 IEEE IOLTS*, Thessaloniki, 2017, pp. 126-131.

[34] E. I. Vatajelu and L. Anghel, "Fully-connected single-layer STT-MTJ-based spiking neural network under process variability," *NANOARCH,* 2017, pp. 21-26