A variability study of PCM and OxRAM technologies for use as synapses in neuromorphic systems

Daniele Garbin

UNIVERSITÉ GRENOBLE ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Nanoélectronique et nanotechnologies**

Arrêté ministériel : 7 août 2006

Présentée par

Daniele GARBIN

Thèse dirigée par **Prof. Gérard GHIBAUDO** et
codirigée par **Dr. Barbara DE SALVO**

préparée au sein du **CEA-LETI**
dans **l'École Doctorale d'Électronique, Électrotechnique, Automatique et Traitement du Signal**

Étude de la variabilité des technologies PCM et OxRAM pour leur utilisation en tant que synapses dans les systèmes neuromorphiques

Thèse soutenue publiquement le **15 décembre 2015**,
devant le jury composé de :

M Daniele IELMINI
Prof., Politecnico di Milano, Italie, Rapporteur
M Giacomo INDIVERI
Prof., Swiss Federal Institute of Technology in Zurich, Suisse, Rapporteur
M Ian O'CONNOR
Prof., Ecole centrale de Lyon, Président
Mme Elisa VIANELLO
Dr. Ing., CEA-Leti Grenoble, Co-encadrant de thèse
M Quentin RAFHAY
MCF, Université Grenoble Alpes (IMEP-LAHC), Co-encadrant de thèse
M Gérard GHIBAUDO
DR, CNRS Université Grenoble Alpes (IMEP-LAHC), Directeur de thèse
Mme Barbara DE SALVO
HDR, Dr. Ing. CEA-Leti Grenoble, Co-directeur de thèse, Invité
M Olivier BICHLER
Dr. Ing. CEA-List Gif-sur-Yvette, France, Invité

Abstract

Title: *A variability study of PCM and OxRAM technologies for use as synapses in neuromorphic systems.*

The human brain is made of a large number of interconnected networks which are composed of neurons and synapses. With a low power consumption of only few Watts, the human brain is able to perform computational tasks that are out of reach for today's computers, which are based on the Von Neumann architecture. Neuromorphic hardware design, taking inspiration from the human brain, aims to implement the next generation of non-Von Neumann computing systems. In this thesis, emerging non-volatile memory devices, specifically Phase-Change Memory (PCM) and Oxide-based resistive memory (OxRAM) devices, are studied as artificial synapses for use in neuromorphic systems. The use of PCM devices as binary probabilistic synapses is proposed for complex visual pattern extraction applications. The impact of the PCM programming conditions on the system-level power consumption is evaluated. A programming strategy is proposed to avoid the PCM resistance drift. It is shown that, using scaled devices, it is possible to reduce the synaptic power consumption. The OxRAM resistance variability is evaluated experimentally through electrical characterization, gathering statistics on both single memory cells and at array level. A model that allows to reproduce OxRAM variability from low to high resistance state is developed. An OxRAM-based convolutional neural network architecture is then proposed on the basis of this experimental work. By implementing the computation of convolution directly in memory, the Von Neumann performance bottleneck is avoided. The robustness of the neuromorphic system to OxRAM variability is demonstrated for complex visual pattern recognition tasks such as handwritten characters and traffic signs recognition.

Résumé

Titre : *Étude de la variabilité des technologies PCM et OxRAM pour leur utilisation en tant que synapses dans les systèmes neuromorphiques.*

Le cerveau humain est composé d'un grand nombre de réseaux interconnectés, dont les neurones et les synapses en sont les briques constitutives. Caractérisé par une faible consommation de puissance, de quelques Watts seulement, le cerveau humain est capable d'accomplir des tâches qui sont inaccessibles aux systèmes de calcul actuels, basés sur une architecture de type Von Neumann. La conception de systèmes neuromorphiques vise à réaliser une nouvelle génération de systèmes de calcul qui ne soit pas de type Von Neumann. L'utilisation de mémoires non-volatile innovantes en tant que synapses artificielles, pour application aux systèmes neuromorphiques, est donc étudiée dans cette thèse. Deux types de technologies de mémoires sont examinés : les mémoires à changement de phase (Phase-Change Memory, PCM) et les mémoires résistives à base d'oxyde (Oxide-based resistive Random Access Memory, OxRAM). L'utilisation des dispositifs PCM en tant que synapses de type binaire et probabiliste est étudiée pour l'extraction de motifs visuels complexes, en évaluant l'impact des conditions de programmation sur la consommation de puissance au niveau système. Une nouvelle stratégie de programmation, qui permet de réduire l'impact de la dérive de la résistance des dispositifs PCM (dit « drift ») est ensuite proposée. Il est démontré qu'en utilisant des dispositifs de tailles réduites, il est possible de diminuer la consommation énergétique du système. La variabilité des dispositifs OxRAM est ensuite évaluée expérimentalement par caractérisation électrique, en utilisant des méthodes statistiques, à la fois sur des dispositifs isolés et dans une matrice mémoire complète. Un modèle qui permets de reproduire la variabilité depuis le niveau faiblement résistif jusqu'au niveau hautement résistif est ainsi développé. Une architecture de réseau de neurones de type convolutionnel est ensuite proposée sur la base de ces travaux éxperimentaux. La tolérance du système neuromorphique à la variabilité des OxRAM est enfin démontrée pour des tâches de reconnaissance de motifs visuels complexes, comme des caractères manuscrits ou des panneaux de signalisations routières.

Contents

Acknowledgments

This PhD thesis was prepared at the Université Grenoble Alpes, CEA LETI and IMEP-LAHC. First of all, I would like to thank my PhD thesis directors Prof. Gérard Ghibaudo and Dr. Barbara De Salvo, for their precious guidance and wisdom. Then, I would like to thank my advisors Dr. Elisa Vianello and Dr. Quentin Rafhay for supporting me and helping me improve during these three years. I thank all the LCM team: Luca Perniola, Eric Jalaguier, Gabriele Navarro, Véronique Sousa, Gabriel Molas, Alain Persico, Christelle Charpin, Sophie Bernasconi, Carine Jahan, Rémi Coquand, Etienne Nowak, Laurent Grenouillet, Cathérine Carabasse, Jean-François Nodin, Guillaume Bourgeois, Jérôme Lozat and Khalil El Hajjam. I would like to thank Dr. Olivier Joubert and the LabEx Minos for supporting my PhD thesis under Grant ANR-10-LABX-55-01.

I am deeply grateful to Dr. Olivier Bichler for our nice and fruitful collaboration. I would also like to thank Dr. Christian Gamrat for welcoming me during the time that I spent in CEA LIST. I thank Alain Lopez, Jacques Cluzel, Denis Blachier, Carlo Cagli, Giovanni Romano, Olga Cueto, and all the people from LCTE and LICL laboratories for their support. I thank our collaborators from STMicroelectronics for providing most of the samples that I tested during this research.

I thank all the PhD students, interns and postdocs who shared with me coffee breaks, lunches, trips and happy moments during the last three years: Marinela, Thanasis, Manan, Boubacar, Quentin, Thomas, Gabriele, Giorgio, Thérèse, Jérémy, Florian, Yann, Sebastien, Amine, Julien, Issam, Sarra, Heimanu, Rémi, Luca, Niccolò, Thilo, Julia, Mourad, Mouhamad, Giuseppe, Cécile, Adam, Daeseok, Luc, Martin, Davide, Marco, Aurore, Fabien, Loïc, Romain, Vincent, Jose, Mathilde, Corentin, Anouar, Mathias, Patricia, Alexandre, Anthony, Louise, Jessy and all those people that i may have forgotten in this list. I would also like to thank Sabine, Brigitte, Sylvaine and Malou for their help during these three years.

I am grateful to Prof. Ian O'Connor, Prof. Daniele Ielmini and Prof. Giacomo Indiveri for accepting to review this thesis and attending my PhD defense.

Last but not least, I thank my family for their support and for always being there for me.

Introduction

The human brain is made of a large number of interconnected networks which are composed of neurons and synapses. With a low power consumption of only few Watts, the human brain is able to perform computational tasks that are out of reach for today's computers, currently are based on the Von Neumann architecture. Neuromorphic hardware design, taking inspiration from the human brain, aims to implement the next generation of non-Von Neumann computing systems. Neuromorphic systems are designed to perform, in a power efficient way, those tasks at which the human brain is excellent, as for example the recognition of complex visual and auditory patterns.

Emerging Non-Volatile Memory (NVM) devices have been studied in the recent years as possible solutions to implement artificial synapses in neuromorphic hardware systems. In this work, emerging NVM devices, in particular Phase-Change Memory (PCM) and Oxide-based resistive memory (OxRAM) devices, are studied having in mind the central role that they will play in future memory and computing architectures. We investigate the use of these devices for the implementation of artificial synapses, with a special focus on device variability and its impact on the performance of neuromorphic computing systems.

Manuscript outline

In Chapter 1, we introduce the context and motivation behind the research conducted during the preparation of this PhD. Given the interdisciplinarity involved in this project, this chapter describes in depth the basics concepts that are needed to contextualize this research, in the framework of both conventional memory and neuromorphic computing architectures.

In Chapter 2, we focus on PCM technology, one of the most mature among the emerging non-volatile memory technologies. PCM devices offer the possibility of multilevel programming by gradually changing the size of the crystalline portion of the active phase-change material. We analyze the drawbacks related to the use of the multilevel PCM synapse approach. Therefore, driven by the motivation to overcome the limitations associated to the multilevel programming, we explore by simulations the use of PCM synapses operated in binary mode, where only two

resistance states are exploited. The use of the proposed binary PCM synapse is studied in a neuromorphic system designed for complex visual pattern extraction.

In Chapter 3, we investigate the binary operation of OxRAM devices. Since variability is the main drawback of OxRAM technology, we carry out an extensive work of electrical characterization on single bitcells and on 16 kb memory array, in order to understand the source of variability. Starting from the electrical characterization results, we develop a simplified trap-assisted tunneling model to reproduce the OxRAM variability from low (LRS) to high resistance state (HRS), highlighting the continuity of the mechanisms involved in the variability. We carry out this analysis with a dual goal. On one hand, the developed model provides an insight on the source of variability in OxRAM, suggesting technology guidelines for the improvement of reliability. On the other hand, the computational efficiency of the developed model allows to simulate large memory arrays and take into account the synaptic variability corresponding to a wide range of programming conditions in neuromorphic system simulations.

In Chapter 4 we propose an OxRAM-based synapse design that combines together the advantages of multilevel and binary approaches. Using such synapses, we propose a hardware implementation of a convolutional neural network (CNN) for complex visual applications such as handwritten digits and traffic signs recognition. We investigate the tolerance of the proposed network to both temporal and spatial synaptic variability.

In Chapter 5 we provide a general conclusion of the research carried out for this thesis. Finally, we provide a perspective on the future work that needs to be done for the further progress of the research on neuromorphic hardware.

Chapter 1

Emerging Non Volatile Memories and Neuromorphic Systems

In this chapter, we introduce the context and motivation behind the research conducted during the preparation of this PhD. On one side, emerging Non-Volatile Memory (NVM) devices are studied having in mind the central role that they will play in the memory architectures of the future. On the other side, a novel application of NVM devices, which has gained a large interest over the last few years, is investigated: the implementation of artificial synapses in brain-inspired computing architectures. Given the interdisciplinarity involved in this project, this chapter describes in depth the basics concepts that are needed to contextualize this research, in the framework of both conventional memory and neuromorphic computing architectures.

1.1 The semiconductor memory market

The design of today's computing systems is based on the Von Neumann architecture [1]. In this architecture, a marked distinction exists between the role of the Central Processing Unit (CPU) and the Memory Unit (MU). The CPU is in charge of performing the arithmetic operations, logic functions, control tasks and input/output operations that are specified by a set of instructions, i.e. a computer program, which is stored in the MU. The MU contains both the code of the computer programs and the data. Data comprise the information that has to be processed by the CPU and the results of the computation [1].

The simplest architecture for organizing memory is the *flat* memory architecture. In this architecture, data are stored in a single, large memory unit block in the form of array. However, the memory access time and the power consumption associated to the access to information increase with the size of the memory array. Hence, memory power and access time dominate the total power and performance when a large storage is required for computation [2]. In fact, a gap exists between processor and

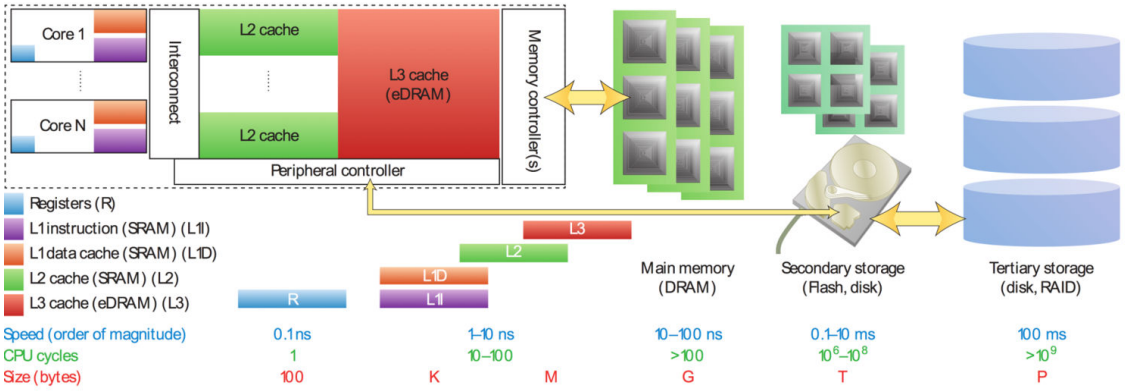


Figure 1.1. The memory hierarchy in computers. Small amounts of high-performance volatile and expensive memory are close to the CPU. Large amounts of slower, non volatile and low-cost storage units are far from the CPU at the bottom of the hierarchy. Source: [4].

memory in terms of performance: computation performance is typically limited by how fast the data in memory can be accessed, with latency and bandwidth being the main limiting factors. This gap is commonly referred to as the memory bottleneck [2].

In more advanced memory architectures, memory is not *flat*. It is structured as a *hierarchy* of volatile and non-volatile memory devices, in order to achieve an optimal trade-off between cost and performance. The goal of this memory hierarchy, shown in Figure 1.1, is to mitigate the problem of the memory bottleneck, bridging the performance gap between the fast CPU and the slower memory and storage technologies, keeping the system costs down [3]. As illustrated in Fig. 1.1, at the top of the hierarchy, close to the CPU unit, is the memory that is accessed most frequently. Static Random Access Memory (SRAM) is the technology of choice because it allows the fastest operation speed. However, due to the large silicon area required, it is also the most expensive technology [4]. The technology adopted for the main memory is typically the Dynamic Random Access Memory (DRAM), which often resides in a different chip than the CPU because the technology process is different. For over 30 years, SRAM [5] and DRAM [5] technology have dominated the memory market [6]. Both SRAM and DRAM, however, are volatile memories, i.e. the information stored in memory is lost when the device is turned off. At the bottom of the memory hierarchy, magnetic Hard Disk Drives (HDDs) have been used for over 50 years [7] as a first choice for non-volatile storage solutions. Since the advent and explosive growth of portable devices such as music players and cellular phones, however, Flash memory [8], [9] has forced its way into the information storage hierarchy, between DRAM and HDD, as non-volatile storage solution. As shown in Fig. 1.2, the growth of Flash technology has exploded over the last few years,

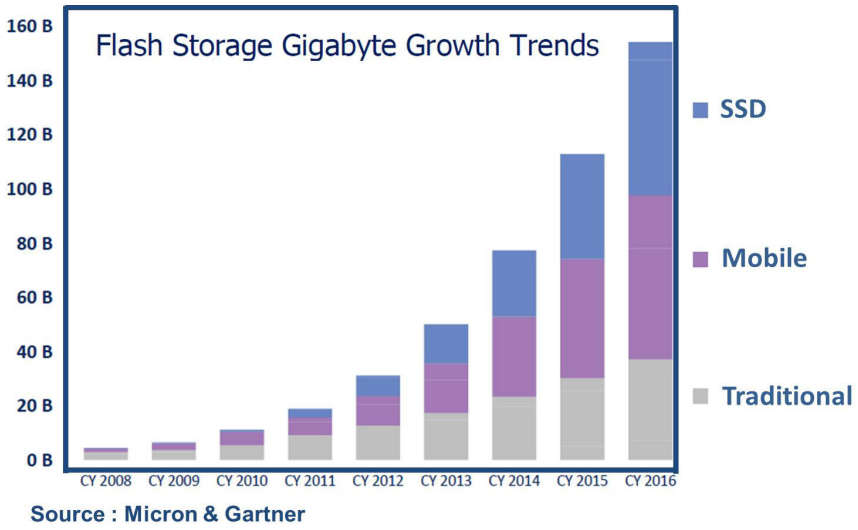


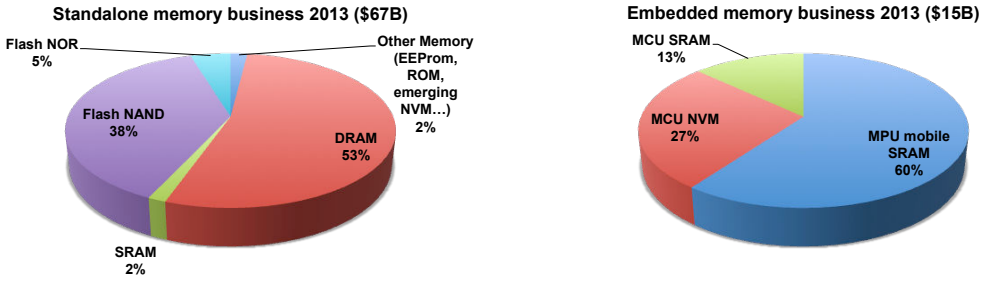
Figure 1.2. Flash storage gigabyte growth trends, source: [12].

and it has become the dominant data storage technology for mobile applications. Enterprise-scale computing systems and cloud data storage systems are also starting to adopt Flash technology to complement the HDD storage units with Solid-State Drives (SSDs) [10]. The Flash memory market in this segment is forecast to grow over the next few years, also thanks to the adoption of vertically 3D stacked cells solutions [11].

However, Flash technology is facing challenges in scaling due to intrinsic physical limitations related to the technology, such as floating gate interference [13], reduced coupling between control and floating gate [14], short channel effects [13] and small electron charge in the floating gate [15], [16]. The emergence of a non-volatile memory technology able to combine at the same time high performance, high density and low cost can potentially lead to deep changes in the memory/storage hierarchy [3]. A non-volatile memory with latency compared to DRAM would be a game changer in storage tiering [17]. For these reasons, research efforts are being done in order to find new non-volatile memory solutions, with better scalability compared to Flash and possibility of vertical 3D stacking. Such a technology would allow to reach the highest possible densities achievable with future technology nodes [3] and offer the possibility to mitigate the problem of the memory bottleneck.

Considering the semiconductor memory market, two main types of memory business can be considered:

- The standalone memory market, with focus on density and performance. It is a very concentrated market with 5 Integrated Device Manufacturers (IDMs) holding 95% of the total business: Samsung, Micron/Intel, SK Hynix, Toshiba and SanDisk.



(a) (b)

Figure 1.3. (a) Standalone and (b) embedded memory market in 2013. Source: [18].

- The embedded memory market, with low power consumption and high thermal stability being some of the most restricting specifications. There are two types of embedded memory, depending on the level of system integration. I System on Chip (SoC) such microcontrollers (MCUs) for smart cards, automotive or mobile microprocessors (MPUs) for portable systems. II System in Package (SiP) where a number of integrated circuits are enclosed in a single module (package). The embedded memory market is more fragmented, with foundries manufacturing the bulk part of the total production [18].

As shown in Fig. 1.3a, Flash NAND technology and DRAM currently dominate the standalone memory market, representing about 90% of overall memory sales. A low-cost emerging memory technology with densities larger than Flash and speed comparable to DRAM can possibly conquer a large portion of the market. After 2020, emerging NVM could also replace SRAM in the MCU and MPU embedded memory business [18] (Fig. 1.3b).

1.2 Emerging non-volatile memory technologies

In the quest for innovative non-volatile memory solutions, different technologies have emerged in research over the last 15 years [19], [20]. These technologies are free from the limitations of Flash, which are low endurance (i.e. a limited number of write operations is possible), need for high voltage supply for programming, long write time and complex erase procedure [4]. Another limitation of Flash is the fact that it is a Front-End-Of-Line (FEOL) technology, difficult to co-integrate with sub-32 nm CMOS [16].

The main emerging non-volatile memory technologies are the following:

- Phase-Change Random Access Memory (PCRAM or PCM);
- Spin-Transfer-Torque Magnetic Random Access Memory (STT-MRAM);

- Conductive-Bridging Random Access Memory (CBRAM);
- metal Oxide resistive Random Access Memory (OxRAM).

These emerging memory technologies store information using physical mechanisms which do not rely on storing charge in a capacitor or the floating gate of a transistor as in the case of SRAM, DRAM and Flash. They are integrated in the Back-End-Of-Line. It is worth noticing that in literature the generic term resistive RAM (RRAM or ReRAM) is often used to generically refer to both OxRAM and CBRAM. In the next sections, an overview of the main emerging non-volatile memory technologies will be given, with details on some important properties and performance aspects.

1.2.1 PCRAM

Phase-Change Random Access Memory (PCRAM, or PCM) working principle is based on the electrical properties of phase-change materials. Phase-change materials, in fact, feature a high contrast in resistivity between amorphous and crystalline phases. The amorphous phase is characterized by a high electrical resistivity, while the crystalline phase features low resistivity [21]. In PCRAM it is possible to switch the material between amorphous and crystalline phase multiple times. Most of phase-change materials are chalcogenides, which are alloys featuring at least one element from the VI group of the periodic table. $Ge_2Sb_2Te_5$ (GST) is one of the most studied chalcogenide phase-change materials. Other examples are GeTe, $GeSeTe_2$, $AgSbSe_2$ [21] and different variants obtained by doping [22], [23] or enrichment of alloying elements [24], [25]. GaSb is an example of phase-change material that is not a chalcogenide [26].

In PCRAM, the phase-change material is switched between amorphous and crystalline phase by Joule heating. Figure 1.4a shows the shape of typical current-voltage characteristics for crystalline and amorphous phases of phase-change materials. Crystallization is achieved by heating the material above its crystallization temperature (SET operation). Amorphization is achieved by melting the material in the liquid state and rapidly quenching it into the disordered amorphous phase (RESET operation). These operations are performed by electrical current pulses: high-power pulses are required for the RESET operation, moderate power but longer duration pulses are used for the SET operation. In order to retrieve the information, low power pulses are used to sense the resistance of the device [21]. Figure 1.4b shows schematically a mushroom-shaped PCRAM cell. The device is composed of a phase-change material sandwiched between a top electrode and a bottom electrode in the form of small cross-section heater plug. The active region is defined as the portion of the phase-change material that actually undergoes phase switching. It is located right above the heater plug, where the temperature reaches the highest value due to high current density.

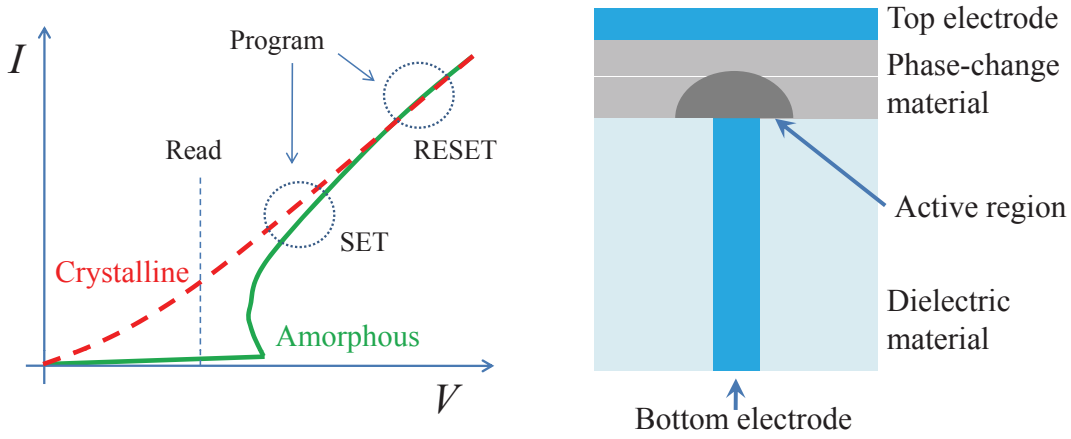


Figure 1.4. (a) Typical current-voltage characteristics of crystalline and amorphous phases of phase-change materials. (b) Schematic cross-section of a phase-change memory cell. Source: [27].

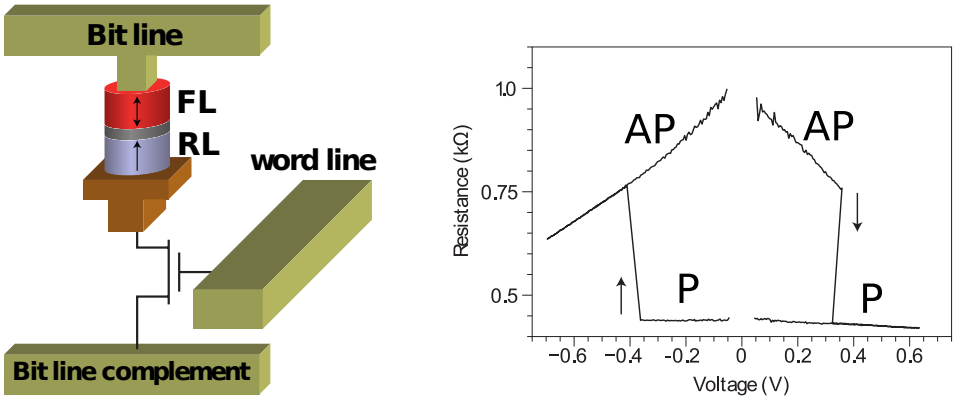
One of the limiting factors for the adoption of PCRAM technology is the relatively high RESET current [28]. However, the programming current scales down with device area. In ultra-scaled devices with 10 nm feature size, the RESET current is shown to decrease down to the microampere range [29]. Furthermore, material and interface engineering can significantly contribute to the reduction of the RESET current [23], [30].

One of the attractive features of PCRAM is the possibility of achieving multilevel-cell (MLC) storage. This means that the device can be programmed into multi-level resistance states, in addition to the full SET and RESET resistance levels. This is obtained by modulating the ratio between the crystalline and amorphous region size within the active region. MLC functionality is an efficient way of decreasing the cost of memory, because it alows to store more information for a given silicon area [31]. However, in phase change materials the amorphous intermediate resistance states drift with time t towards higher resistance values, following a $R(t) = R_0(t/t_0)^{\nu}$ relationship, where R_0 is the resistance at initial time t_0, and ν is the drift coefficient, which depends on the material and on the device morphology [32]. As a consequence, after a certain amount of time has passed after programming, it is a not trivial task to distinguish the programmed states. Only using advanced cell readout methodologies which are intrinsically resilient to resistance drift, a reliable 2 bits/cell storage and data retention at high temperature can be achieved [31].

1.2.2 STT-RAM

In Spin-Transfer-Torque Magnetic Random Access Memory (STT-MRAM) devices, information is stored in the orientation of the magnetization of a nano-scale ferromagnetic layer. Figure 1.5a shows the schematic view of a typical STT-MRAM bit-cell. The main component of STT-MRAM is the Magnetic Tunnel Junction (MTJ), which consists of two magnetic layers separated by a tunneling barrier, composed of a thin layer of insulating MgO. The orientation of the magnetization of the Free Layer (FL) can be switched between two states and is used to store information. The magnetization of the Reference Layer (RL) is permanent and it serves the function of a stable reference for the magnetic orientation [33]. If the orientation of RL and FL are the same, the device is said to be in Parallel (P) state. If the orientation of the two layers is opposite, the device is in the Anti-Parallel (AP) state. The STT-MRAM working principle is based on two phenomena that have been discovered during the last two decades: the Tunnelling Magneto-Resistance (TMR) effect and the Spin-Transfer Torque (STT) effect. The TMR effect [34] is the cause of the resistivity contrast between the resistance R_{P} in the P state and the resistance R_{AP} in the AP state. The resistance of the device can be sensed in order to determine which is the magnetic state of the FL, so the stored information can be retrieved. The STT effect [35]–[37] is the effect which allows to switch the magnetic orientation of the FL. When electrons flow through the MTJ, a torque is exerted on the magnetization of the FL. If the torque is large enough, the magnetic state of the FL can be switched and information is written. The write operation is achieved according to the direction of the current flow. If a positive voltage is applied to the device, a non-polarized current, i.e. featuring electrons with random spin orientation, is injected in the direction that goes from the RL to the FL. Electrons with a spin opposite to the RL magnetization orientation are mostly reflected. Only the electrons with a spin having the same orientation as the RL magnetization will be transmitted by tunneling through the MTJ and will transfer their spin to the FL by STT. The result is that the MTJ will be in the parallel state. If a negative voltage is applied to the device, the electrons are injected from the FL to the RL. At the interface between the tunneling barrier and the RL, electrons with the same orientation as the RL will be transmitted through the MTJ. Electrons with an opposite spin will be reflected back in the FL and switch its magnetization to the AP state. In summary, as shown in Fig. 1.5b, a positive voltage leads to AP-to-P transition. A negative voltage leads to P-to-AP transition [38].

STT-MRAM is expected to have a very high endurance [38]. This is due to the fact that no magnetic degradation mechanism is associated to the switching of the magnetization orientation. In fact, no atoms are moved during write operations, contrary to PCRAM, CBRAM or OxRAM. However, the dielectric breakdown of the MgO tunnel barrier can occur if the voltage across the tunnel barrier exceeds roughly 400 mV [38], [39].

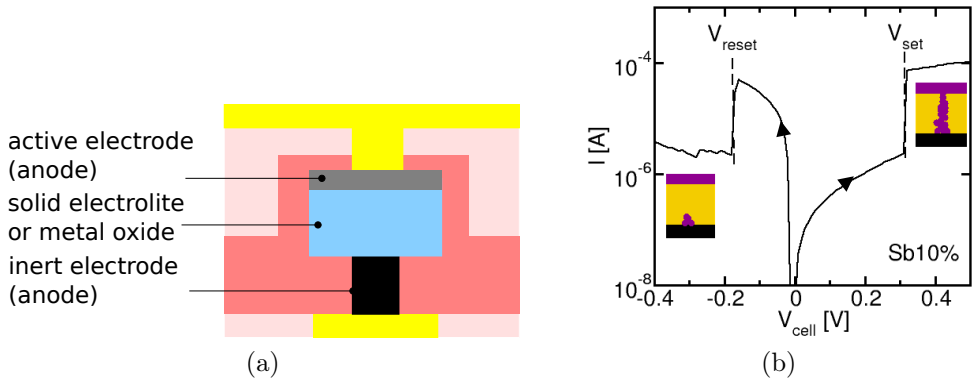


(a) (b)

Figure 1.5. (a) Schematic view of an STT-MRAM bit cell. The MTJ is composed of a permanent reference layer (RL), a tunnel barrier and a free layer (FL) element, with both layers magnetized perpendicular to the plane of the junction. (b) Typical resistance-voltage characteristics, showing switching between antiparallel (AP) to parallel (P) states (positive bias) and vice versa. Source: [38].

1.2.3 CBRAM

As shown in Fig. 1.6a, the structure of Conductive-Bridging Random Access Memory (CBRAM) devices consists of a Metal-Insultator-Metal structure where the top electrode (anode) is electrochemically active or oxidized under positive bias, and the bottom electrode (cathode) is electrochemically inert. The insulating materials between top and bottom electrode can be solid electrolytes [40] or metal oxides [41], [42]. Upon application of a positive voltage on the anode, mobile metal ions from the anode migrate, driven by the electric field, into the solid electrolyte or oxide and reduce on the inert cathode, forming a conductive filaments (CF) composed of element of the top electrode (typically Cu or Ag), bridging top and bottom electrode and bringing the device to the Low Resistance State (LRS, SET operation). When the voltage is reversed, metal ions migrate back to the anode dissolving the CF and bringing the device into the High Resistance State (RESET operation) [43], [44]. Figure 1.6b shows typical current voltage characteristics of CBRAM. A resistance ratio between HRS and LRS higher than 10^6 has been demonstrated by interface engineering of chalcogenide CBRAM with dual-layer electrolyte stack [45], [46]. Due to the fact that the migration of ions is a stochastic process, the configuration of the CF is different after each SET and RESET operation. This results in large resistance variability, especially in the HRS [4].



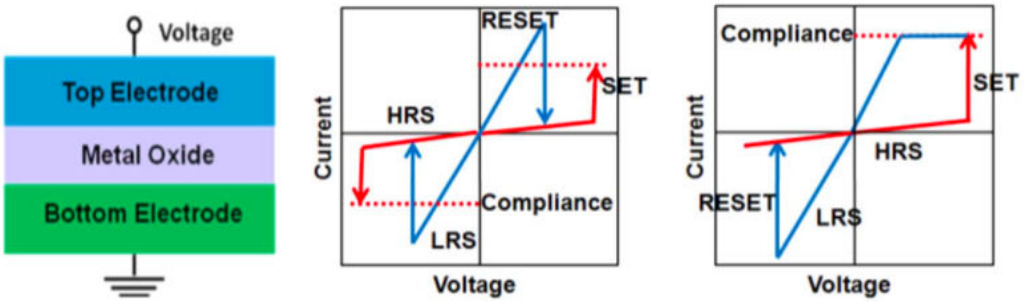
(a) (b)

Figure 1.6. (a) Schematic view of CBRAM device [47]. (b) Typical CBRAM current-voltage characteristics [40].

1.2.4 OxRAM

Similarly to CBRAM, Oxide-based resistive RAM (OxRAM) devices are also composed of a simple MIM structure, where a metal oxide is sandwiched between a top and a bottom electrode as shown in Fig. 1.7a. The application of an electric field on the device induces the creation and motion of oxygen vacancies V_O, resulting in the possibility of repeatedly form and destruct V_O-rich conductive filaments (CFs) in the oxide. This corresponds to a change in the resistance of the device, which can be switched between Low Resistance State (LRS) and High Resistance State (HRS) with SET and RESET operation, respectively. Figure 1.8 is a schematic illustration of the switching processes. Usually for the fresh samples in the pristine resistance state, a forming or electroforming process is needed to form a Conductive Filament (CF) in the oxide layer for the first time [48]. During the forming process, oxygen ions drift towards the top electrode interface driven by the orientation of the electric field. The formation of an interface oxide layer occurs if the top electrode material is oxidizable. Otherwise, oxygen accumulates in the form of nonlattice atoms if the top electrode material is inert. Thus, the top electrode/oxide interface behaves like an oxygen reservoir [49] for the subsequent SET/RESET operations. According to the polarity of the voltage needed to SET and the RESET the device, OxRAM operation is classified into two switching modes: unipolar and bipolar. Figures 1.7b and 1.7c show a schematic of the current-voltage characteristics for the two switching modes.

- In the unipolar switching mode (Fig. 1.7b), the SET and RESET operations depend only on the amplitude of the applied voltage. Thus, they can be achieved using the same programming polarity. For the RESET operation, current flowing through the CF causes Joule heating. The rising temperature activates the thermal diffusion of oxygen ions, which will diffuse away from the CF due to the concentration gradient [50], bringing the device to HRS. If



(a) (b) (c)

Figure 1.7. (a) Schematic of MIM structure of OxRAM devices and (b) Schematic unipolar and (c) bipolar current-voltage characteristics [48].

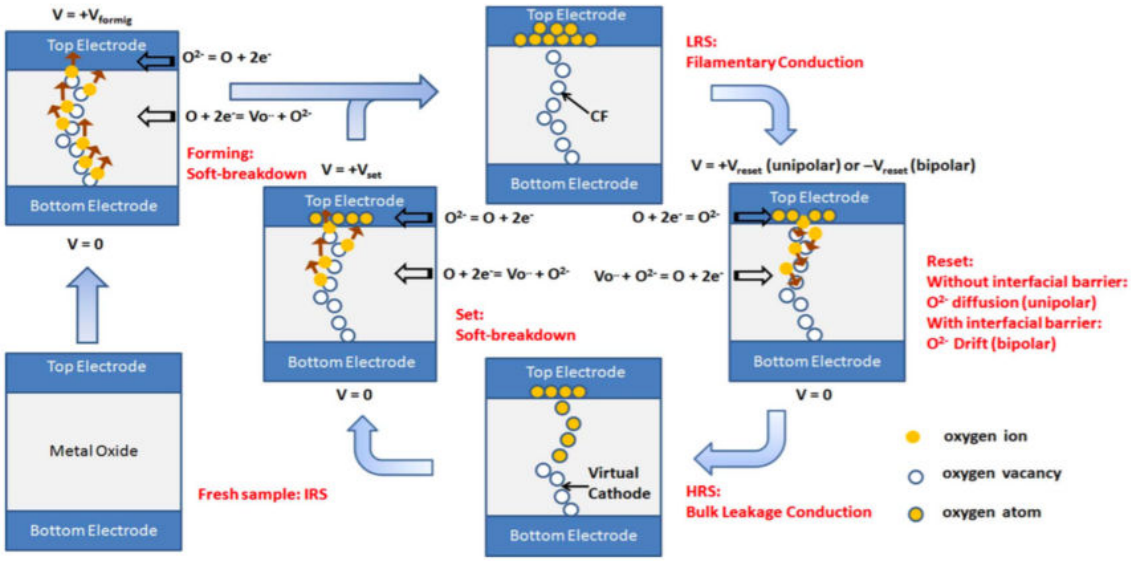


Figure 1.8. Schematic illustration of the working principle of OxRAM. Source: [48].

SET and RESET operations can equally occur at both positive and negative voltage polarities, the unipolar switching mode mode is also called *nonpolar*.

- In bipolar switching (Fig. 1.7c), the SET and RESET operations are performed at reverse voltage polarities. The interfacial oxide layer at the top electrode may present a significant diffusion barrier. In this case, thermal diffusion caused by Joule heating and concentration gradient alone is not sufficient, so a reverse electric field is needed to improve oxygen ions migration in the RESET process.

In order to achieve the required high temperature at the CF for the RESET

process, usually the unipolar devices requires a higher programming current compared to bipolar devices. In both switching modes, the SET operation occurs by dielectric soft breakdown, similarly to the forming operation. SET operation, however, typically requires a lower voltage compared to the forming one [48]. In order to avoid a permanent dielectric breakdown in the forming and set process, a current limitation, or compliance, is needed. The compliance current is usually provided by the semiconductor parameter analyzer in the case of devices composed of the MIM structure only (1R devices). In the case of 1T1R device, the compliance current is obtained via the selection transistor. Self-compliance 1D1R devices, with a diode as selector device, are an attractive solution for high-density crossbar structures [51].

One of the biggest advantages of this technology is the fact that it relies on simple structure and on materials that are widely used in semiconductor processes and current CMOS technologies. Some OxRAM material examples are HfO_x, TiO_x, AlO_x and TaO_x.

Although the device working principle is simple, the physics that govern the functioning of OxRAM devices are not fully understood yet. There are controversies about the shape of the conductive filament and the role that top and the bottom electrodes play in the switching mechanisms. The characteristics of the oxygen vacancies in terms of thermal stability and mobility are topics of intense research [44], because they are related to memory performance and reliability, such as high temperature data retention and speed. Physical observation of real-time formation and dissolution of CF with in-situ TEM is an active research field [52]–[54], because it can give guidelines for the improvement of the variability issue.

OxRAM variability

One of the main issues for the manufacturing and industrialization of OxRAM devices is the reproducibility of their electrical characteristics. Large resistance variations occur in fact not only between devices (device-to-device variability – d2d), but also between consecutive programming cycles of the same device (cycle-to-cycle variability – c2c). The problem of variability has been holding OxRAM technology back from commercialization despite its many attractive features, because it limits the size of the memory array that can be implemented. In fact, as the number of devices increases, the distributions of the devices in LRS and HRS tend to overlap, thus making it impossible to sense the difference between LRS and HRS state and thus retrieve information. The resistance variability in High-Resistance State (HRS) is typically larger than the variability of the Low-Resistance State (LRS) [48]. The HRS variability has been modeled introducing a variation on a tunneling barrier thickness [55], [56]. For LRS, it has been attributed to geometric variability of the conductive filament (CF) shape (i.e. CF radius, constriction point...) [48], [57], [58]. However, a unified model able to reproduce the variability from HRS to LRS [59] is still lacking.

1.2.5 Comparison of NVM technologies

Table 1.1 presents a benchmark of the different emerging nonvolatile memory technologies. Practically unlimited endurance and good speed are advantages of STT-MRAM technology. However the relatively low resistance ratio achievable in MTJ requires a memory cell architecture that limits its device density. OxRAM features better endurance and speed than PCM and CBRAM, but the problem of variability is much worse than that of PCM and STT-MRAM. In addition, all these memories promise to scale further than Flash and DRAM. When these emerging NVM memories were proposed, there was hope that one of them could become the *universal memory*, able to make a revolution in the memory hierarchy by meeting all specifications in terms of power consumption, high temperature data retention, speed, endurance, density, scalability and low cost [60]–[62]. However, as it can be evinced from Table 1.1, researchers now generally agree that the possibility of universal memory technology is not very realistic. Application-driven design imposes different specifications about memory performance at each level of the memory hierarchy. These specifications require trade-offs in device characteristics that are hard to obtain with an individual memory technology [4].

1.3 Neuromorphic systems

In addition to a drastic change in the organization of the memory hierarchy in traditional Von Neumann computing architectures, emerging non-volatile memories have been indicated as key players in a computation paradigm shift, beyond the traditional Von Neumann architecture, thanks to their use as nanoscale artificial synapses in neuromorphic hardware [64].

Neuromorphic hardware refers to an emerging field of computing systems design. It takes inspiration from biological neural networks that exist in mammalian nervous system and cerebral cortex. Research in neuromorphic hardware is interdisciplinary, requiring knowledge from computational neuroscience, neurobiology, machine learning, computer science, VLSI design and nanotechnology [65]. Unlike conventional Von Neumann computing architectures, in neuromorphic architectures memory and processing are not isolated tasks. They are interleaved entities, and memory participates in the task of processing the information [66]. Figure 1.9 shows the spectrum of models of computing, from the traditional program-centric Von Neumann-like architectures to emerging data-centric, learned computation models. In the era of the internet of things, with 10 billion devices networked together today (50 billion by 2020) [67], a huge amount of data has to be processed. New models of computation that learn from data, rather than executing instruction provided by programmers, are thus fundamental. The human brain is an example provided by nature of a computing system that learns from data in an efficient way. As

		PCRAM	STT-MRAM	CBRAM	OxRAM
Feature Size F (nm)	Demonstrated	45	65	20	5
	Projected	8	16	5	<5
Cell Area	Demonstrated	$4F^2$	$20F^2$	$4F^2$	$4F^2$
	Projected	$4F^2$	$8F^2$	$4F^2$	$4F^2$
Programming Voltage (V)	Demonstrated	3	1.8	0.6	1
	Projected	<3	<1	<0.5	<1
Programming Time (ns)	Demonstrated	100	35	<1	<1
	Projected	<50	<1	<1	<1
Programming Energy (J/bit)	Demonstrated	$6 \cdot 10^{-12}$	$2.5 \cdot 10^{-12}$	$8 \cdot 10^{-12}$	$< 1 \cdot 10^{-12}$
	Projected	$1 \cdot 10^{-15}$	$1.5 \cdot 10^{-13}$	N.A.	$1 \cdot 10^{-16}$
Read Voltage (V)	Demonstrated	1.2	1.8	0.2	0.1
	Projected	<1	<1	<0.2	0.1
Retention Time	Demonstrated	>10yr	>10yr	>10yr	>10yr
	Projected	>10yr	>10yr	>10yr	>10yr
Endurance (nb. cycles)	Demonstrated	10^9	$> 10^{12}$	10^{10}	10^{12}
	Projected	10^9	$> 10^{15}$	$> 10^{11}$	$> 10^{12}$

Table 1.1. Comparison of the performance of the different emerging nonvolatile memory technologies according to the 2013 International Technology Roadmap for Semiconductor (ITRS) [63], with projections for year 2026.

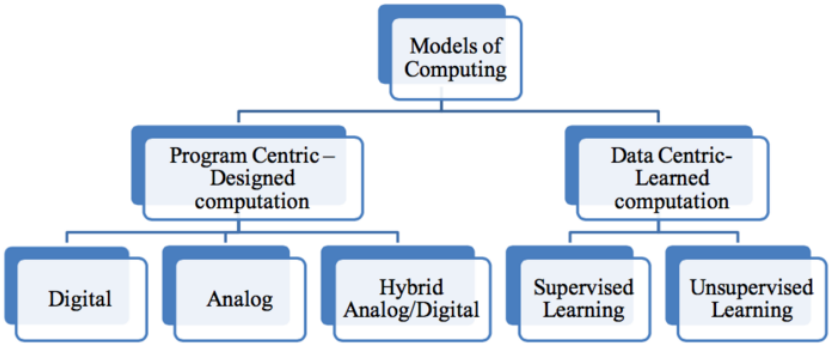


Figure 1.9. Taxonomy for traditional and emerging models of computation. Source: [63].

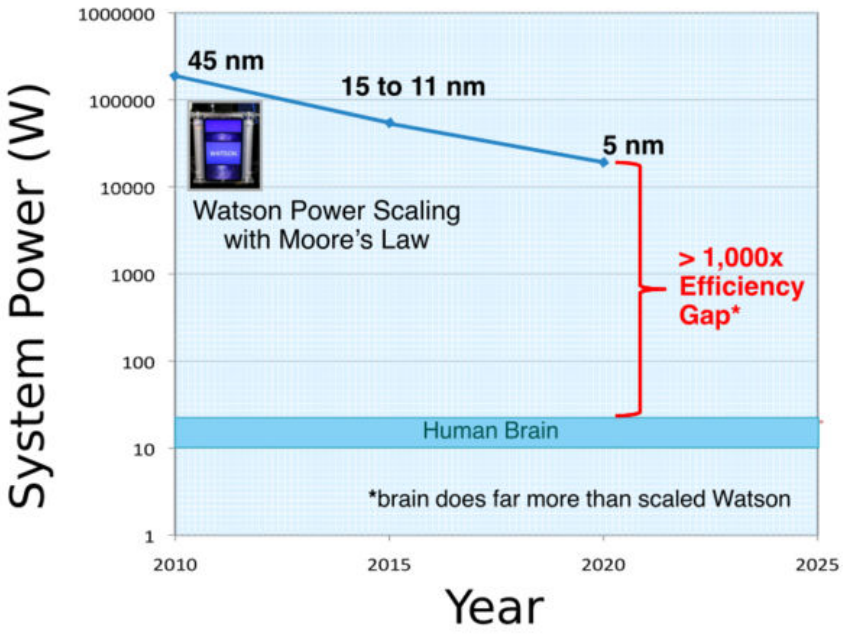


Figure 1.10. Comparison between power consumption of extremely scaled IBM Watson supercomputer and the human brain [68].

shown in Fig. 1.10, even with extreme scaling, power consumption associated to the Von Neumann computing architecture is orders of magnitude larger than the power required by the human brain. The invention of new architectures is thus required in order to face this challenge, bridging the gap of efficiency that exists between conventional computing architectures and the human brain. In the quest for more efficient computation, neuromorphic hardware has been proposed as a new generation of computing systems, with a complementary role with respect to Von Neumann machines (Fig. 1.11).

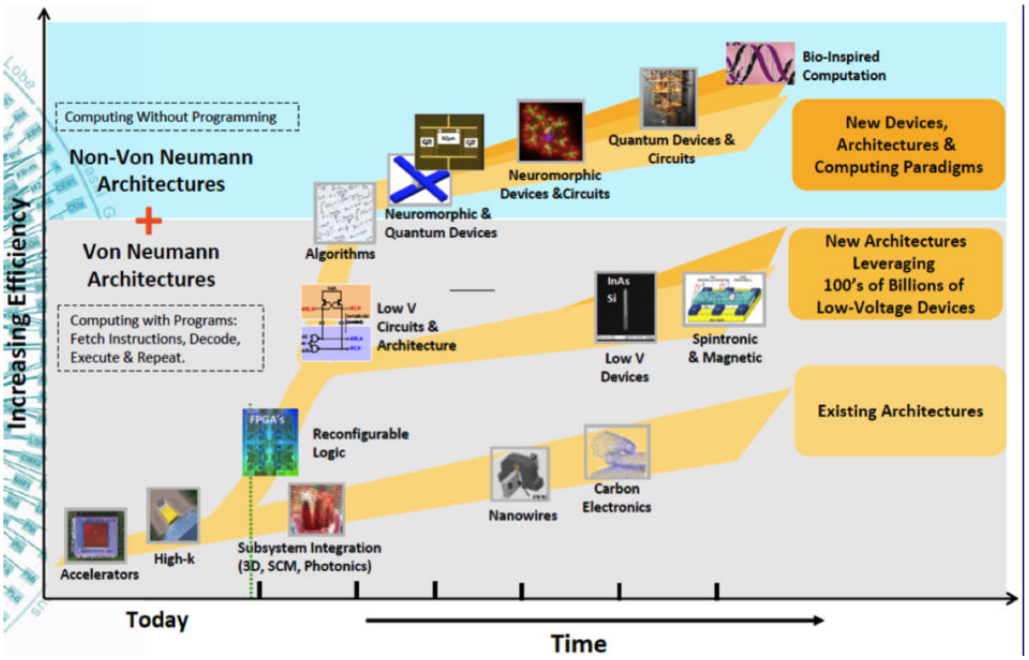


Figure 1.11. Proposed future computing roadmap with emerging non-Von Neumann architectures [68].

Historically, the interest on neuromorphic computing systems originated in the 1940s, with the presentation of the computational model for neural networks developed by McCulloch and Pitts [69]. In the late 1940s Hebb made the hypothesis that brain plasticity is at the basis of the human learning mechanism [70]. Researchers started applying these concepts to computational models in 1948 with Turing's B-type machines [71]. In 1957, Rosenblatt developed the perceptron algorithm for image recognition [72], implemented in hardware as the "Mark I Perceptron" or the first neuromorphic machine. Over the next years, the field was relatively stagnant because of the limitations of computational machines that processed neural networks [73]. The emergence of greater computational efficiency, together with advances with the backpropagation algorithm [74], revived the research activity in neuromorphic computing. During the 1980s, parallel distributed processing systems started to be adopted to simulate large-scale neural networks [75]. Mead introduced VLSI design concepts for the design of bio-inspired systems [76], with the design of the first silicon artificial retina and neuro-inspired learning chips on silicon. Neurocomputers, i.e. dedicated hardware implementations of processors specialized for neural computations, emerged in the period from 1980s to early 1990s. The ZISC (Zero Instruction Set Computing) processor [77] was proposed by IBM. The ETANN (Electrically Trainable Artificial Neural Network) chip, featuring 10240 floating-gate synapses was presented by Intel [78]. Other examples of neurocomputers from that period are L-Neuro by Philips, ANNA by AT& T, SYNAPSE 1 by Siemens [79], and

MIND-1024 by CEA [80]. Research advancements in neuroscience during the 1990s, particularly the interest in synaptic plasticity [81] and unsupervised learning rules like spike timing dependent plasticity (STDP) [82] represented a turning point in the field [83]. The progress in the field of emerging non-volatile resistive memory technologies brought new life to research in neuromorphic hardware in the 2000s.

In the next section, we will briefly discuss the characteristics of biological neural networks, composed of neurons and synapse. This is useful to understand which characteristics have to be emulated in order to efficiently implement in hardware a bio-inspired architecture.

1.3.1 Neurons and Synapses

The human brain is composed by a large number of interconnected networks, where the fundamental building blocks are neurons and synapses. It is estimated that in the human brain there are about 10^{11} neurons, and 10^{15} synapses [84]. Neural networks perform different intelligent functions inside the brain such as perception of stimuli, recognition, movement, speech.

The neuron is an electrically excitable cell that processes and transmits information through electrical signals. Neurons are connected to each other via the synapses, to form neural networks. The signals that are exchanged between neurons are called action potentials or spikes. As shown in Fig. 1.12, a neuron consists of three main parts: the dendrites, the soma and the axon.

- The dendrites are the input vectors through which signals are received. The dendrites allow the cell to receive signals from a large (>1000) number of neighboring neurons.
- The main body of the neuron is the soma. It performs an integrate-and-fire function: as positive and negative signals (exciting and inhibiting, respectively) reach the soma from the dendrites, the membrane voltage of the cell is affected.
- Once the membrane voltage of the soma reaches a certain threshold value, the neuron produces a spike wich is transmitted along the axon to all other connected neurons dendrites.

The synapse is the connecting unit between the axon of a presynaptic neuron (pre-neuron, i.e. the neuron that is sending a spike), and a post-synaptic neuron (post-neuron, i.e. the neuron which is receiving the signal). In a synapse, the voltage spike of the presynaptic neuron activates the voltage-controlled calcium channels present in the presynaptic membrane. The rapid influx of Ca^{2+} into the presynapse triggers the release of chemical substances (the neurotransmitters) in the synaptic cleft. Neurotransmitters bind to receptors located on the membrane of the postsynaptic cell. The binding of the transmitters to the postsynaptic receptors causes ionic

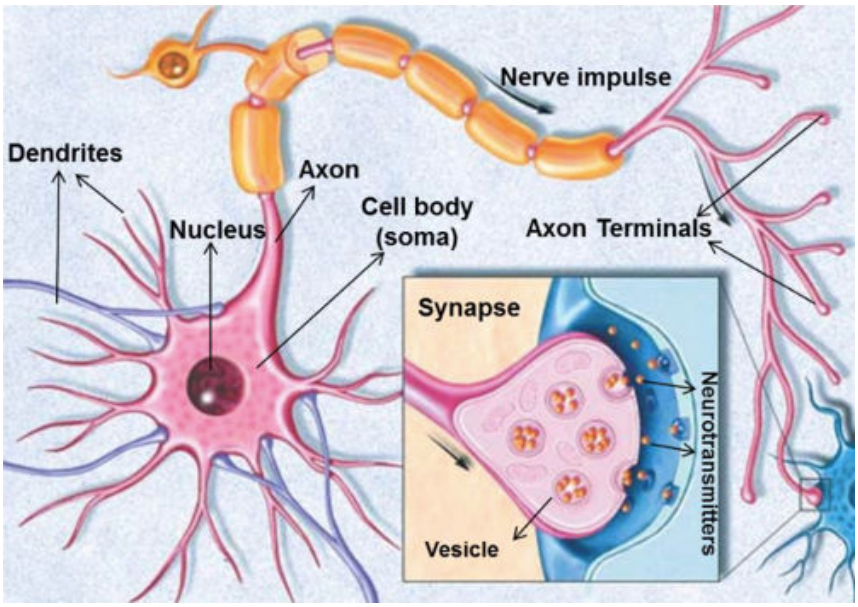


Figure 1.12. Schematic view of the basic structure of a neuron cell. Inset shows a zoom of the biological synapse. Source: [85].

channels to open or close, thus changing the ability of ions to flow into or out of the postsynaptic neuron. The selective permeability of these channels allow ions to move along their electrochemical gradient, inducing a ionic current that changes the membrane potential of the neuron (Post-Synaptic Potential, PSP). The change can be positive (Excitatory Post-Synaptic Potential, EPSP) or negative (Inhibitory Post-Synaptic Potential, IPSP).

In the post-neuron, all ionic currents incoming from multiple synapses are summed over time and when a threshold potential is reached, an action potential (spike) is generated and sent along the axon. After that, the membrane potential of the neuron goes back to the resting threshold potential. An important characteristics of synapses is the fact that they are *plastic*: their weight, i.e. the efficiency in the transmission of signals through the synapse, changes over time according to the relative timing at which pre-synaptic and post-synaptic spikes occur. This plasticity, which is described in more detail in Section 1.3.5, is a key factor in learning and remembering.

One of the simplest artificial neuron model is, the Integrate and Fire (IF) neuron model. Figure 1.13 shows the concept of a simple IF neuron. It sums over time (integrates) the incoming excitatory and inhibitory signals inside the neuron integration block using a capacitor. More advanced designs also work with this principle [86]. This integration leads to an increase in the membrane potential of the neuron V_{mem}. When V_{mem} reaches the threshold value V_{th}, the neuron generates an output spike. After the neuron has fired the membrane potential is restored to a resting value, by

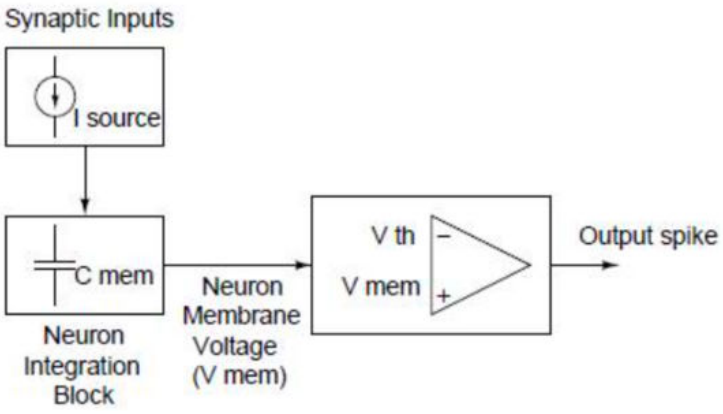


Figure 1.13. Schematic image shown the basic concept of an Integrate and Fire neuron [86].

discharging the the capacitor C_{mem}.

Many designs for the hardware implementation of artificial neurons on silicon, based on standard VLSI CMOS technology, have been proposed in the literature [86]. Research activity is being carried out to optimize power and area efficiency of neurons. Some example feature the use of non-volatile memories [87].

However, given the fact that the number of synapses is about 4 orders of magnitude larger than the number of neurons, the real challenge is to find an efficient design for the synapse, in order to be able to integrate large-scale neural networks on chip. The hardware implementation of artificial synapses is discussed in the next section.

1.3.2 Non-volatile memory devices as artificial synapses

Multiple solutions to implement artificial synapses using available VLSI devices such as Flash, DRAM and SRAM have been proposed in the literature [66]. These approaches have the advantage of relying on already available standardized design tools and a mature fabrication process. However, some limitations exist with this approach [65]. Flash devices are not an ideal candidate for the implementation of bio-inspired learning rules because they are 3-terminal devices, while real synapses are 2-terminal. During synaptic learning individual synapses may undergo weight modification asynchronously, which is not straightforward to achieve with the addressing schemes required for Flash arrays. Flash devices also require high operating voltages in order to program the cell. In many cases, complex pre-synaptic circuitry is required to implement timing dependent learning rules. This is necessary in the case of NOR Flash, because of the differences in the physics involved in the writing and erasing of the floating gate devices, but not in the case of NAND Flash. Furthermore, Flash endurance is limited, which implies a limited amount of learning operations can occur. Synapses based on DRAM technology are volatile and require refresh cycles

to retain the synaptic weight, since the information is stored as charge accumulated on a capacitor. Typically the implementation of learning rules based on DRAM synapse requires more than 10 additional transistors [88], [89]. The capacitor element itself is also area-consuming. The SRAM based synapses are affected even worse by the problem of large area consumption and are also volatile. When the network is turned off, the synaptic weights stored in SRAM are lost, so they need to be stored to a nonvolatile memory unit during or after the learning, which leads to additional power and area consumption. The limitations of available VLSI technologies for the implementation of artificial synapses provided the motivation for research in synaptic emulation using emerging non-volatile memory technologies. Recent research in nanoscale devices and materials has demonstrated the possibility of emulating the behavior of real synapses in artificial neural networks, and in particular to reproduce their plasticity and non-volatility characteristics [66], [90]–[104]. The basic idea behind this approach is to emulate the behavior of the synapse, which is a communication channel featuring variable efficiency, as a tunable resistor, implemented with a non-volatile memory (NVM) device. Some advantages of using emerging NVMs as artificial synapses are low-cost, full CMOS compatibility, high density, low-power consumption, high endurance, high temperature retention [28], [48]. NVM devices are 2-terminal, as in the case of real synapses, and offer the possibility of 3D integration.

Two main device categories can be identified for the implementation of artificial synapses: multilevel and binary devices.

Multilevel

In the multilevel (or analog) approach, the possibility of programming individual NVMs at multiple resistance levels is exploited. Some examples feature the use of OxRAM and CBRAM devices, where multilevel resistance levels are obtained by tuning compliance current during SET operation, or modulating the applied voltage [96]. However, this implementation is not ideal from a practical perspective. It requires the adoption of complicated neuron spike shapes [90], or the generation of spikes with increasing amplitude while keeping a history of the previous state of the synaptic device, leading to additional overhead in the neuron circuitry.

A better candidate for the multilevel approach is PCRAM technology, which offers the possibility of gradually increasing the conductance of the device by applying identical SET pulses, gradually increasing the size of the crystalline region in the active phase-change material. However, the reset process is not gradual but abrupt.

This led to the proposition of the use of two PCRAM devices per synapse, in the *2-PCM* approach proposed in [101] and recently adopted in [104]. These two devices are connected in a complementary configuration, where each device has an opposite contribution to the neuron's integration. When the equivalent synapse needs to be potentiated, the Long Term Potentiation (LTP) PCRAM device is partially crystallized with a weak SET operation. This increases the equivalent

weight of the synapse. On the contrary, when the synapse must be depressed, it is the Long Term Depression (LTD) PCM device that undergoes partial crystallization. Since the contribution of the to the neuron's integration is negative, the equivalent weight of the synapse is decreased. With this solution, since gradual crystallization is achieved with successive identical voltage pulses, the pulse generation scheme is greatly simplified. However, a systematic refresh scheme is needed to reset the devices by retaining the weight of the synapse.

Binary

With the binary approach, only two resistance levels per NVM device are used: the low and high resistance states (LRS and HRS). The advantage of this approach resides in the fact that it relies on programming schemes that are by all means similar to the ones used for conventional memory applications. Since only two sates of the device are exploited, simple SET and RESET pulses are required, optimized for speed and power consumption. It has been demonstrated that for some applications, such as the detection of cars driving in different lanes of a motorway [105], a single device associated to a stochastic learning rule (Section 1.3.5), is enough to achieve detection rates comparable to the ones obtained with multilevel synapses. In Chapter 4 we will illustrate how, by connecting n devices in parallel, it is possible to obtain a multilevel conductance behavior using binary devices. Since parallel conductance sum up, the conductance of the equivalent synapses ranges from the sum of the n conductance in the HRS to the sum of all the n conductance in the LRS. The use of multiple devices is necessary for applications that are more complex than detection, such as visual pattern recognition. This strategy comes at the cost of an increased number of devices needed to build a synapse. The binary approach offers the advantage of a simple programming methodology for the NVM devices, in which standard SET and RESET pulses, optimized for high endurance and low-power consumption, are used to switch the device resistance from LRS to HRS and vice versa.

1.3.3 Fully connected neural networks

The artificial synapses described in Section 1.3.2 have been proposed in the literature for the implementation of artificial neural networks composed of CMOS neurons and NVM-based synapses [66], [90]–[104]. The network topology that has been mostly investigated in the literature is the fully connected neural network. In this neural network topology, neurons are organized in layers. The first neuron layer is connected to the input of the network, while the last neuron layer represents the output of the system. The neuron layers between input and output are generally referred to as *hidden* layers. For a system designed for visual applications such as pattern detection or recognition, the raw data to be processed can be a static picture [96] or a video [101]. For auditory application the raw data is sound [105]. The raw

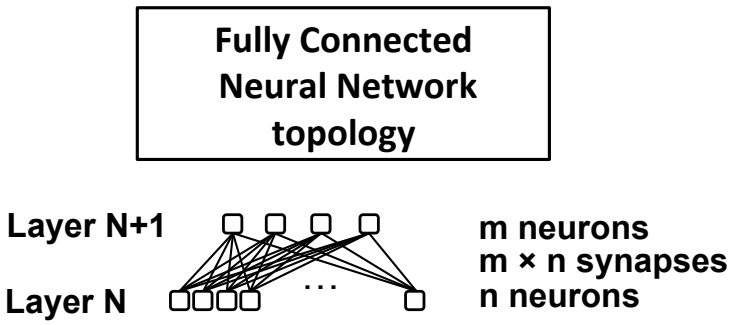


Figure 1.14. Fully connected neural network topology. Each neuron is connected to every neuron of the upper layer by a large number of synapses.

data is converted into voltage spikes with a given encoding rule, and fed as an input to the network. The conversion of the data to "spike language" understandable by the network can be implemented with a simple algorithm, such as linear conversion from pixel brightness to spike frequency. It can also be obtained with bio-inspired sensors such as artificial retina [106] or cochlea [107], or even electroencephalography (EEG) recording [95]. In the fully connected neural network topology, each neuron is connected to every neuron of the next layer as shown in Fig. 1.14. The spike signals propagate through from input to output through the hidden layers of the network, undergoing a transformation that is defined by the weight of the synapses. The output neuron layer can be composed of a single or multiple neurons. A single neuron is used if the network is used to detect a pattern in time. An example is detecting a specific sound pattern hidden by white noise [105]. Multiple output neurons can be used to perform a classification task, such as the classification of the sound of different vowels [95], the orientations of a segment [96] or the shape of simple visual patterns [100]. Figure 1.15 shows an example of a neuromorphic system with fully connected topology and multiple output neurons. The input of the network is connected to a bio-inspired artificial retina sensor, using Address-Event Representation (AER) data. The artificial retina records a video of cars passing on different lanes of a motorway. The neuromorphic system is used to detect cars passing on different lanes: when a car is driving on a given lane, a corresponding output neuron is activated. This allows to extract information about when and which lane a car is driving on.

1.3.4 Convolutional neural networks

Fully connected neural network topologies are often limited to a maximum number of hidden layers equal to one or two. Further increasing the number of layers explosively increases the complexity of the network and the number of required synapses, without necessarily improving the performance of the network for pattern

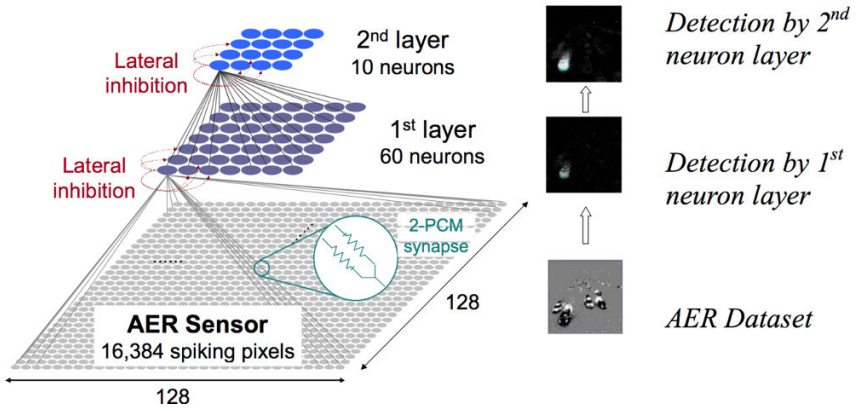


Figure 1.15. Fully connected neuromorphic system for visual pattern extraction from video of cars driving on different lanes [101].

recognition applications. Convolutional Neural Networks (CNNs), often referred to as deep neural networks, are composed of a cascade of many layers. The first layers of a CNN are convolutional layers, with a topology schematized in Fig. 1.16. Neurons of a convolutional layer are organized in feature maps. Each neuron of a feature map is connected to a small subset of neurons (receptive field) of the previous layer. A small set of synapses (kernel, or filter bank) is shared among different neurons to connect layer N and $N+1$ through a convolution operation. Figure 1.17 illustrates the operation of convolution where layer N is a handwritten digit 4, and the kernel feature is a diagonal edge. The kernel corresponds to a feature that has to be localized in the input image. A peak in the convolution signal means that the feature is present in the input pattern, and the feature map indicates where the feature is present in the input field. At each convolutional layer, the input pattern undergoes a transformation to a higher, more abstract representation. In the case of image recognition applications, for examples, the learned kernel features in the first convolutional layer typically represent simple edges or segments with a given orientation. The features of the second layer typically represent particular arrangements of edges in more complex shapes. The kernel features of the next layer layer may represent more complex combinations that correspond to parts of objects. After the convolutional layers, a classifier with fully connected topology is used to classify objects as combinations of the different parts extracted by the previous convolutional layers.

CNNs are based on the property that many natural signals feature a hierarchic structure, where higher-level complex features are a composition of lower-level simple ones. In the examples of visual images, local combinations of edges are arranged into motifs, motifs are arranged into parts composing different objects. Similar hierararchic structures exist in natural speech signal, where different sounds compose phones, which in turn form phonemes, then syllables, then words and finally full

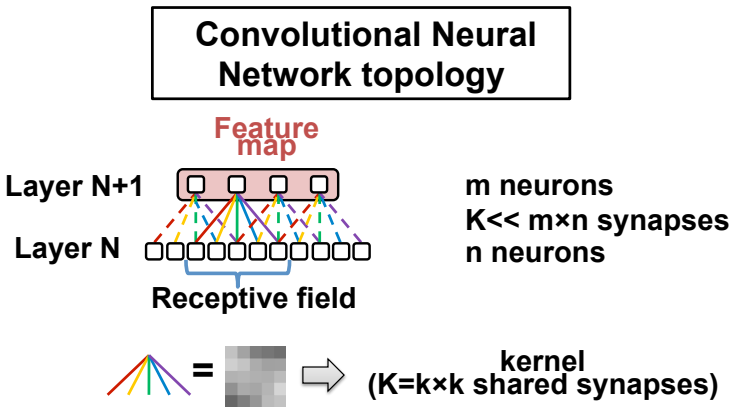


Figure 1.16. Convolutional neural network topology. A small set of synapses (kernel) is shared among different neurons to connect layer N and $N+1$ through a convolution operation.

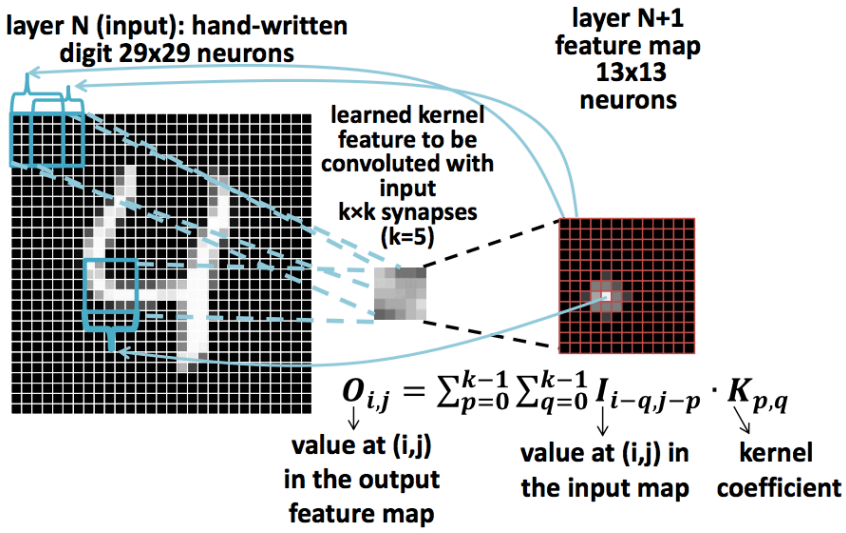


Figure 1.17. Schematic illustration of the convolution operation between an input image representing the handwritten digit "4" and a kernel feature representing a diagonal edge. The resulting feature map holds information about where the kernel feature is present in the input image.

complex sentences [108]. The organization of convolutional layers in CNNs are inspired by the complex cells in visual neuroscience [109], and the CNN hierarchic structure is inspired by the neuronal hierarchy in the visual cortex [110].

Software implementations of CNNs were originally developed in the early 1990s and used for applications such as speech recognition [111] and document reading [112]. Since the early 2000s CNNs have been applied with great success in applications such as traffic sign recognition [113], the analysis of biological images [114], and the detection of faces, complex text, pedestrians on the streets and human bodies in

natural images [115]–[120]. A major recent practical success of software implementations of CNNs is the face recognition software proposed by Facebook, which is based able to match human performance in people's faces recognition tasks [121].

Hardware implementations of CNNs, exploiting the energy efficiency of NVMs as discussed in Chapter 4, would open the way to advanced complex pattern recognition in smart and portable devices, where low power consumption is a crucial factor to take into account.

1.3.5 Learning

In the previous sections, the concept of synapses as connections with tunable weight in artificial neural networks has been introduced. In this section, we will introduce the concept of learning, i.e. how the weight of each synapse of a neural network is defined in the network, starting from the input stimuli.

Supervised learning

In the framework of artificial neural networks, the most common form of machine learning is supervised [108]. In supervised training, the network learns from data but an external supervision is needed, in the form of a labeled training data set, to guide the learning process towards correct results. The backpropagation algorithm is one of the most used supervised learning algorithm. In order to explain this algorithm, let's consider a hypothetical system designed for image classification, where images of objects have to be classified into n different categories. At the initial state, the network is untrained, i.e. the synaptic weights are random. In order to determine the good synaptic weights, first a training data set has to be collected. It consists of a large set of examples, where each object image is labeled with its corresponding category. Then the network undergoes training, which is done in software simulation. A flowchart illustrating the algorithm that is executed is presented in Fig. 1.18. At every step of the training (training epoch), the machine is shown one of the images of the training data set and produces an output. The output of the network is in the form of neuron activity of the n output neurons, one for each object category. It is desired that only the output neuron associated to the category of the image that we show is activated. However, this is unlikely to happen before training, because we still need to "teach" to the network which object belongs to which category. In order to quantify the goodness of the network output, the *objective function* δ is determined as schematically represented in Fig. 1.19a. It measures the error, i.e. the distance between the actual network output y and the desired output pattern z. The internal adjustable parameters of the machine are then adjusted to reduce this error. These adjustable parameters are the synaptic weights, which define the transformation performed by the machine from input to output. In a typical deep-learning system such as a convolutional neural network, there may be tens or hundreds of millions

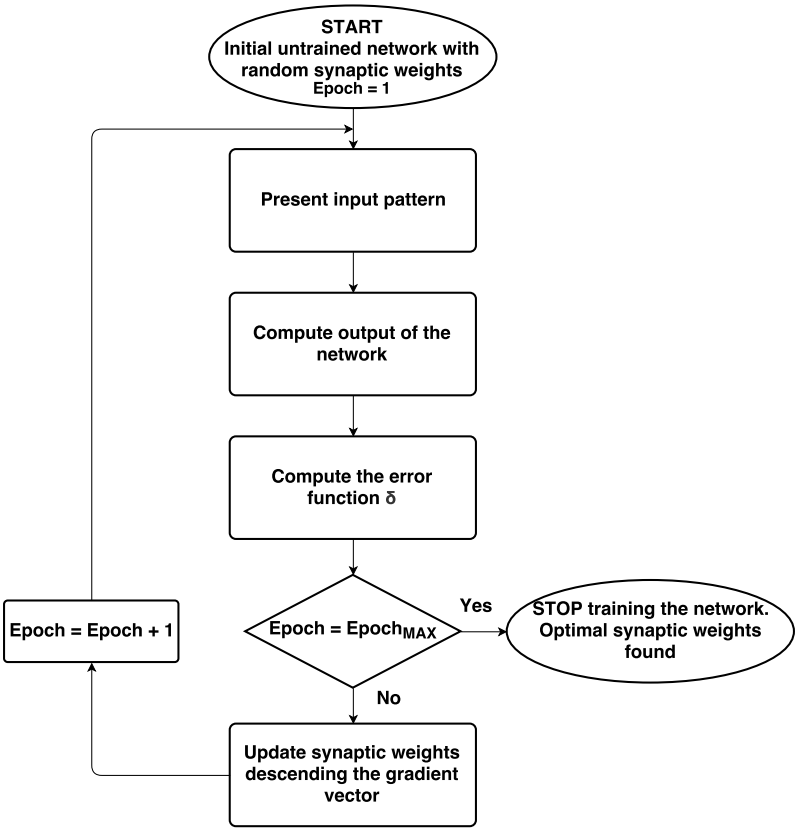
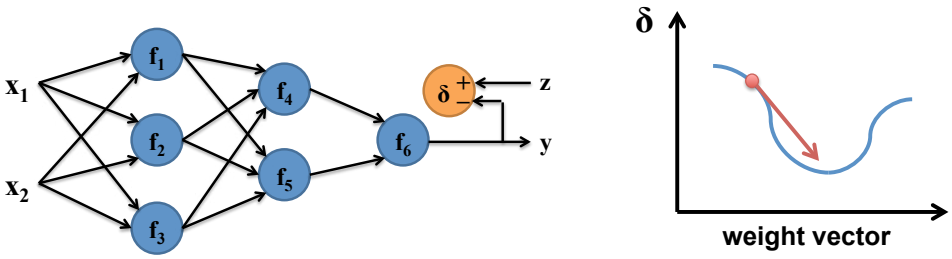


Figure 1.18. Flowchart of the supervised learning algorithm.

of these synaptic weights. The set of synaptic weights of the system is called the weight vector. To properly adjust the weight vector, the backpropagation algorithm computes a gradient vector. The gradient vector describes, for each weight, by how much the error would decrease (or increase) if the synaptic weights were changed by a tiny amount. The key insight of the backpropagation algorithm is that gradient vector of the objective function can be computed backwards from the output to the input of the network. The weight vector is then adjusted in the opposite direction to the gradient vector. The objective function is a complex function in the multidimensional weight space, featuring multiple local minima and maxima. Figure 1.19b gives a schematic example of an objective function δ as a function of a one-dimensional weight vector. Thanks to the gradient vector we can follow the direction of steepest descent in this multidimensional function, towards a global minimum where the error is the lowest on average [108].

The optimal weight vector determined with backpropagation, which makes sure that the network performs the best on the training data set, is also the one that most likely will perform well on new images that do not belonging to the training data set, with images that the network didn't "see" before, in a process which is



(a) (b)

Figure 1.19. (a) Schematic representation of the computation of the objective function δ as the distance between the desired output of the network z and the actual network output z. Adapted from [122]. (b) Illustration of the gradient descent process, performed to minimize the error of the network thanks to the backpropagation algorithm.

called generalization. Once the weight vector is defined with training in computer simulation, it is possible to import the weight in the artificial neural network. In the case of an artificial neural network where the synapses are implemented with NVM devices, each device is programmed to the resistance level determined by the backpropagation learning algorithm.

Unsupervised learning

While supervised learning offers the possibility to achieve excellent performance, similar to the human one, even on very complex recognition tasks such as face recognition [121], its main limitation is the need of a labeled training data set, which can require hundreds of millions of elements in order to achieve excellent performance [108]. If we take inspiration from human learning, we realize that it is largely unsupervised: the structure of the world is learned by observing it, not by using a huge database of labeled examples, as it happens in supervised learning. Even though human learning is still relatively obscure and still object of research, the biological process known as Spike Timing Dependent Plasticity (STDP) is widely believed to play a key role in learning and storing information in the brain [82].

Spike timing dependent plasticity (STDP) is a biological process or learning-rule that changes and adjusts the weight of each synapse based on the time difference between the spiking of post- and pre-synaptic neurons. According to STDP, if the post-synaptic neuron spikes right after the pre-synaptic neuron, the synapse is potentiated, i.e. its weight increases (Long Term Potentiation, LTP). On the contrary, if the post-synaptic neuron spikes before the pre-synaptic neuron, the synaptic connection is depressed or weakened (Long Term Depression, LTD) [123]. Fig. 1.20 shows the experimentally observed STDP rule in cultured hippocampus neuron cells [81]. As a consequence of this learning rule, the synaptic weights in the

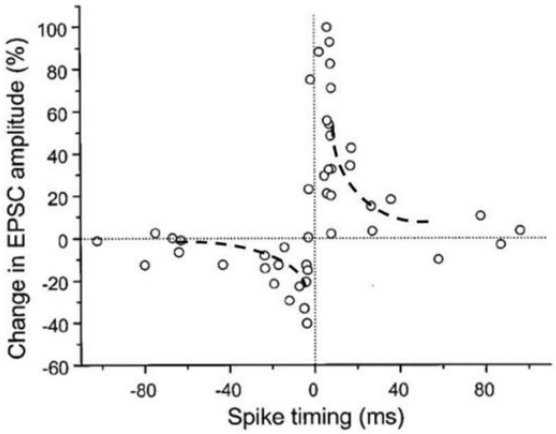
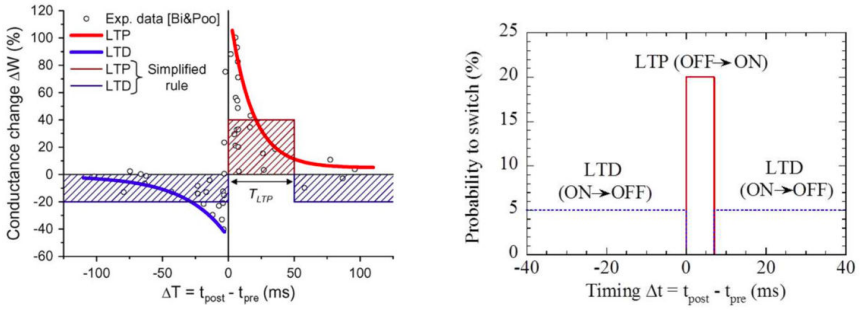


Figure 1.20. Experimentally observed STDP rule in cultured hippocampus neurons. Change in Excitatory Post Synaptic Current (EPSC) amplitude is indicative of change in synaptic strength or conductance. A spike timing Δt is defined as time difference between spikes from the post-synaptic neuron and the pre-synaptic neuron, $\Delta t = t_{\text{post}} - t_{\text{pre}}$. $\Delta t < 0$ implies LTD while $\Delta t > 0$ implies LTP [81].

brain change over time, and the brain rewires itself with a learning rule that conveys a concept of causality: inputs that might be the cause of the post-synaptic neuron's excitation are made even more likely to contribute to neuron spiking in the future, whereas inputs that are not the cause of the post-synaptic spikes are made less likely to participate in the future.

Unsupervised STDP-inspired learning rules have been successfully proposed in neuromorphic systems with artificial NVM synapses. In some examples, STDP is obtained by the overlapping of complex spike shapes [90]. While this approach closely resembles what is observed in biology, it often requires complicated CMOS circuitry to obtain such complex spike shapes. This approach thus comes at the cost of larger area and power consumption.

A simplified deterministic learning rule, shown in Fig. 1.21a, has been proposed in [101] with multilevel PCRAM synapses. It features the advantage of relying on simple spike design. A stochastic version of STDP associated to binary NVMs such as CBRAM, OxRAM and STT-MRAM has been also proposed in [105], [124]–[126]. It is shown in Fig. 1.21b. It is based on a functional equivalence [127] that exists between multilevel deterministic synapses and binary probabilistic synapses. When a LTP or LTD occurs, instead of partially changing the conductance of the the synapse, stochastic STDP specifies probability $p < 1$ of changing it totally. And if several NVMs are connected in parallel, a multibit synapse can be emulated.



(a) (b)

Figure 1.21. (a) Deterministic [101] and (b) stochastic [105] simplified STDP learning rule.

1.3.6 Applications

Bio-inspired computing systems and neuromorphic hardware has a very wide set of potential applications. Software based artificial neural networks are already being used efficiently in fields such as image classification, pattern extraction, face recognition, machine learning, machine vision, self-driving cars, robotics, optimization, prediction, natural language processing (NLP) and data-mining [128]–[134]. Analysis of big-data, web searches, data-center applications, and smart autonomous systems are new emerging fields where neuromorphic hardware can play a significant role [63]. Neuromorphic concepts are also being explored for defense and security applications such as autonomous navigation, unmanned aerial vehicles, cryptography [135]. Neuromorphic hardware have also health-care related applications such as future generation prosthetics and brain-machine interfaces [136]. The hardware implementation of deep convolutional neural network can open the way to power efficient recognition tasks, such as predicting the activity of potential drug molecules [137], analysing data from particle accelerator [138], [139], analysing brain circuits [140], and predicting the effects of DNA mutations DNA on gene expression and disease [141], [142]. Deep convolutional neural networks have also achieved very promising results in natural language understanding [143], particularly topic classification, sentiment analysis, question answering [144] and language translation [145], [146].

1.4 Conclusion

Bio-inspired computing systems where artificial neural networks are emulated in software or implemented in hardware with traditional von Neumann architectures, such as Digital Signal Processors (DSPs), Graphic processing Units (GPUs) and Field-Programmable Gate Arrays (FPGAs), have have shown strong limitations in terms of power consumption, scalability and reconfigurability [147]. The true

potential of bio-inspired systems can be realized with the implementation on optimized special purpose hardware which can provide direct one-to-one mapping with the learning algorithms running on it [67]. Emerging resistive memory technologies (as PCM, OxRAM...) are expected to change not only the conventional Von Neumann memory hierarchy. They will also play a key role in the hardware implementation of neuromorphic systems, thanks to their exceptional properties of density, scalability, nonvolatility and low power consumption.

Chapter 2

Neuromorphic Systems based on PCRAM synapses

In this chapter, we investigate the use of phase-change random access memory (PCRAM, or PCM) devices as artificial synapses. After considering the pro's and con's for the adoption of a multilevel synapse approach, we propose the use of PCM device as binary synapse as a simple and efficient solution. We test the functionality of the proposed approach through large-scale neural network simulations. The use of the proposed binary PCM synapse is studied in a neuromorphic system designed for complex visual pattern extraction. We explore unsupervised learning adopting a probabilistic STDP learning rule. Different PCM programming schemes for architectures with- or without-selector devices are provided. The system-level simulations show that such a system can solve a complex real-life video processing problem (vehicle counting) with high recognition rate (>94%) and low power consumption. We also study the impact of the resistance window on the power consumption of the system during the learning phase. The problem of resistance drift in PCM devices is also addressed, and we propose a programming strategy for the mitigation of this issue.

2.1 Introduction

As discussed in Chapter 1, emerging resistive memories will not only play a key role to reduce the memory hierarchy gap, in the framework of conventional Von Neumann computing. They will also be a key enabling factor for the hardware implementation of artificial neural network.

In this chapter, we will focus on the possibility of implementing unsupervised learning using PCM devices as synapses. Among the different emerging non volatile memory technologies, PCM is one of the most mature. The main advantages of this technology are scalability, reliability and low variability [64], [148]–[152]. Among its

attractive features, the possibility of programming devices to multilevel resistance states makes PCM an appealing candidate for the realization of artificial synapses [153], [154]. For this reason, numerous research groups [101], [104], [153], [155], [156] have investigated the use of PCM devices as artificial synapses, adopting the multilevel approach introduced in Section 1.3.2. Numerous neuron spike schemes have been proposed in the literature to implement unsupervised STDP-based learning with PCM artificial synapses. In this section, we review the main programming techniques proposed in the literature.

The scheme proposed in [153] and shown in Fig. 2.1a is a multi-pulse scheme, in which the pre-synaptic spike consists of SET and RESET pulse trains with varying voltage amplitudes. The post-synaptic spike is an individual pulse which, by overlapping with one of the pulses of the pre-synaptic pulse train, has the effect of programming the device to a state that depends on the relative timing at which pre-synaptic and post-synaptic spikes have been fired. The weakness of this programming scheme is the fact that it uses unnecessary pulses in the pre-synaptic spike [64]. The majority of pulses in the pre-synaptic train is not actually used to program the synapse, because only one of them overlaps with the post-synaptic spike. These extra pulses might have the effect of disturbing the synapse and other synapses connected to the same neuron. The large quantity of unnecessary programming pulses is also related to an unwanted large power consumption [65]. This is due to the charging of interconnect metal lines, which is not negligible if the considered synapse array is large. A single programming pulse scheme, proposed in [100], [156] and illustrated in Fig. 2.1b, addresses the drawbacks of the multi-pulse scheme. This approach relies on the use of a communication signal between the pre-synaptic neuron and the post-synaptic neuron. One important characteristic of PCM devices is the fact that gradual change of conductance can be achieved using identical pulses only in one direction, i.e. from high to low resistance state (Long Term Potentiation, LTP), by gradually increasing of a tiny amount the size of the crystalline region of the phase change material (Section 1.2.1) applying identical consecutive SET pulses. The RESET process, however, is an abrupt process because it is achieved by melting the whole crystalline region and then quenching it into the amorphous phase. If one wishes to obtain a gradual Long Term Depression (LTD), i.e. gradually increase the resistance state from low to high resistance values, RESET pulses with increasing amplitude need to be used. So the single-pulse programming scheme, even if it avoids unnecessary programming pulses, still requires the generation of non-identical voltage pulses, with amplitude changing at every spike. The desired spike voltage is obtained dynamically as a function of spike timing according to the STDP rule. In order to implement this functionality in hardware, a complex neuron design has to be adopted. Each neuron's circuitry has to keep track of the spike timing, i.e. the time difference between the last time the neuron spiked and the time it received the communication signal. Once this time difference is computed, the

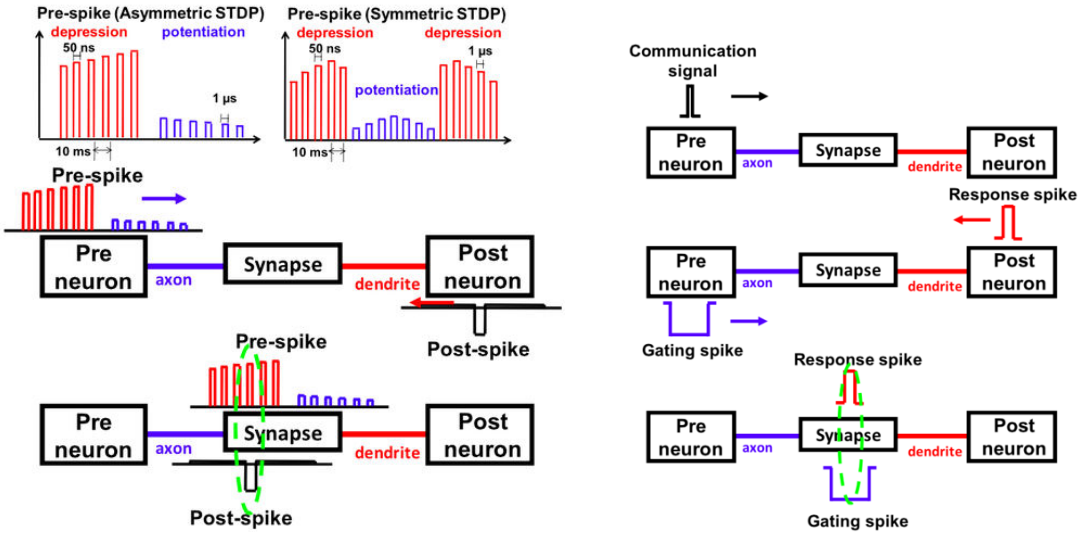


Figure 2.1. (a) Multi-pulse and (b) single-pulse programming scheme for PCM synapses. [64].

desired amplitude of the programming pulse is determined with the STDP rule and the spike is generated.

In the next subsection, we will discuss how these limitations can be overcome by adopting two devices per synapses with the "2-PCM" synapse approach, at the cost of the introduction of a refresh scheme.

2.1.1 The 2-PCM Synapse refresh scheme

The "2-PCM synapse" approach, briefly presented in Section 1.3.2, is an alternative solution proposed for the first time in [101], [157] and recently adopted in [104]. It is schematically illustrated in Fig. 2.2a. An important advantage of the "2-PCM Synapse" approach is the following: since it is based mostly on the crystallization operation of PCM devices (SET), it allows defining a programming methodology that uses identical neuron spikes, composed of single pulses, to obtain both gradual LTP and LTD, thus requiring a simpler neuron design. This advantage comes at the cost of a synaptic density reduced by a factor 2, because two PCM devices per synapse are used instead of one.

The conductance of each PCM device gradually increases during the learning, both for LTP and LTD operations. It tends to saturate towards the highest achievable conductance, in which the totality of the active phase change material is crystallized. Therefore, a *refresh* mechanism is introduced to reduce the conductance of LTP and LTD devices, while keeping the weight of the equivalent synapse unchanged. A schematic representation of the refresh operation is shown in figure Fig. 2.2b. As

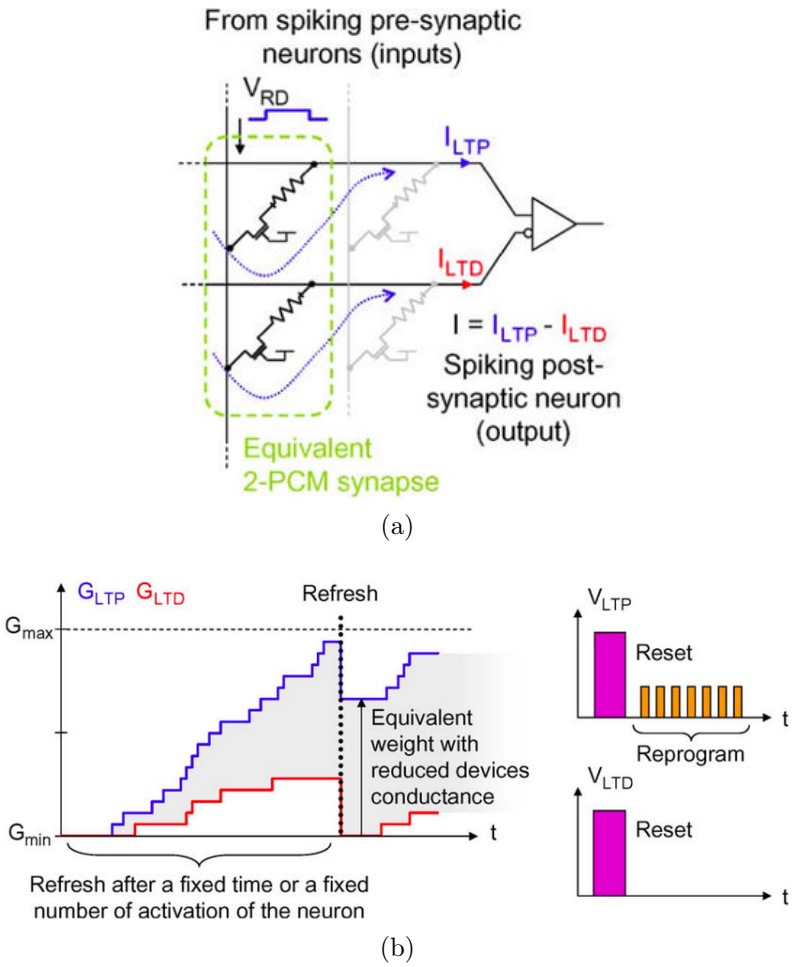
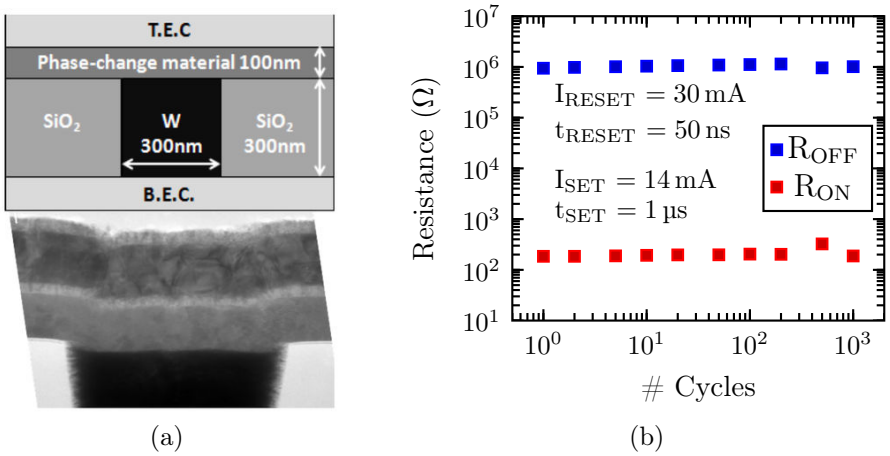


Figure 2.2. Illustration of the refresh scheme required by the 2-PCM synapse approach. Source: [157]

soon as one of the two devices reaches the maximum conductance value, a RESET operation is performed on both devices. At this point both devices are in the high resistive, amorphous state. In order to recover the equivalent synaptic weight before the refresh operation, a series of SET operations is performed on the device that had the highest conductance, until the equivalent weight is restored. One has to know the average number of conductance levels N obtainable before full crystallization of the phase-change material in use, for a given SET pulse shape (duration and amplitude). A refresh operation is scheduled after N LTP or LTD operations, using for example a simple counter. This approach does not require permanent monitoring of the state of the LTP and LTD devices [65], however this approach still requires a rather complex neuron design in order to restore the equivalent synaptic weight at each refresh operation.



(a) (b)

Figure 2.3. (a) Scheme of the studied GST PCM device (top) and cross-section TEM image (bottom) [23]. (a) Experimental results of 10^3 SET-RESET cycles.

In the next sections, we will hence discuss how PCM devices can be used in binary mode in order to simplify the programming scheme and therefore avoid the drawbacks of the multilevel approach.

2.2 PCM binary synapse

We propose here a different approach for PCM synapses, i.e. a binary probabilistic one, where only two states are used for the synaptic weights and the switching of the device between these two states is governed by a probabilistic learning rule. This study was carried out to simplify system programming by avoiding the refresh operation previously required in the 2-PCM approach, and to optimize the synaptic power consumption.

In order to extract the characteristics of PCM devices operated in binary mode and be able to model the PCM synapses in artificial neural network simulations, electrical characterization was performed. PCM devices composed of a 100 nm thick $Ge_2Sb_2Te_5$ (GST) phase-change layer were studied. The phase change material was deposited at room temperature by plasma-assisted sputtering on a cylindrical tungsten heater plug with a diameter of 300 nm. Figure 2.3a shows a Transmission Electron Microscopy (TEM) image of the studied devices. Devices could be repeatedly switched between low resistance (ON) and high resistance (OFF) states with SET and RESET programming pulses. Figure 2.3b shows the switching operation between ON and OFF states for 10^3 cycles using the programming conditions indicated in figure.

Based on the Resistance–Current (R–I) characteristic shown in Fig. 2.4, it appears

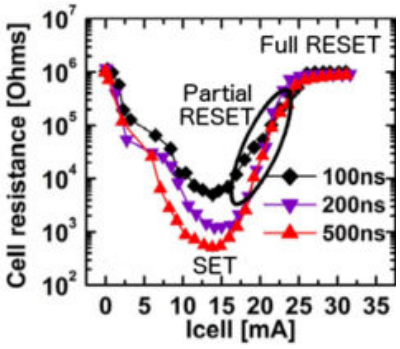
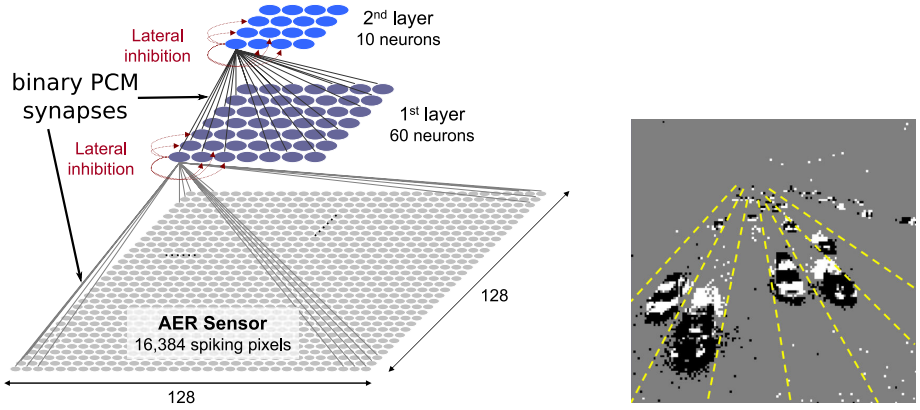


Figure 2.4. Resistance–current characteristics of GST-based PCM device.

that it is possible to program the device with different R_{ON} and R_{OFF} values. Different R_{ON} values can be achieved by using different pulse-widths for the SET operation. By tuning the RESET current I_{RESET} (i.e. by tuning the gate voltage of the selector transistor in a 1T-1R architecture) different values for R_{OFF} can be achieved. In a selector-free 1R configuration, different values for the RESET current I_{RESET} can be achieved for instance by using materials with different resistivity for the heater plug or by just tuning the voltage amplitude of the RESET pulse. In the next section, we will study how the binary device, featuring tunable SET and RESET states, is employed as artificial synapse in a neuromorphic system.

2.3 Neuromorphic Architecture

The double-layer artificial neural network illustrated in Fig. 2.5a is introduced in this section. In this neuromorphic system, binary PCM synapses are used to achieve full connectivity between CMOS neuron layers. The input of the network is composed of a bio-inspired artificial retina sensor [106]. The artificial retina is composed of 128×128 spiking pixels or neurons. The sensor working principle can be summarized as follows. The artificial retina is sensitive to the luminosity change in its visual field. Each pixel generates an event, or spike, each time the relative change of its illumination intensity exceeds a positive or a negative threshold. Therefore, depending on the sign of the change in intensity, events can be of either type ON if the sign is positive or type OFF if the sign is negative. The working principle of this device is thus different compared to standard video sensors, where the video is recorded as a temporal sequence of static images, frame by frame. The sensor is used to record a video of cars passing on six different lanes of a motorway. Figure 2.5b shows a frame extracted from the video, where the six different lanes have been highlighted. A first neuron layer composed of 60 neurons is connected with fully connected topology to the input of the network. There are two synapses per pixel, one for each event type, ON or OFF. The output of the network is composed of a



(a) (b)

Figure 2.5. (a) Schematic of the fully connected neuromorphic system studied in simulation. (b) An example of one of the frames of the input video, showing cars passing on multiple lanes of a motorway. The separation between lanes (in yellow color) has been added to illustrate the distinction among different lanes and is not present in the original input video.

second layer of 10 neurons, where each neuron is connected to each neuron of the previous layer via one PCM synapse. The total number of synapses in the studied system is thus $N_{\text{synapses}} = 2 \cdot 128 \cdot 128 \cdot 60 + 60 \cdot 10 = 1\,966\,680$.

The neuromorphic system here described is used to detect cars passing on different lanes: when a car is driving on a given lane, the corresponding output neuron is activated. This allows to extract information about when and which lane a car is driving on. In order to make sure that different neurons become sensitive to different lanes, avoiding that one neuron becomes sensitive to the full set of lanes, a particular learning strategy is adopted, i.e. competitive learning. Competitive learning is obtained by implementing lateral inhibition, i.e. when a post-synaptic neuron fires, the integration of incoming spikes in the other post-synaptic neurons of the same layer is disabled [158]. Therefore, we avoid that all neurons become sensitive to the full input frame and we make sure to differentiate the neurons sensitivity. This neuromorphic system is analogous to that proposed in [101] and described in Section 1.3.3. The difference is that in this approach, only 1 PCM device per synapse is used, and the learning rule has been adapted to the new proposed synapse. The neuromorphic system proposed here is associated to a probabilistic STDP learning rule. Its adoption is based on a functional equivalence [127] existing between a multilevel deterministic learning rule and a binary probabilistic one. This equivalence is schematically represented in Fig. 2.6: when a long term potentiation or depression occurs, instead of gradually changing the conductance of the the synapse with probability $p = 1$, probabilistic STDP has a probability $p < 1$ of switching the synaptic totally from one state to the other. A similar approach has been

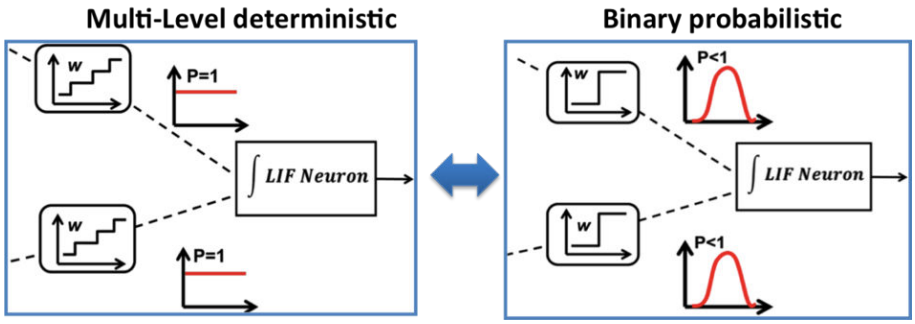


Figure 2.6. Illustration of the equivalence existing between multi-level deterministic and binary probabilistic synapses. p indicates the probability of change in conductance or switching. Adapted from [105].

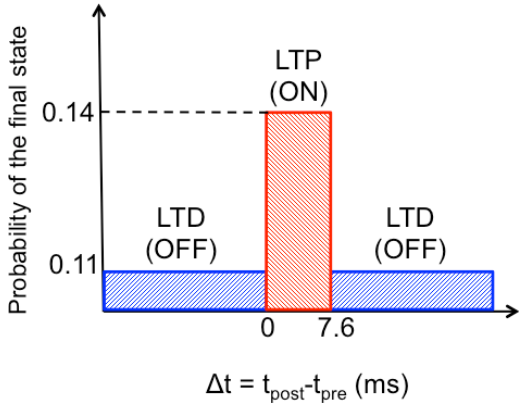


Figure 2.7. Binary probabilistic learning rule adopted for the simulations.

adopted in previous works, such as in [105], [159]–[161]. This approach is motivated by biology research [162], presenting some evidence that STDP learning could be in part a stochastic process. As shown in the learning rule scheme of Fig. 2.7, if the time difference between the spike of a post-synaptic neuron and the spike of a pre-synaptic neuron is smaller than the long term potentiation (LTP) time window t_{LTP}, then the PCM synapse has a given probability to switch to the ON state (LTP or synaptic-potentiation). Otherwise, the PCM device has a distinct given probability to switch to the OFF state (LTD or synaptic depression). Neuron parameters, including switching probability and spike timing values shown in Fig. 2.7, have been determined using the genetic evolution algorithm described in [163]. The desired switching probabilities can be enforced in the system extrinsically by using a Pseudo-Random Number Generator (PRNG) circuit, similarly to [105]. The PRNG circuit controls the probability of LTP and LTD with a 2-bit signal.

In order to implement the full connectivity of the CMOS neuron layers in the artificial neural network, we suggest two possible architectures, shown in Figs. 2.8a

and 2.8b: a matrix structure with a selector transistor for each PCM device and a selector-free crossbar structure. The crossbar (selector-free) architecture offers the highest possible integration density, however it can be only implemented if the programming conditions for the chosen PCM technology ensure that there are no unwanted device disturbs (see Section 2.3.1). On the other hand, the matrix architecture, while requiring more area for selectors, is not sensitive to disturb issues. It should be noted that the two architectures and programming schemes are valid not only for PCM technology-based synapses, but can be extended to other unipolar devices, as for example unipolar oxide-based resistive memories [164] described in Section 1.2.4.

2.3.1 Operation of the system

The operation of the proposed neuromorphic system can be classified in two different modes: learning-mode and read-mode.

Learning mode

In learning mode the synaptic programming is enabled, following the STDP rule. The spiking activity recorded with the artificial retina is presented to the network for a period of time $t_{\text{LEARNING}} = 680\,\text{s}$. The network learns from data in unsupervised fashion, and different output neurons become sensitive to the passing of car on different lanes. It is only during the learning mode that the synaptic weights is changed with SET and RESET operations.

Two different programming schemes are proposed for the two different studied architectures:

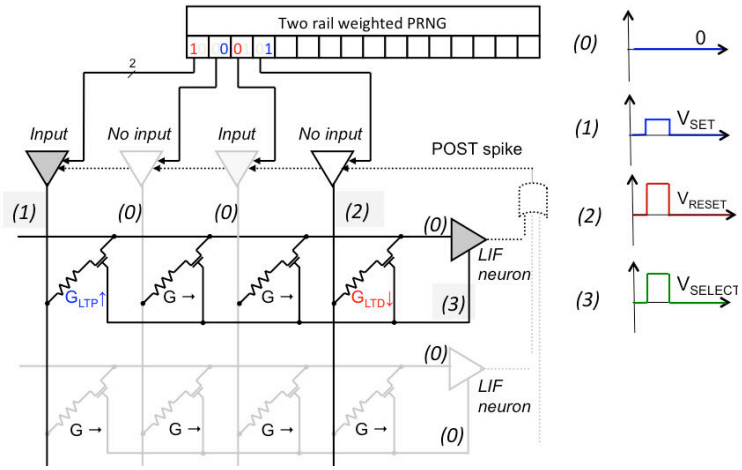
- *Programming scheme for matrix structure with selectors (Fig. 2.8a).* In the learning mode, when a post-synaptic neuron fires, all the incoming synapses are activated by means of the select transistor. Concurrently, a write-mode signal is sent to all pre-synaptic neurons so that the following happens: if a pre-synaptic neuron fired recently, i.e. during the LTP time window, a V_{SET} signal has a p_{LTP} probability to be applied to the device. On the contrary, if the last activity of the pre-synaptic neuron is outside the LTP window, there is a p_{LTD} probability for a reset pulse to be applied. In the example of Fig. 2.8a), the following scenario is represented. The post-synaptic neuron in the first row of the memory array fires. As a consequence, all the PCM devices in the first row of the matrix are selected, while all other rows are not selected. The post-synaptic sends a write-mode signal to the pre-synaptic neurons. The pre-synaptic neurons can be categorized in two categories, labeled as “Input” and “No input”. The pre-synaptic neurons marked as “Input” received an input spike in the recent past, within the LTP time-window Δt defined by the STDP

rule. On the contrary, the pre-synaptic neurons marked as "No input" did not receive any input spike within the LTP window. The "Input" neurons are thus possible candidates for performing an LTP operation. The actual decision whether an LTP operation actually occurs or not is made according to the value of the first bit of the 2-bit signal from the PRNG block. If the bit is equal to 1 (as in the case of the first column), then LTP occurs, by applying a SET voltage with amplitude V_{SET}. The conductance G of the PCM synapse thus increases. If the bit is equal to 0 (third column), no programming pulse is applied, so the conductance is unchanged. Similarly, the "No input" neurons are possible candidate for performing an LTD operation. The actual decision whether an LTD operation actually occurs or not is made according to the value of the second bit of the 2-bit signal from the PRNG block. If the bit is equal to 1 (as in the case of the fourth column), then LTD occurs, by applying a RESET voltage with amplitude V_{RESET}. The conductance G of the PCM synapse thus decreases. If the bit is equal to 0 (second column), no programming pulse is applied, so the conductance is unchanged.

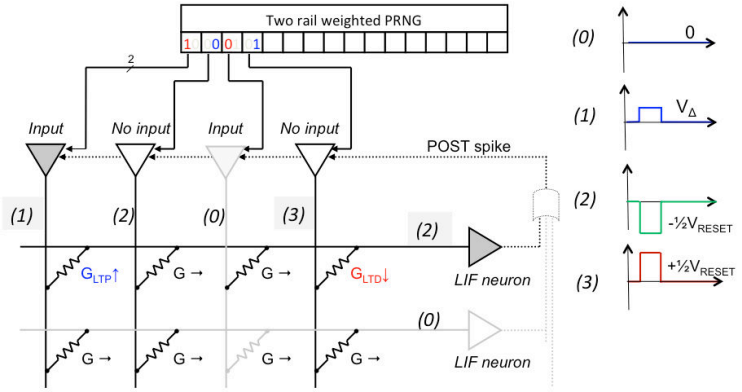
- *Programming scheme for selector-free crossbar structure (Fig. 2.8b).* Whenever a post-synaptic neuron fires, a feedback pulse $-\frac{1}{2}V_{\mathrm{RESET}}$ is fed back to all the synapses connected to it on the same row. If $\frac{1}{2}V_{\mathrm{RESET}} < V_{\mathrm{SET}}$, the signal does not affect the resistive state of the connected synapses as its amplitude is less than the programming threshold. At the same time, a write-mode signal is provided to all pre-synaptic neurons so that they will fire according to the probabilities given by the STDP rule and implemented by means of the PRNG block. If a pre-synaptic neuron was active in the LTP window, there is a p_{LTP} probability for a $V_{\Delta} = V_{\mathrm{SET}} - \frac{1}{2}V_{\mathrm{RESET}}$ signal to be fired. It will interact with the feedback signal so that the actual voltage drop across the corresponding synapse is $V_{\mathrm{SET}} = V_{\Delta} - (-\frac{1}{2}V_{\mathrm{RESET}})$ and the synapse is switched to the ON state. The amplitude of the V_{Δ} pulse on its own is not large enough to program the other connected synapses. If a pre-synaptic neuron's last activity is outside the LTP time window, its output will be a $+\frac{1}{2}V_{\mathrm{RESET}}$ pulse with a p_{LTD} probability or a $-\frac{1}{2}V_{\mathrm{RESET}}$ pulse with a $(1-p_{\mathrm{LTD}})$ probability. The positive pulse will interact with the feedback resulting in a pulse of amplitude $V_{\mathrm{RESET}} = +\frac{1}{2}V_{\mathrm{RESET}} - (-\frac{1}{2}V_{\mathrm{RESET}})$, while the negative pulse will result in a voltage drop across the device equal to 0 V, thus keeping the resistance of the cell unaltered.

Read mode

After the learning phase, the learning achieved by the network is evaluated in read mode by showing to the network an 85 s long video recording. When the circuit is in read mode, the synaptic programming is disabled. Spiking activity propagates



(a) Architecture with selector device



(b) Selector-free architecture

Figure 2.8. (a) Architecture with selector device and (b) selector-free architecture.

through the network, from the input to the output layer. SET and RESET operations are not performed. Only spikes, consisting of read pulses, occur. Whenever an input event occurs, a read voltage pulse (or spike) is applied by the pre-synaptic neurons across the PCM synapses. In the architecture with selector, all transistors are turned on by applying a select voltage on the gate terminals. The input events contribute to the current integration of the post-synaptic neurons according to the weight (i.e. the resistive state) of the synaptic connections. When the integration threshold of a post-synaptic neuron is reached, a spike is fed forward to the next neural layer or the output of the circuit.

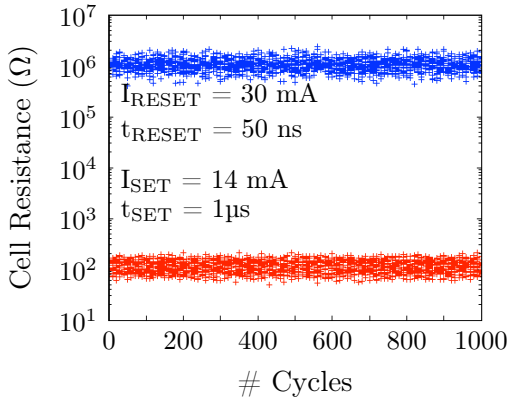


Figure 2.9. Simulated values of the binary PCM synapses resistance over 1000 cycles.

2.3.2 System performance

We tested the functionality of the proposed system with the event-driven "XNET" simulation tool for spiking neural networks presented in details in [165], [166]. The synapses were modeled by implementing lognormal distributions with mean values for R_{ON}=110 Ω, R_{OFF}=1.06 MΩ extracted from data presented in Fig. 2.3b. Figure 2.9 shows the simulated values of binary PCM devices over 1000 cycles. In all simulations a 10% device-to-device standard deviation was implemented for both resistive states. A 5% cycle-to-cycle standard deviation was also implemented for R_{ON} and a 10% cycle-to-cycle standard deviation was implemented for R_{OFF}. The activity of each output neuron is compared to a reference data set, in which the timing of the passing of cars in each lane is labeled by hand. In this way, a score or detection rate can be computed for each output neuron with respect to the reference data of each lane. The detection rate takes into account both false negatives, (i.e. missed cars) and false positives, i.e. neuron spiking even if no car passed on the considered lane at a given time. Since there are 10 output neurons and 6 lanes, the 6 best matching neuron-lane couples are selected and taken into account for the score computation. In Table 2.1 the score of the output neurons of the neuromorphic system here proposed is presented and compared to the results previously obtained with the "2-PCM" multi-level deterministic approach reported in [101]. The final sensitivity patterns of the 6 neurons that became sensitive to the 6 lanes are shown in Fig. 2.10. Cars in the first and sixth lanes, being at the edge of the visual field of the retina, activate less pixels compared to those on other lanes. The detection rate for the central lanes is equal to 95% on average, and it is comparable to the results obtained with the multilevel-deterministic approach shown in [101]. This confirms the functional equivalence between the two systems. The score obtained by the corresponding neurons is low ($< 85\%$), so it has been omitted. In the next section, we will describe how the choice of programming conditions, and the corresponding values of R_{ON} and

lane	detection rate (%)	
	2–PCM ref. [101]	Binary Probabilistic [this work]
1st	N.A.	N.A.
2nd	100	97
3rd	89	93
4th	89	94
5th	96	96
6th	N.A.	N.A.
Total synaptic power consumption	112 µW	73 µW

Table 2.1. Cars detection rate by lane as obtained with the system based on the multilevel-deterministic PCM synapses [101] and the binary probabilistic PCM synapses presented in this chapter.

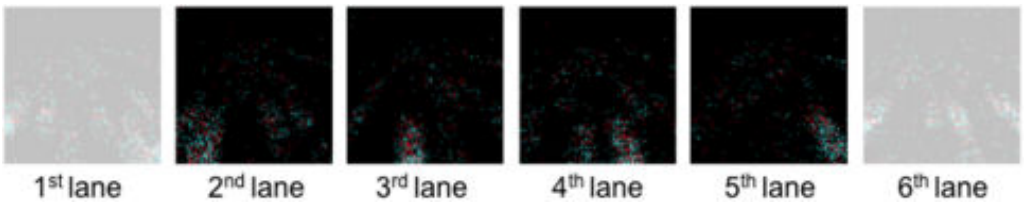


Figure 2.10. Sensitivity maps of 6 neurons at the end of learning phase. The neurons became selective to cars passing on the central lanes.

R_{OFF}, affect the power consumption of the system.

2.4 Power consumption analysis

As discussed in Section 2.2, by varying the programming conditions for the PCM synapses it is possible to obtain different values for R_{ON} and R_{OFF}. Based on this consideration, a set of parametric simulations has been carried out in order to study the impact of the PCM programming conditions on synaptic power consumption, and provide guidelines to optimize the power consumption in the studied neuromorphic system. First, a set of simulations has been performed changing the SET conditions together with the corresponding R_{ON} values, while keeping $R_{OFF} \approx 1\,M\Omega$ constant. Specifically, the SET pulse width has been varied from $t_{SET} = 100\,ns$ up to $t_{SET} = 1\,\mu s$, with corresponding R_{ON} values ranging from $R_{ON} \approx 100\,\Omega$ to $R_{ON} \approx 10\,k\Omega$. Then, a

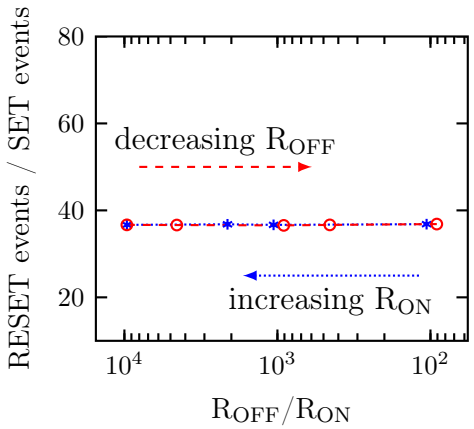


Figure 2.11. Ratio between the number of RESET and SET events as a function of the resistance window R_{OFF}/R_{ON}.

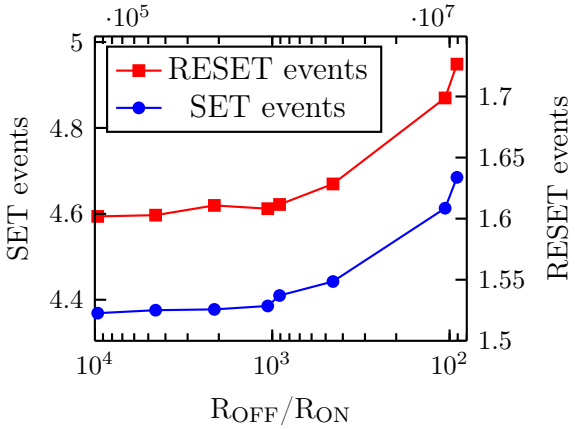


Figure 2.12. Number of SET and RESET events as functions of the resistance window R_{OFF}/R_{ON}.

set of simulations changing RESET conditions and the corresponding R_{OFF} values has been carried out. The reset current has been varied from I_{RESET}=17 mA to I_{RESET}=30 mA. In this set of simulations, the low resistance state value $R_{ON}\approx 100\,\Omega$ has been kept constant. In all simulations, the system remained functional and the average detection rate was $\geq 94\%$. In the next section the results of the parametric simulations in term of power consumption are discussed, considering both learning mode and read mode.

2.4.1 Learning mode power consumption

First, power consumption in learning mode has been evaluated. In order to do so, the number of SET and RESET events performed during the learning phase has been recorded, during the learning time $t_{LEARNING} = 680\,s$.

Section 2.4.1 shows that the ratio between the total number of RESET and SET events remains constant when the resistance window, defined as the ratio R_{OFF}/R_{ON}, changes. Indeed, this means that the programming activity is dominated by the input stimuli and the STDP learning rule. However, as shown in Section 2.4.1, the absolute number of both SET and RESET events increases when the resistance window is reduced. This can be explained with the fact that, when the resistance window is reduced, the R_{OFF} value is closer to the R_{ON} value. So, the contribution of the synapses in OFF state to the current integration at the post-synaptic neurons is larger. This means that the threshold of the integrate and fire neuron is reached more frequently. As a consequence, the number of times that neurons fire is larger, leading to an increased number of LTP and LTD events.

Once the number of SET (N_{SET}) and RESET operations (N_{RESET} is known, it is possible to estimate the power consumption associated to the SET and RESET switching events, i.e. the synaptic power consumption P_{LEARNING} during the learning time $t_{\mathrm{LEARNING}} = 680\,\mathrm{s}$, using the following approximated formulas:

$$E_{\mathrm{SET}} \approx V_{\mathrm{SET}} \cdot I_{\mathrm{SET}} \cdot t_{\mathrm{SET}} \tag{2.1}$$

$$E_{\mathrm{RESET}} \approx V_{\mathrm{RESET}} \cdot I_{\mathrm{RESET}} \cdot t_{\mathrm{RESET}} \tag{2.2}$$

$$E_{\mathrm{LEARNING}} = E_{\mathrm{SET}} \cdot N_{\mathrm{SET}} + E_{\mathrm{RESET}} \cdot N_{\mathrm{RESET}} \tag{2.3}$$

$$P_{\mathrm{LEARNING}} = \frac{E_{\mathrm{LEARNING}}}{t_{\mathrm{LEARNING}}}, \tag{2.4}$$

where V_{SET} (V_{RESET}), I_{SET} (I_{RESET}) and t_{SET} (t_{RESET}) are the voltage amplitude, current and time width associated to each SET (RESET) programming pulse.

Section 2.4.1 shows that when $\mathrm{R_{OFF}}$ is decreased while keeping $\mathrm{R_{ON}}$ fixed (red curve) it is possible to reduce the synaptic programming power during learning by 32%. This is explained by the fact that smaller current values are required to obtain smaller $\mathrm{R_{OFF}}$ values. It should be noted that the high current values considered in these simulations, in the order of tens of mA, are due to the large PCM test structures (300 nm heater plug) studied in this work. In ultra-scaled state of the art devices [29], the reset current can be reduced to µA, so giving rise to programming power in the order of nW. Weakening the SET state (increasing $\mathrm{R_{ON}}$, blue curve) does not translate into a reduction of the programming power. This is explained by two reasons: 1) the STDP rule is strongly dominated by LTD, i.e. RESET operations, rather than SET operations; 2) when the resistance window is decreased, the number of RESET events increases (see Section 2.4.1). So, the effect of weakening the SET conditions gets compensated by the increased number of RESET events.

2.4.2 Read mode power consumption

In order to compute the power consumption associated to the read mode operation, the system spiking activity has been recorded during the read-mode time, with duration $t_{\mathrm{TEST}} = 85\,\mathrm{s}$. Total read energy has been computed as the sum of the energy associated to each read pulse or spike.

$$E_{\mathrm{READ}} = \sum_i V_{\mathrm{READ}}^2 \cdot \frac{1}{R_i} \cdot t_{\mathrm{READ}} \tag{2.5}$$

$$P_{\mathrm{READ}} = \frac{E_{\mathrm{READ}}}{t_{\mathrm{TEST}}}, \tag{2.6}$$

where V_{READ} is the amplitude of the read spikes, t_{READ} is the read pulse width and R_i is the resistance of each synapse, that can be equal to either $\mathrm{R_{ON}}$ or $\mathrm{R_{OFF}}$. It should be noted that the variation of the $\mathrm{R_{ON}}$ values plays a bigger role in the

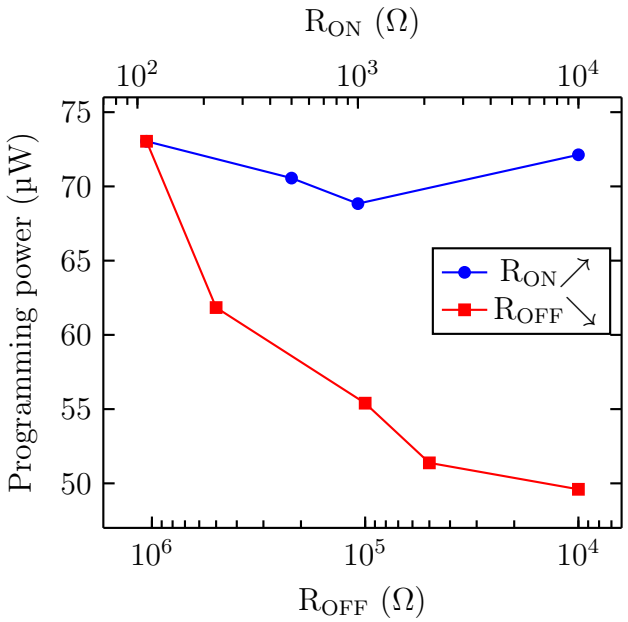


Figure 2.13. Programming power as a function of decreasing R_{OFF} (R_{ON}=110 Ω constant, red line, squares) and increasing R_{ON} (R_{OFF}=1.06 MΩ constant, blue line, circles).

determination of the power consumption in read mode, as it determines the most important contribution to the current flowing into the synapses at each read pulse since R_{ON} is orders of magnitude smaller than R_{OFF}. As shown in Section 2.4.2, increasing the R_{ON} value (blue curve), it is possible to reduce the power consumption for read operations by 99%. Variation of R_{OFF} value, on the other hand, determines a negligible variation in the read mode power consumption.

2.5 Resistance drift

As discussed in Section 1.2.1, one of the main issues of PCM technology is the resistance drift, which is a change of the device resistance value over time. Considering the devices in the RESET state, structural relaxation occurs in the amorphous phase regions of the calchogenide material [167]. As a consequence, the resistance of PCM device increases with time. Resistance drift in the crystalline SET state of GST-based PCMs is shown to be much smaller than in the amorphous phase [167], [168].

Resistance drift has a limited impact during the learning phase of a neuromorphic system. In fact during learning, because of the STDP learning implementation, spiking activity and corresponding LTP and LTD events dynamically change as a function of the resistive state of the synapses [65]. A resistance drift towards high resistance state is balanced-out by an increased LTP activity, which brings the

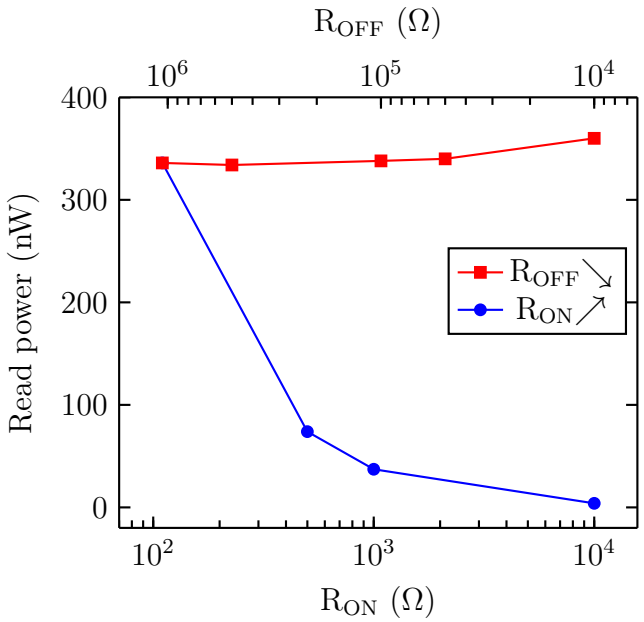


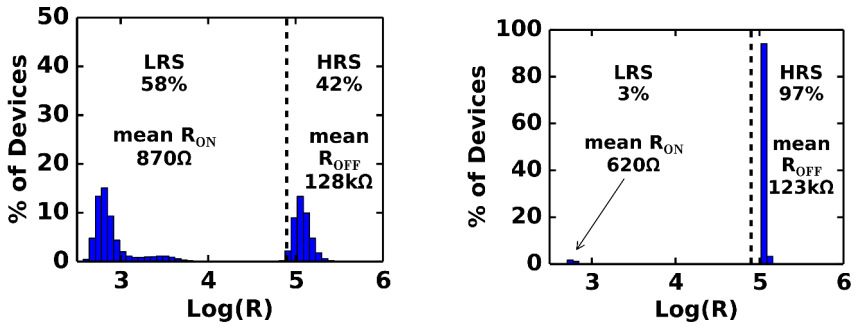
Figure 2.14. Read power as a function of decreasing R_{OFF} (R_{ON}=110 Ω constant, red line, squares) and increasing R_{ON} (R_{OFF}=1.06 MΩ constant, blue line, circles).

resistance of the devices back to the low resistance state. So the main effect of drift during learning phase is a delay in the learning time.

However, resistance drift effect might be much more detrimental in the read mode operation of the neuromorphic system, when the uncontrolled change in the resistance state of the synapses cannot be compensated by the plastic programming activity characteristic of the learning phase. For this reason, the performance of systems featuring drifting synapses might change over time, compared to the original performance of the network obtained right after the learning phase.

Based on these considerations, XNET learning simulations of the car video application discussed in the previous sections have been conducted. This study has been performed using both multilevel deterministic (2-PCM) and binary probabilistic approaches. These simulations have been carried out to evaluate how many devices, at the end of the learning phase, are in the high resistance state and thus in the resistance range affected by drift. Figure 2.15a and Fig. 2.15b show the final synaptic resistance distributions of the PCM devices, corresponding to the multilevel and binary PCM synapses, respectively.

Results show that, in the case of the 2-PCM synapse approach (Fig. 2.15a), about 60% of the devices are in the SET state, i.e. in the resistance range not affected by drift. This is due to the fact that, in the 2-PCM synapse approach, both potentiation (LTP) and depression (LTD) are obtained by SET events. This result is valid even if the learning rule is predominantly governed by LTD as in the studied case. So,



(a) (b)

Figure 2.15. Comparison of the distributions of synaptic resistance states for (a) the 2-PCM synapse approach and (b) binary PCM approach at the end of the visual learning simulation.

at the end of the learning phase, the majority of PCM devices are programmed in the low resistance range which is rather immune to the drift problematic [32], [168], [169]. For this reason, the 2-PCM synapse approach seems inherently robust to the drift problematic. However, a non-negligible amount of devices (about 40%) still lies in the region which is affected by drift. When the learning is over and the system is used in read mode, the resistance value of these synapses will change over time. This implies that the response of the network will also change over time, together with the performance of the network in read mode, which is an undesired effect.

In the case of the binary PCM synapses shown in Fig. 2.15b it appears that, at the end of the learning phase 97%, of the devices is in the RESET state, i.e. in the resistance state which is affected by drift. This result suggests that in theory the problem of drift would severely affect neuromorphic systems based on the binary PCM approach. Section 2.5.1 presents a programming strategy that we conceived in order to mitigate the problem of drift.

2.5.1 Drift mitigation strategy

In the case of binary PCM Synapse architecture the impact of drift can be fully mitigated if the reset state of the PCM devices is tuned carefully to a partial RESET states, where the drift problem is negligible. In order to evaluate this strategy, we performed XNET simulations for the Binary-PCM synapse architecture using 3 different PCM RESET states, keeping the SET state constant with $R_{ON} \approx 600\,\Omega$:

1. negligible drift region, mean $R_{OFF} = 20\,k\Omega$;
2. low drift region, mean $R_{OFF} = 30\,k\Omega$;
3. high drift region, mean $R_{OFF} = 123\,k\Omega$.

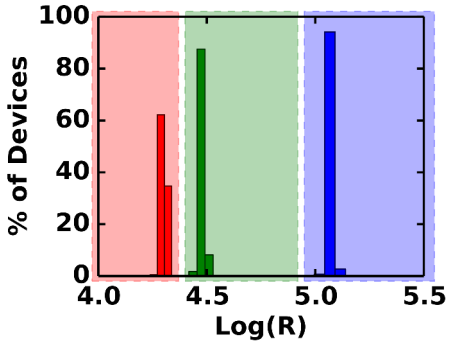
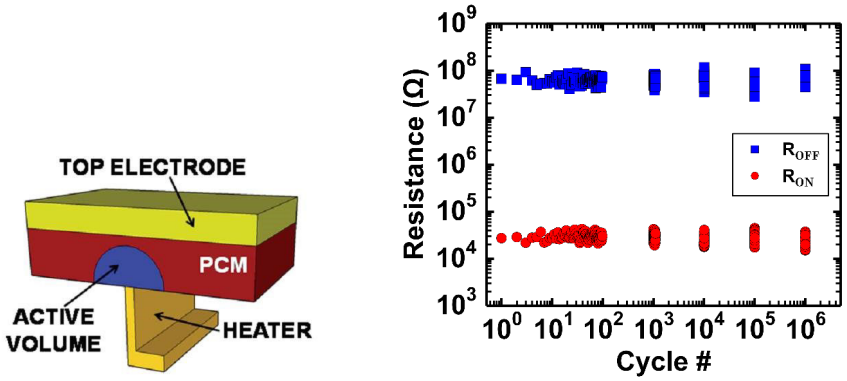


Figure 2.16. Distribution of synapses in RESET state for the binary PCM synapse approach with mean R_{OFF} values of 20 kΩ, 30 kΩ and 123 kΩ.

Quantity	*'2-PCM* Synapse'	*Binary-PCM* Synapse	
	Roff = 128 kΩ	Roff = 20 KΩ	Roff = 30 KΩ
Total Read	4.97×10^9	2.48×10^9	2.48×10^9
Total Set	3.86×10^8	5.13×10^5	4.84×10^5
Total Reset	1.56×10^7	1.90×10^7	1.78×10^7
Frequencies of events (event / device / sec)			
F read/d/s	1.9	1.9	1.9
F set/d/s	0.14	0.00038	0.00036
F reset/d/s	0.0058	0.014	0.013
Energy/ Power Consumption			
Set Energy	33.6 mJ	0.1 mJ	1.6 mJ
Reset Energy	37 mJ	36.2 mJ	42.6 mJ
Total Energy	71.4 mJ	36.3 mJ	44.2 mJ
Total Power	105 µW	53.8 µW	65 µW
Read Energy*	402 µJ	51 µJ	48 µJ
Read Power*	600 nW	75 nW	70 nW

Table 2.2. Comparison of learning statistics for the different simulated architectures [170].

The final synaptic resistance distributions for the PCM synapses in the 3 cases are shown in Fig. 2.16. Table 2.2 compares the learning statistics for the multilevel 2-PCM and the binary PCM approach with partial RESET states immune to drift. The energy consumption decreases in the case of binary-PCM synapse as the current required to program partial-reset states (20 kΩ and 30 kΩ) is much smaller compared to the current required to program the RESET strong reset state (128 kΩ). During read mode, the power consumption of the binary PCM approach is much smaller because, at the end of learning, the majority of the devices is in the high resistance state, contrary to the 2-PCM approach, as shown in Figs. 2.15a and 2.15b.



(a) (b)

Figure 2.17. (a) Schematic representation of the scaled wall storage structure PCM device [25]. (b) SET and RESET experiment for 10^6 cycles on scaled wall storage PCM device.

2.6 Simulations using scaled devices

In the previous sections, we carried out our analysis based on electrical results obtained from test devices featuring a large heater plug. The corresponding programming current is thus very large, in the order of tens of mA. In this section, we will show how, using state-of-the-art PCM devices, it is possible to reduce the programming current and, as a consequence, the synaptic programming energy. For this reason, electrical characterization has been performed on "wall storage" PCM memory cells, fabricated in 90 nm technology node in the framework of the R&D project between STMicroelectronics and CEA-LETI [25]. The device structure is schematically represented in Fig. 2.17a. The smallest feature size of the heater is not limited by the photolithographic resolution. The active phase-change material is GST. Figure 2.17b shows the R_{ON} and R_{OFF} values for one device corresponding to 10^6 SET-RESET cycles. The experimental resistance values of the PCM devices have been used to model, in XNET simulation, the resistance distributions of the PCM synapses during the learning of the car video. The characteristics of the STDP learning rule have also been optimized by genetic evolution algorithm [163] in order to adapt the learning to the new synaptic distributions. Specifically, the LTP time window and LTD probabilities have been increased from $t_{LTP} = 7.6\,\text{ms}$, $p_{LTD} = 0.11$ to $t_{LTP} = 13.4\,\text{ms}$, $p_{LTD} = 0.21$. Therefore, the network activity is larger in terms of SET and RESET events in the case of the wall storage device. Simulation results are presented in Table 2.3, where the learning statistics of the two studied neuromorphic systems, with large heater synapses and scaled wall storage synapses, are compared. Results show that, even if the programming activity is increased in the case of the scaled synapses, the total programming energy associated to SET and RESET events is

Quantity	Large heater synapses	Wall storage synapses
p_{LTP}	0.14	0.14
p_{LTD}	0.11	0.21
t_{LTP}	7.6 ms	13.4 ms
Nb. SET pulses	$4.5 \cdot 10^5$	$8.9 \cdot 10^5$
Nb. RESET pulses	$1.6 \cdot 10^7$	$4.7 \cdot 10^7$
Nb. Read pulses	$2.48 \cdot 10^9$	$2.48 \cdot 10^9$
Energy associated to SET events	0.4 mJ	0.2 mJ
Energy associated to RESET events	47.3 mJ	4.3 mJ
Total energy (SET + RESET)	47.7 mJ	4.5 mJ
Total power (SET + RESET)	70.1 µW	6.6 µW
Read energy	43 µJ	0.3 µJ
Read power	64 nW	0.5 nW

Table 2.3. Comparison of PCM learning statistics obtained for large heater devices and scaled wall storage devices.

reduced by one order of magnitude, from 70 µW to 7 µW. Furthermore, the power consumption associated to the read mode is reduced by three orders of magnitude, from 64 nW to 0.5 nW. This is due to the smaller current associated to the read operations with the scaled devices. The most important contribution to the to the total read current is in fact given by the read operations on devices in the SET state, because their resistance is smaller compared to the RESET state. Since for the scaled devices the mean R_{ON} is much larger compared to the heater plug devices (57 kΩ vs. 110 Ω), the resulting read current is smaller.

2.7 Conclusion

In this chapter, we investigated the use of PCM devices as synapses in a fully connected artificial neural network. We presented the limitations associated to the use of the multilevel synapse approach. Therefore, driven by the motivation to overcome the limitations associated to the multilevel approach, we explored the use of PCM synapses in binary mode. Based on the results obtained from electrical

characterization, we performed simulations of a large scale artificial neural network for complex visual application, tuning the resistance levels of the SET and RESET states according to the selected programming conditions. Programming schemes for architectures with- or without-selector devices are provided. The proposed programming schemes avoid the use of complex refresh schemes and unnecessary programming pulses required by multilevel PCM synapses. Simulation results show that the learning mode power consumption associated to the studied neuromoprphic can be dramatically reduced if the RESET state of the PCM devices is tuned to a relatively low resistance. Read-mode power consumption, on the other hand, can be minimized by increasing the resistance values for both SET and RESET states of the PCM devices. We also investigated the issue of PCM resistance drift and we proposed a strategy to mitigate this problem. We also observed that, using scaled devices, it is possible to dramatically reduce the power consumption thanks to the smaller programming current. In conclusion, we successfully demonstrated the interest of using PCM devices in binary mode in a neuromorphic system for visual applications.

Chapter 3

OxRAM technology: failure mechanisms and variability

In this chapter, the main features of the HfO_2-based OxRAM technology, one of the most promising emerging non-volatile memory (NVM) technologies, are presented. Experimental results on endurance failure mechanisms are discussed and a programming methodology to improve endurance at low operating current is proposed. A physical model able to explain OxRAM variability in both Low Resistance State (LRS) and High Resistance State (HRS) is presented. This study is carried out with a dual aim. From the point of view of conventional memory applications, variability is indeed one of the limiting factors for the adoption of OxRAM technology in commercial products. The understanding of the source of resistance variations of OxRAM can thus give guidelines to reduce this issue. From the point of view of neuromorphic computing, OxRAM devices are an ideal candidate for the implementation of artificial synapses. The development of a model able to reproduce device variability for a wide range of programming conditions can be used to study the impact of synaptic variability at the system level.

3.1 Device structure

The devices studied in this chapter feature a metal-insulator-metal (MIM) structure composed of an HfO_2 layer between a TiN/Ti top electrode and a TiN bottom electrode (Fig. 3.1) [171]. A bitcell is composed of 1 Transistor − 1 Resistor (1T1R) structure, where the access transistor is used to select the cell when integrated into an array and limit the current flowing through the device during programming. The electrical characterization was performed on both 1T1R bitcells and on a 28 nm CMOS digital testchip that contains 16 Circuits Under Test (CUTs) of 1 kb each, plus a digital controller (Fig. 3.2a,b), fabricated in the framework of an R&D project between STMicroelectronics and CEA-LETI [172]–[174].

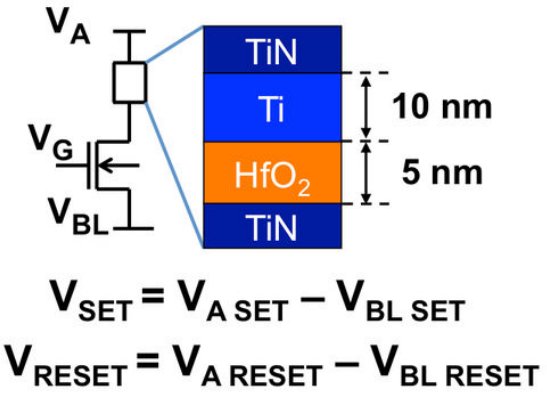


Figure 3.1. 1T-1R device schematic.

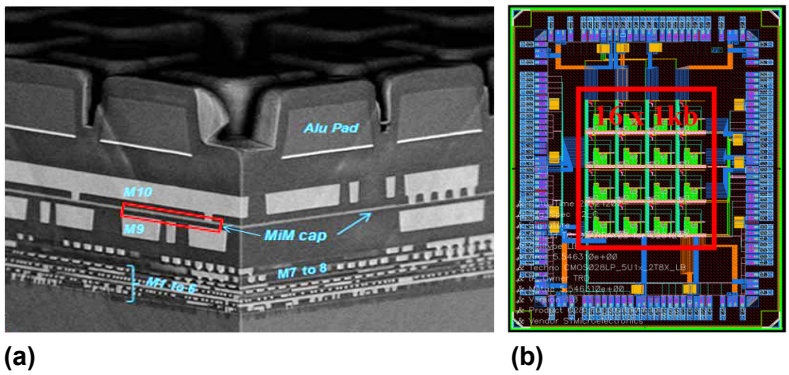


Figure 3.2. (a) SEM cross section of CMOS 28 nm stack including MIM device, (b) 16 kb circuit demonstrator layout [172].

3.2 Device operation

Typical I−V characteristics and switching behavior of a 1T1R bitcell are shown in Fig. 3.3. The operation of the device is bipolar, i.e. opposite voltage is necessary for the switching of the device from low to high resistance state and vice versa. Devices initially in Pristine Resistance State (PRS), featuring very high resistances, typically larger than 1 GΩ, are subjected to an electroforming operation (forming) by applying a positive voltage (≈ 2 V) across the device (red curve). This operation induces a soft breakdown of the oxide layer thus creating a conductive filament (CF) rich in oxygen vacancies (V_O) [48]. After the forming operation, the device is established in the Low Resistance State (LRS). The RESET operation, consisting in the application of an opposite voltage V_{RESET} across the device, partially disrupts the CF by recombination of oxygen vacancies with oxygen ions, hence leading to the High Resistance State (HRS) (green curve). In the SET operation, a positive voltage is applied on the top electrode (blue curve). This reforms the CF, bringing the device back in LRS. After

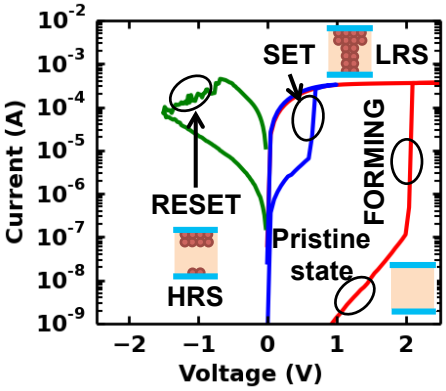


Figure 3.3. Typical Current-Voltage OxRAM characteristics. FORMING, SET and RESET operations are highlighted.

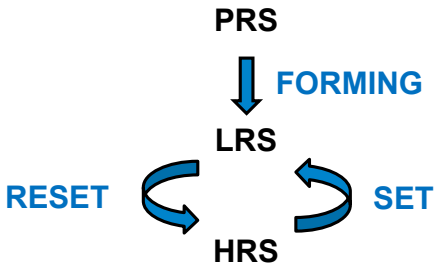


Figure 3.4. Flow chart of device operation.

the initial forming step, the device can be switched multiple times between LRS and HRS with RESET and SET operations as schematically depicted in Fig. 3.4. Table 3.1 reports the three main programming parameters that are associated to the three programming operations. After each SET and RESET operation, a low-field measurement of the device resistance is performed (READ operation): a voltage $V_{READ} = 0.1\,V$ is applied on the top electrode and the corresponding read current flowing through the device is measured.

3.3 Endurance: failure mechanisms

Endurance is defined as the number of SET/RESET cycles that a device can sustain while remaining functional, i.e. maintaining a significant resistance contrast between the ON/OFF states. When SET or RESET operations are not effective to switch the resistance state of the devices, the device is not functional and a *failure* occurs.

In the context of the use of OxRAM devices as artificial synapses, a good endurance is of great importance, because it allows a longer learning phase in a neuromorphic system, if the application requires a long learning phase. At the same time, it is

Operation	**Programming parameter**	**Symbol**
FORMING	Compliance current imposed during FORMING operation	$I_{C\ FORMING}$
RESET	Voltage applied to the device during RESET operation	V_{RESET}
SET	Compliance current imposed during SET operation	$I_{C\ SET}$

Table 3.1. Parameters associated to the three programming operations of OxRAM devices

also important to reduce the energy required for SET and RESET operations, in order to reduce power consumption during device operation. The impact of SET and RESET operations on endurance has been thoroughly investigated in literature. Chen *et al.* [175] provided experimental evidence of two opposite endurance failure mechanisms, observed as a function of the SET programming current $I_{C\ SET}$ for a given reset voltage V_{RESET}. The experiments were carried out on a memory stack similar to the one studied in this thesis, composed of TiN/Hf/HfO_2/TiN [175].

1. Low programming current $I_{C\ SET}$: an endurance failure with resistance stuck at HRS, with resistance values larger than 10 MΩ. This value is closer to pre-forming resistance values than to typical HRS resistance obtained after RESET operation. SET operations are not useful anymore to bring the resistance of the device back to LRS. An example is reported in Fig. 3.5a.

2. High programming current $I_{C\ SET}$: in this case, the failure occurs with the resistance of the device stuck at LRS. The RESET operation is not functional to bring the device resistance back to HRS. An example of this type of endurance is presented in Fig. 3.5b.

Chen *et al.* [175] also demonstrated that, by carefully balancing SET and RESET conditions, it is possible to improve the endurance as shown in Fig. 3.5c, achieving an endurance larger than 10^{10} SET-RESET cycles. This however implies that, for a given value of V_{RESET}, the value of $I_{C\ SET}$ must be increased to an intermediate value. This leads to an increased power consumption with respect to the conditions described in point 1.

In this work, we explore an alternative approach to improve endurance without increasing the power consumption during cycling. We start from the observation

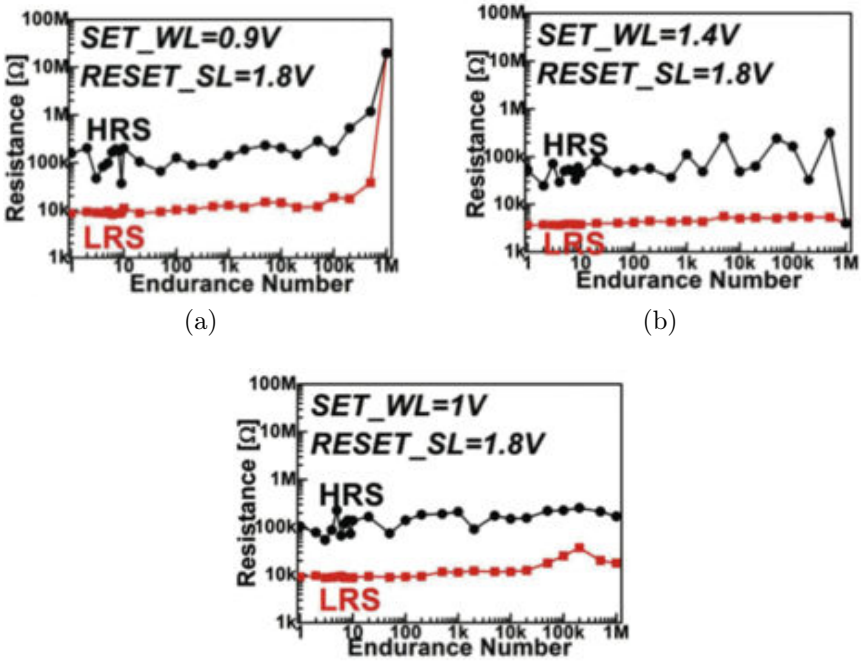


Figure 3.5. Example of: (a) endurance failure mechanism 1, where the resistance is stuck at pre-forming, high resistive state; (b) endurance failure mechanism 2, where the resistance is stuck at low resistive state. (c) Selecting balanced SET/RESET conditions allows improving the endurance but requires increasing $I_{C\ SET}$, thus increasing power consumption. Adapted from [175].

that, in order to reduce the average power consumption P, it is necessary to use an $I_{C\ SET}$ as small as possible for a given reset voltage V_{RESET} and programming pulse width t, as shown in Fig. 3.6. The approximated equations that have been used to estimate the average power consumption are the following:

$$E_{SET} \approx I_{C\ SET} \cdot V_{SET} \cdot t \tag{3.1}$$

$$E_{RESET} \approx I_{RESET} \cdot V_{RESET} \cdot t \tag{3.2}$$

$$E_{TOT} = E_{SET} + E_{RESET} \tag{3.3}$$

$$P = \frac{E_{TOT}}{T} \tag{3.4}$$

where E_{SET} and E_{RESET} are the energies required for SET and RESET operations, respectively. $I_{C\ SET}$ is the compliance current imposed by the select transistor during the SET operation. I_{RESET} is the current flowing through the device during RESET, and assumed equal to $I_{C\ SET}$, according to the universal RESET characteristics of OxRAM devices [176]. $V_{SET} = 1\,V$ and V_{RESET} are the voltage drops across the 1T1R

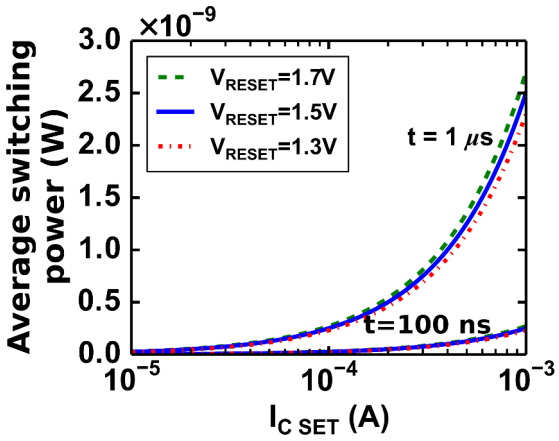


Figure 3.6. Estimated switching power as a function of $I_{C\,SET}$, obtained for 3 different values of V_{RESET}, for a pulse width of $t = 100\,ns$ and $t = 1\,\mu s$.

structure during SET and RESET operations, respectively. T is the time frame over which the average is computed, assumed equal to 1 s. As a consequence, the endurance failure mechanism that will be encountered is mechanism 1, i.e. resistance stuck at HRS.

Fig. 3.7 shows the typical bitcell behavior during an endurance experiment. The programming current is $\approx 230\,\mu A$. If long pulse widths equal to $1\,\mu s$ for both RESET and SET are used, no failure occurs and more than 10^8 cycles endurance can be achieved (Fig. 3.7 (a)). If shorter pulses with height 1 V and width 100 ns are used, a failure in endurance is observed after around 10^5 cycles, with the resistance of the device stuck at high resistance state Fig. 3.7 (b). The SET failure is an evidence of the unbalance between RESET and SET conditions, which is accentuated at short programming time. Specifically, the value of $I_{C\,SET} \approx 230\,\mu A$ is too low for the adopted $V_{RESET} = 1.3\,V$ at $t = 100\,ns$.

While the role of SET and RESET programming conditions has been discussed in literature in depth [175], little importance has been given to the role of the FORMING operation on the endurance performance. We can schematically represent our hypothesis of what happens in the OxRAM device during FORMING and first RESET operation with the simple schematic pictures reported in Fig. 3.8. During FORMING operation, the oxygen vacancy-rich Conductive Filament (CF) is formed in the insulator layer. The size cross-section Φ of the CF is believed to be proportional to the current flowing through the cell $I_{C\,FORMING}$ [177]. During the 1st RESET operation, the CF is partially dissolved in the region with length L_{GAP}. Higher the reset voltage V_{RESET}, higher L_{GAP} [171], [178]. This hypothesis is supported by the fact that higher V_{RESET} values correspond to higher HRS resistance values. The portion of the CF that remains after the first RESET operation would work like a reservoir (RES) for the formation/dissolution of the CF in the following SET/RESET

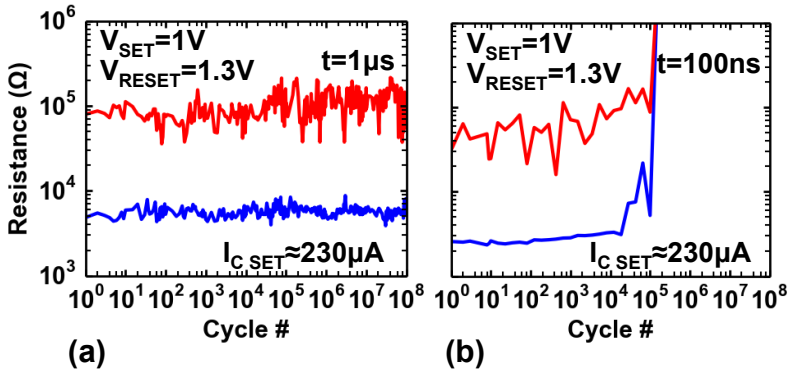


Figure 3.7. (a) Endurance test for programming pulse width $t = 1\,\mu s$. The device can be successfully switched for more than 10^8 cycles. (b) Using shorter programming pulse width ($t = 100\,ns$), an early SET failure with device stuck at HRS is observed.

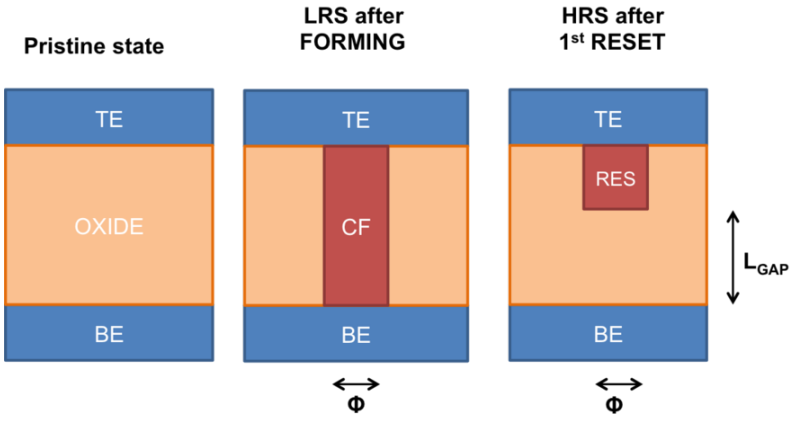


Figure 3.8. Schematic view of pristine device, conductive filament after forming and V_O reservoir in HRS after first RESET operation.

cycles. The presence of this reservoir is justified by the fact that the SET voltage is much lower with respect to the forming one. Moreover this reservoir can justify the lower RESET resistance value with respect to the initial pristine PRS state.

Having this simple scenario in mind, failure in endurance presented in Fig. 3.7b can be explained as follows. The gradual increase of the high resistance level during cycling suggests a gradual decrease of this reservoir. At every SET/RESET cycle a small fraction of RES is lost, until it becomes too small to create a continuous conductive filament during SET operation and the cell remains stuck in pre-forming state. In partial support of this hypothesis, experimental results also provided in reference [175] show that it is possible to recover the failed device by applying a new FORMING step, which can be interpreted as re-forming the reservoir.

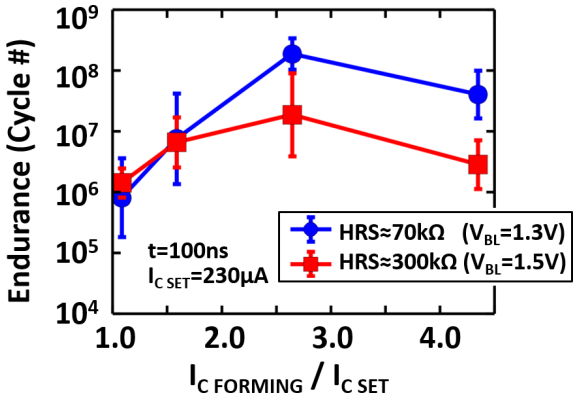


Figure 3.9. Endurance as a function of the ratio $I_{C\ FORMING}/I_{C\ SET}$, obtained for different values of V_{RESET} for $t = 100$ ns. Each point corresponds to the mean value on about 4 cells. An experimental optimal value of the ratio is observed at $I_{C\ FORMING}/I_{C\ SET} \approx 2.7$.

3.3.1 Endurance improvement for low programming current

Based on the previous considerations, to improve endurance for a given SET/RESET condition, we propose to increase the ratio between the reservoir and the conductive filament size. Higher is the ratio between the reservoir and the conductive filament size, better will be the endurance. In order to achieve this, a tailored forming operation is here suggested. It features a compliance current during forming ($I_{C\ FORMING}$) higher than the compliance current used during subsequent SET operations ($I_{C\ SET}$), in order to increase the size of the vacancies reservoir generated during the forming operation. Endurance performances (number of cycles obtained with LRS $<10\,k\Omega$) as a function of the ratio $I_{C\ FORMING}/I_{C\ SET}$ obtained with a strong (red curve, square symbols) and weak (blue curve, circle symbols) RESET conditions, are shown in Fig. 3.9. Each point corresponds to the mean value computed on about 4 cells. Worse endurance performance for strong RESET conditions can be explained by an early depletion of the vacancies reservoir generated during the forming operation. Higher $I_{C\ FORMING}/I_{C\ SET}$ values allow to increase the oxygen reservoir and consequently to improve the endurance performances. However $I_{C\ FORMING}$ values higher than 900 µA degrade the oxide properties and consequently the cell reliability. Thus an experimental optimal value of $I_{C\ FORMING}/I_{C\ SET}$ of about 2.7 is observed. In Fig. 3.10, more than 10^8 cycles without failure have been demonstrated for weak RESET conditions with an optimal value of the ratio $I_{C\ FORMING}/I_{C\ SET} \approx 2.7$ and a SET pulse height 1 V and width 100 ns. Since a higher current is used only at the forming step and not during the subsequent SET/RESET operations, the impact of the higher $I_{C\ FORMING}$ on the total power consumption is negligible.

Moreover, we demonstrated that the devices can withstand $>10^8$ read cycles without drift for both LRS and HRS. Figure 3.11 shows the read current response

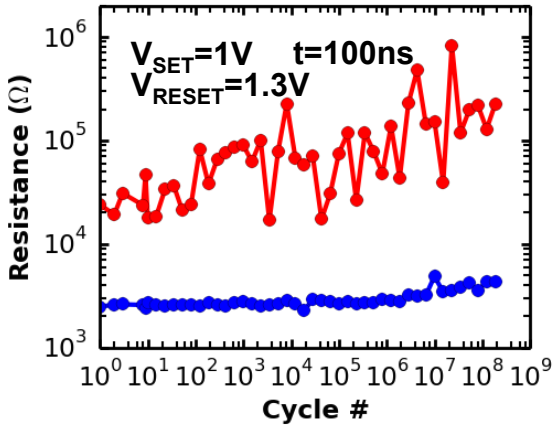


Figure 3.10. Endurance test for short programming pulse width $t = 100\,\text{ns}$ after ad hoc forming operation with $I_{C\ FORMING}/I_{C\ SET} \approx 2.7$. More than 10^8 SET-RESET cycles are achieved.

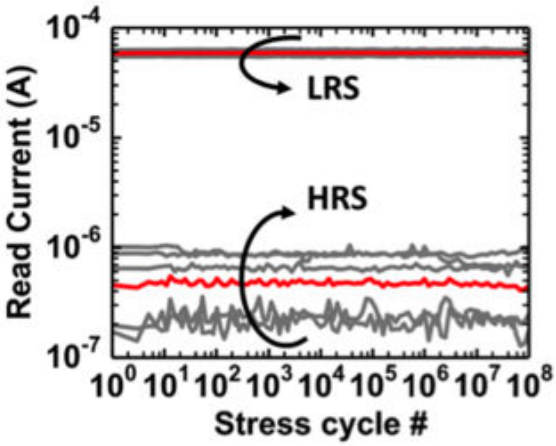


Figure 3.11. Read current response to pulse voltage stress corresponding to 10^8 read cycles. Pulse width is equal to 1 µs, amplitude 0.1 V.

to pulse voltage stress corresponding to 10^8 read cycles. The experiment was interrupted for time limitations, not because of device failure. Pulse width is equal to 1 µs, amplitude 0.1 V. Red curve corresponds to the average over multiple devices (grey curves).

3.4 Variability

As introduced in Section 1.2.4, variability is one of the main issues that limits OxRAM technology to be adopted in large memory arrays. As an example of the order of magnitude of variability, Fig. 3.12 shows the values of LRS and HRS resistances of a

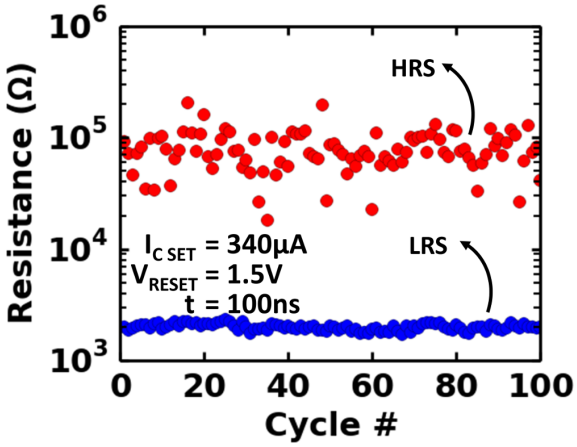


Figure 3.12. Resistance values of LRS and HRS during 100 corresponding SET and RESET operations.

single device measured with a READ operation, corresponding to 100 consecutive SET-RESET cycles, performed with 100 ns pulse-width and 1 V-1.5 V pulse-height, respectively. The HRS exhibits significant variability with resistance values ranging over one decade between 20 kΩ and 200 kΩ). The LRS variability appears to be smaller, with points tightly distributed around the 2 kΩ value. In order to extensively characterize the properties of variability, we performed a thorough work of electrical characterization. This has been done with the goal of collecting a large statistics of resistance distributions for both LRS and HRS on multiple devices. To summarize the results on variability obtained for different programming conditions, Fig. 3.13 reports the typical Cumulative Distributions Function (CDF) of the resistance in LRS and HRS. A single curve is obtained for 100 SET-RESET cycles (as presented in Fig. 3.12), while the different curves are obtained by respectively tuning the values of the programming current during SET ($I_{C\,SET}$ fixed by the gate word-line voltage of the transistor, ranging from 10 µA to 340 µA) and the reset voltage (pulse amplitude of the bit-line voltage V_{RESET} ranging from 1.3 V to 1.7 V). By tuning the SET and RESET conditions it is possible to tune the resistance values of LRS and HRS, respectively. The distributions are well approximated as lognormal. In order to prove this, we performed statistic Pearson's χ^2 tests, which confirmed the goodness of fit of the resistance distributions as lognormal with 95% confidence. The mean HRS value hence is controlled by the reset voltage, while the standard deviation of the distribution remains nearly constant, independently of the values of V_{RESET}. In LRS, on the contrary, it appears that both the mean value and the standard deviation of the distributions change with $I_{C\,SET}$. As shown in Fig. 3.13, the statistic estimators μ_R and σ_R are extracted for each curve, i.e. for each SET/RESET condition. For a given resistance distribution, the value of μ_R is extracted as the 50% of the distribution, while σ_R is determined as the difference between 70% and

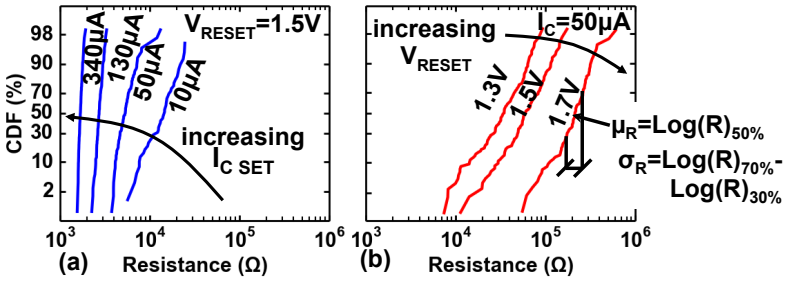


Figure 3.13. Experimental cumulative distributions of (a) LRS changing $I_{C\ SET}$ while keeping V_{RESET} constant and (b) HRS, changing V_{RESET} while keeping $I_{C\ SET}$ constant. The extraction of mean value (μ_R) and standard deviation (σ_R) are indicated.

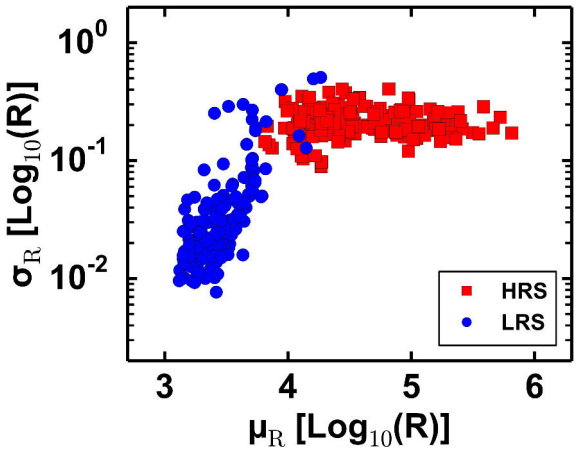


Figure 3.14. Experimental σ_R vs. μ_R, showing the variability evolution from LRS to HRS.

30% of the distribution. Plotting the value of σ_R as a function of μ_R leads to the graph $\sigma_R(\mu_R)$ in Fig. 3.14. This plot, which is representative of more experiments than those presented in Fig. 3.13, demonstrates a continuous evolution of resistance variability from a constant HRS value to a decreasing LRS value. Similar results were obtained in the literature by Fantini *et al.* [57] with a similar memory stack TiN/Hf/HfO_2/TiN. The difference is that a plateau in variability was not observed: σ_R increased with μ_R also in HRS. More importantly, the continuity in the evolution of variability from LRS to HRS was not put in evidence. Similar results have also been obtained by Ambrogio *et al.* in [179], with OxRAM devices also featuring a TiN/Ti/HfO_2/TiN memory stack. A comparison between the results obtained by the Politecnico di Milano group and our results are reported in Fig. 3.15. A similar trend is recognizable and a continuity in the evolution of the variability is observed in both cases, whatever the process technology.

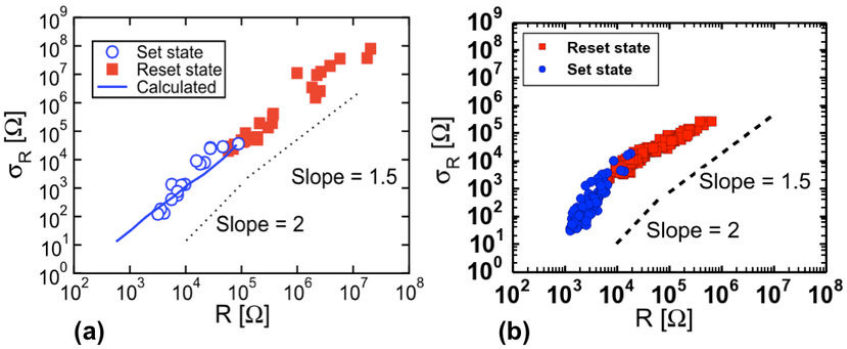


Figure 3.15. Comparison between experimental results of σ_R vs. μ_R obtained independently by (a) Ambrogio *et al.* and published in ref. [179] and (b) the results presented in this thesis in Fig. 3.14 (computed in linear scale) [59].

3.5 Variability Modelling: 3D resistor network approach

In this section, we address the modeling of OxRAM variability based on the experimental results provided Section 3.4.

In the literature, two different approaches have been used to model the variability of LRS and HRS.

- Resistance variability in LRS has been interpreted as the results of variations of the filament size and geometry in the framework of the quantum point contact model [57], [58], [180].

- In HRS, resistance variability has been interpreted as the results of variation in the length of the tunneling barrier in the framework of trap-assisted tunneling model [55], [56], [181] and of the Poisson fluctuation of the number of defects in the gap region [179].

The experimental observation of the continuity in the evolution of variability from LRS to HRS, discussed in Section 3.4, provided us the motivation to explore a modeling approach valid for both LRS and HRS.

The scenario described in Fig. 3.16 has been employed to define the CF shape: after forming operation, the V_O concentration ($<V_O>$) is constant along the CF length (Fig. 3.16a). After RESET, a portion of the filament is disrupted in the L_{GAP} region and the $<V_O>$ profile gradually changes with a truncated normal distribution profile on each side (Fig. 3.16b). The trap position features a constant standard deviation s_{Vo} on the edges. Stronger RESET voltages correspond to longer L_{GAP} regions, hence leading larger resistance. After the SET operation, the length L_{GAP} is reduced. Below a critical length, the two trap distributions overlap, inducing

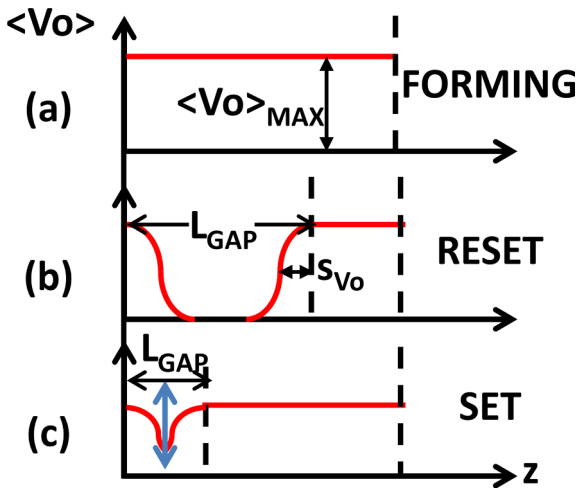


Figure 3.16. Schematic view of the theoretical oxygen vacancy profile along the filament (z direction) after (a) Forming, (b) SET and (c) RESET operation.

a progressive increase of $<V_O>$ in the L_{GAP} region (Fig. 3.16c). The 3D resistor-network approach allows defining the random position of an arbitrary number of V_O's describing the CF in the HfO_2 layer. Table 3.2 provides a summary of the parameters used to model the filament in terms of $<V_O>$ and geometry, as indicated in Fig. 3.18. s_{Vo} can be interpreted as the abruptness of the $<V_O>$ profile along the L_{GAP} region. The maximum value of the oxygen vacancy concentration is a fitting parameter, the only constraint is that the minimum distance between two oxygen vacancies should not be smaller than the interatomic distance (≈ 3 Å [56]). Hence, L_{GAP} is the only parameter used to modulate the resistance value. Lower L_{GAP} values correspond to a higher trap concentration in the CF. Once the positions of the V_O's have been randomly determined for a fixed value of L_{GAP}, according to the probability profile explained in Fig. 3.16, the computation is carried out by solving a 3D resistor network where the nodes of the network are composed by the V_O's and the top and bottom residual portions of the filament. An example of two V_O system is shown in Fig. 3.17. Using Kirchhoff's first law applied on this simplified system, the set of equations 3.5 can be obtained:

$$\begin{cases} (V_1 - V_2)\,G_{12} + (V_1 - V_3)\,G_{13} + (V_1 - V_4)\,G_{14} = 0 \\ (V_2 - V_1)\,G_{12} + (V_2 - V_3)\,G_{23} + (V_2 - V_4)\,G_{24} = 0 \\ (V_3 - V_1)\,G_{13} + (V_3 - V_2)\,G_{23} + (V_3 - V_4)\,G_{34} = 0 \\ (V_4 - V_1)\,G_{14} + (V_4 - V_2)\,G_{24} + (V_4 - V_3)\,G_{34} = 0 \end{cases} \tag{3.5}$$

where V_i is the voltage at the i^{th} node of the resistor network and G_{ij} is the

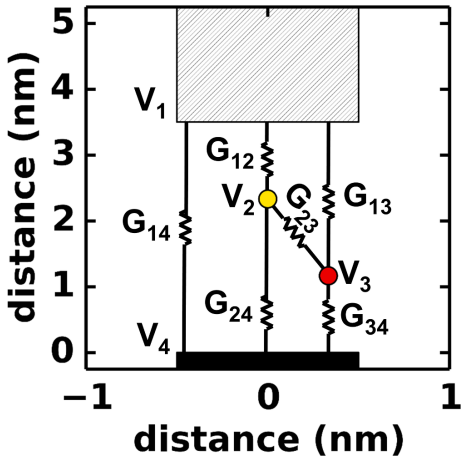


Figure 3.17. Schematic of the resistor network implemented in model, for a simplified configuration with only two traps.

conductance between i and j nodes. Eq. 3.5 can be rewritten as in Eq. 3.6.

$$\begin{cases} V_1 \left(G_{12} + G_{13} + G_{14}\right) - G_{12}V_2 - G_{13}V_3 - G_{14}V_4 = 0 \\ \qquad \vdots \\ V_4 \left(G_{14} + G_{24} + G_{34}\right) - G_{14}V_1 - G_{24}V_2 - G_{34}V_3 = 0 \end{cases} \tag{3.6}$$

Eq. 3.6 can be generalized to the case of a resistor network with n nodes using Eq. 3.7.

$$V_i - \frac{\sum\limits_{i \neq j} \left(G_{ij} V_j\right)}{\sum\limits_{i \neq j} G_{ij}} = 0 \qquad i, j = 1, \ldots, n \tag{3.7}$$

In order to determine the conductance value G_{ij} between two nodes i, j, the following equation is used:

$$G_{ij} = G_0 \, \exp\left[-2 \frac{\sqrt{2 m_{\text{eff}} E_{\text{B}}}}{\hbar} \left(x_{ij} - a\right)\right] \tag{3.8}$$

where $m_{\text{eff}} = 0.1 m_0$ is the electron effective mass in hafnia, $E_B \approx 2\,\text{eV}$ is the energy barrier height, x_{ij} is the distance between traps i and j, $a \approx 3\,\text{nm}$ is the inter-atomic distance, i.e. the minimum distance between traps [56], and G_0 is defined as the quantum of conductance equal to $2q^2/h$. Equation 3.8 hence defines the conductance between two traps as equal to the quantum of conductance if their distance is equal to the interatomic distance a (i.e. two traps next to each other). Else, the conductance is reduced by a factor equal to the tunneling probability between traps, which decreases exponentially with their distance. Using the boundary conditions $V_1 = V_{\text{READ}} = 0.1\,\text{V}$ and $V_n = 0\,\text{V}$ it is possible to solve Eq. 3.7, obtaining voltage and

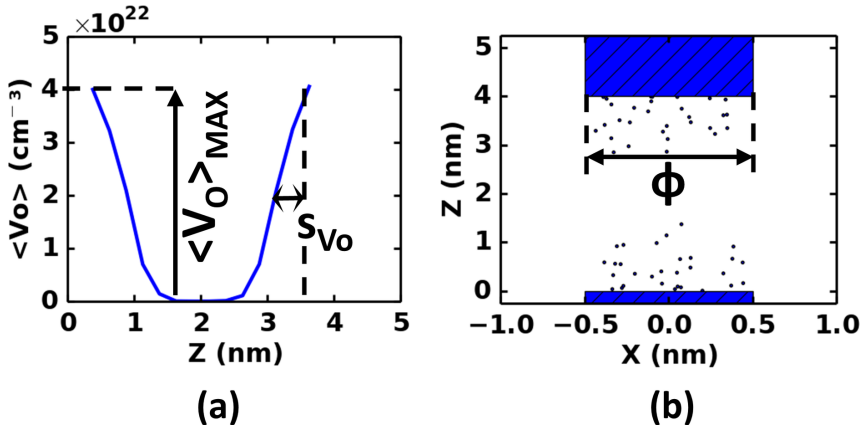


Figure 3.18. Averaged V_O concentration as a function of vertical position along the filament, and position of traps in the gap for one random drawing, illustrating the main parameters used in the model (and listed in Table 3.2).

current values at every node. It is thus possible to extract the equivalent resistance of the filament.

It is worth noticing that our modeling approach is a simplification of more complete models provided in [56], [182]. In our approach, the position of oxygen vacancies is not the result of a self consistent SET/RESET simulation. In other words, our model does not allow to reproduce the full IV curve of OxRAM deivces. The validity of the model is limited to the computation of the trap-assisted tunneling current corresponding to small values of voltage typical of READ operation ($V_{\mathrm{READ}} = 0.1\,\mathrm{V}$). In this range, the shape of the energy barrier between traps can be considered rectangular and the Eq. 3.8 is valid. For larger values of voltage, a more computationally expensive model should be adopted, taking into account the electric field, electron-phonon coupling and lattice relaxation.

Figs. 3.19 (a) and (b) illustrate two different filament configurations for two different L_{GAP} values, corresponding to LRS and HRS respectively. In Figs. 3.19 (c) and (d) the simulated trap concentration profiles corresponding to LRS and HRS are shown. The light grey curves represent single random configurations, while the dark blue lines are the average profiles of traps over multiple cycles, reproducing the scenario represented schematically in Fig. 3.16. Figs. 3.19 (e) and (f) show the lateral cross sections of the filament in LRS and HRS, respectively. The color code indicates the voltage values along the filament obtained solving Eq. 3.7 and Eq. 3.8 for a read voltage equal to 0.1 V. In the case of LRS the voltage at the traps changes continuously along the simulated filament constriction. In HRS, on the other hand, an abrupt voltage drop occurs between top and bottom residual filament portions. Fig. 3.20 reports the evolution of the voltage profile of the traps along the length of the simulated CF constriction, from LRS to HRS, corresponding to four different values of L_{GAP}. For a resistance value of $\approx 8\,\mathrm{k}\Omega$, corresponding to $L_{GAP} = 2\,\mathrm{nm}$

Parameter	Description	Default value
$<V_O>_{MAX}$	Peak value of V_O concentration distribution	$4 \times 10^{22}\,cm^{-3}$
s_{Vo}	Standard deviation of V_O concentration distribution	4 Å
Φ	Filament diameter	1 nm
L_{GAP}	Distance between top and bottom CF residual portions	function of R

Table 3.2. Summary of the parameters used for the model.

(Fig. 3.20a), the OxRAM device is in LRS and a linear voltage profile is obtained. For two intermediate resistance states, $\approx 20\,k\Omega$ and $\approx 80\,k\Omega$, corresponding to $L_{GAP} = 2.5\,nm$ and $L_{GAP} = 3\,nm$ respectively (Fig. 3.20b-c), the resistivity along the CF is not constant and consequently the voltage profile is non-linear. In the case of a high resistive state of $\approx 200\,k\Omega$, ($L_{GAP} = 3.5\,nm$, Fig. 3.20d) a voltage drop occurs in the middle of the L_{GAP} region, where no traps are present and the resistivity of the oxide is the largest. In the latter case, the conduction mechanism is dominated by tunneling between the two residual filament portions.

3.6 Continuity of variability from LRS to HRS: model calibration

The previous equations is used in a Monte Carlo like code, which generates multiple sets of random trap configurations, and hence generates statistical sets of resistances for a given L_{GAP}. This statistic, for one L_{GAP}, is used to build CDF graph, like in the experimental case, as presented in Fig. 3.21, which shows the simulated CDFs of resistance obtained for values of L_{GAP} increasing from 1 nm to 4 nm. The extraction of statistic estimators μ_R and σ_R can be performed on simulated CDFs as in the case of the experimental results.

In order to calibrate the model, a set of simulations has been performed varying the parameters listed in Table 3.2. Fig. 3.22 reports a comparison between experimental results (symbols) and simulations (solid lines). The dependence of the variability on the parameters listed in Table 3.2 has been evaluated: increasing $<V_O>$ decreases σ_R in HRS due to the reduction of possible spatial combinations of traps (Fig. 3.22a); increasing s_{Vo} increases σ_R in HRS and displaces the transition between the two regimes (Fig. 3.22b). Finally increasing Φ lowers σ_R and displaces the transition

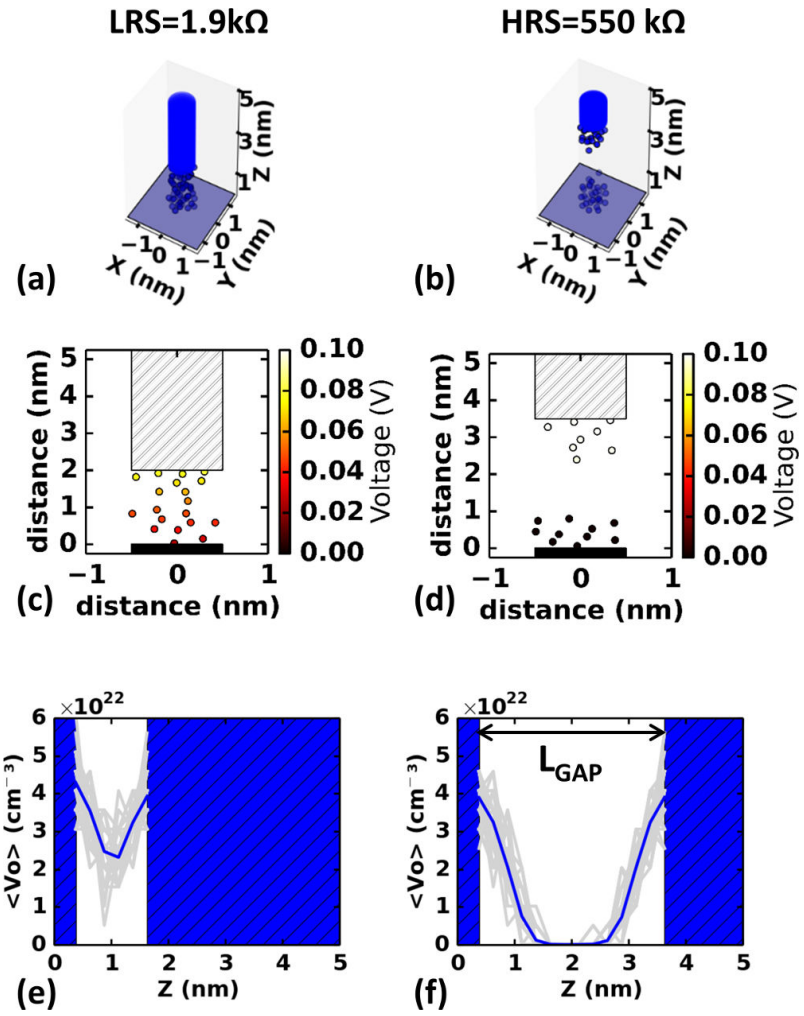


Figure 3.19. (a), (b) 3-dimensional view of simulated CF in LRS and HRS. (c), (d) Lateral cross-section of CF with voltage values calculated using Eq. 3.7 and Eq. 3.8. (e), (f) Oxygen vacancy concentration $<V_O>$ profiles for LRS and HRS. Grey curves correspond to individual random drawings; blue curves correspond to the average of these drawings over different cycles.

point (Fig. 3.22c). The model has thus been calibrated and the best fit values listed in Table 3.2 have been obtained. Fig. 3.23 reports a comparison between experimental and simulation results obtained with calibrated parameters.

The modeling approach presented in this work is valid in both LRS and HRS using the same set of equations. It reproduces very well the experimental results presented in Section 3.4, thus explaining the continuity in the evolution of variability from LRS to HRS.

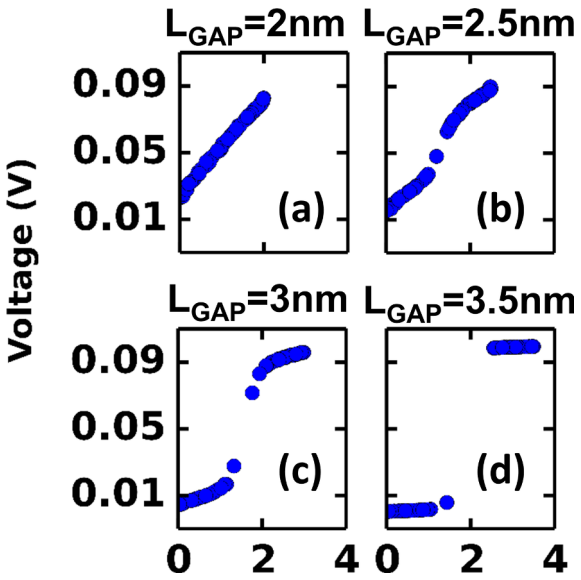


Figure 3.20. Simulated voltage profile of traps in the L_{GAP} region for L_{GAP}= 2 nm (a), 2.5 nm (b), 3 nm (c), 3.5 nm (d).

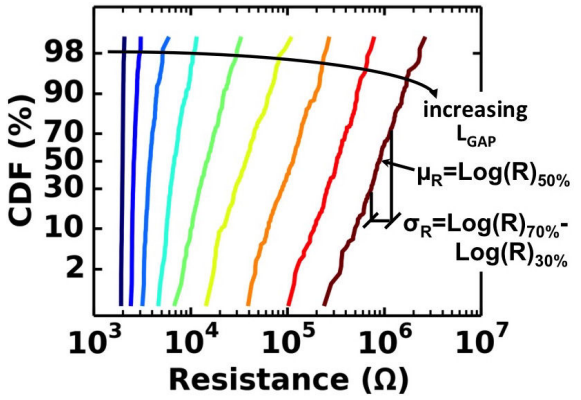


Figure 3.21. Simulated cumulative distributions of resistance obtained for values of L_{GAP} increasing from 1 nm to 4 nm. It is possible to observe continuity in the transition between LRS and HRS.

3.7 Variability from 28 nm memory array demonstrator

3.7.1 Cycle-to-cycle variability

In this section, we focus on the *cycle-to-cycle (temporal)* variability, i.e. on the resistance variations occurring at each SET/RESET cycles observed on the same device. In order to evaluate experimentally the cycle-to-cycle variability of OxRAM resistance over a large statistics, 10^4 consecutive SET-RESET have been performed

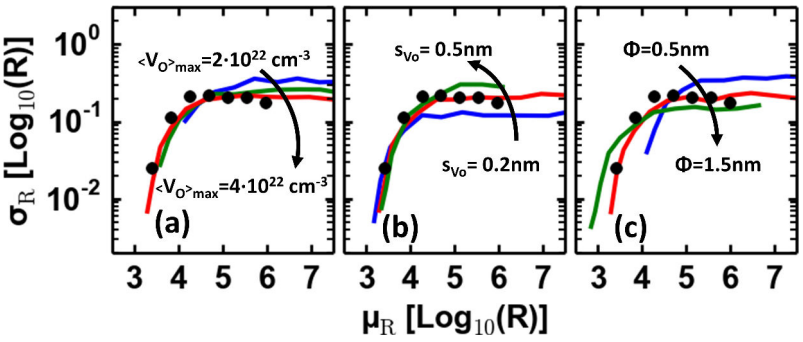


Figure 3.22. Variability dependence on (a) $<V_O>_{MAX}$, (b) width of the $<V_O>$ profile distribution s_{Vo} and (c) CF diameter. Black points are geometric mean of experimental data, simulations in solid lines.

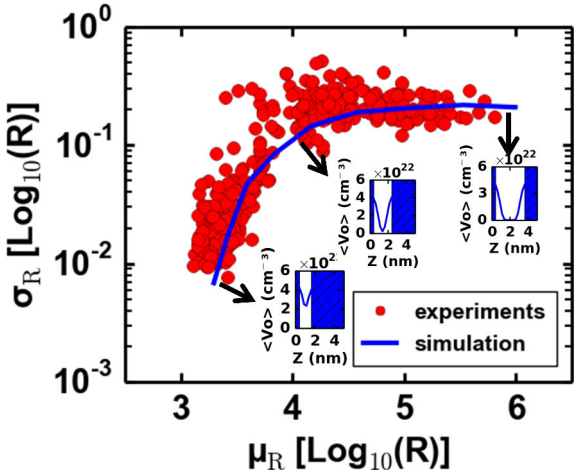


Figure 3.23. Experimental and simulated standard deviation of the resistance as a function of the mean resistance. The simulation curve has been obtained using parameters reported in Table 3.2.

on the same bitcell (Fig. 3.24). The corresponding CDFs are reported in Fig. 3.25a (symbols). The model presented in the previous section has been used to perform 10^4 simulations, keeping the value L_{GAP}= 1.5 nm and L_{GAP}= 3.5 nm constant for LRS (Fig. 3.25b) and HRS (Fig. 3.25c), respectively. The corresponding CDFs are reported in Fig. 3.25a (lines). Simulations well reproduce the experimental distributions. Cycle-to-cycle variability is thus interpreted as the result of the random placement of oxide defects V_O in the CF constriction region, which is partially re-formed and disrupted at each SET-RESET cycle, for a fixed value of L_{GAP}.

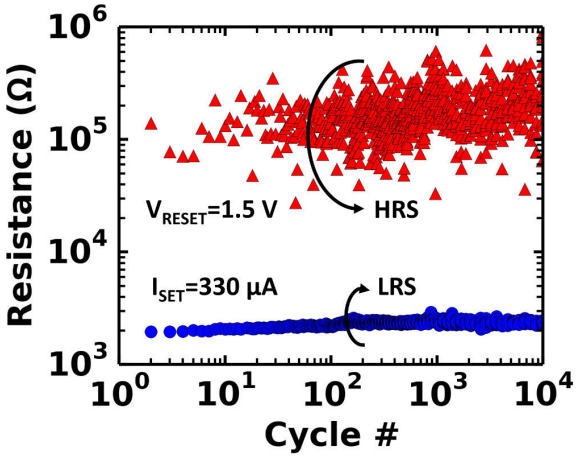


Figure 3.24. Resistance values of LRS and HRS during 10^4 SET and RESET operations on single bitcell.

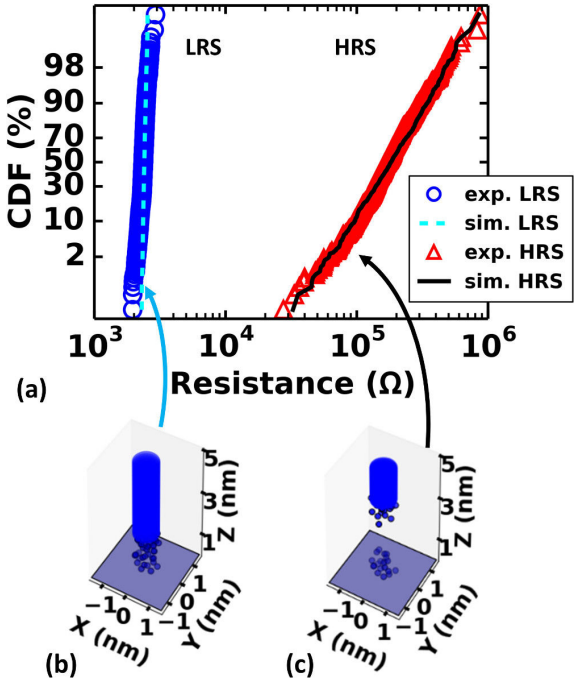


Figure 3.25. (a) Experimental and cumulative distribution of resistance values for LRS and HRS. (b), (c) 3D view of the CF in LRS and HRS, respectively.

3.7.2 Device-to-device variability

The *device-to-device (spatial)* variability has been also evaluated. Device-to-device variability is defined as the resistance variation observed over a large population of devices programmed with the same conditions. In Fig. 3.26(a) cumulative distributions of LRS and HRS extracted from 1 kb OxRAM array [172] statistics are

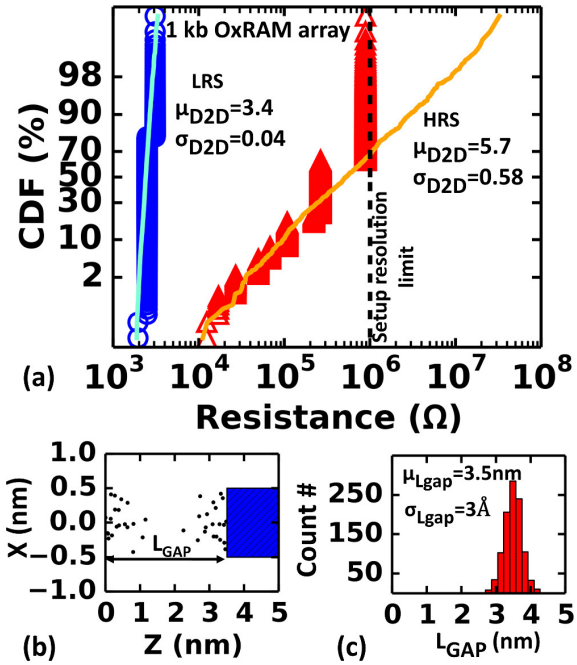


Figure 3.26. (a) Cumulative distributions of LRS and HRS for experimental (symbols) and simulated (solid lines) 1 kb OxRAM array. Discrete steps in the experimental distributions are due to discrete thresholds in read current sensing. The experimental distribution is cut at $\approx 1\,\mathrm{M\Omega}$ due to lower limit in current sensing. (b) Lateral view of CF with $\mathrm{L_{GAP}} = \mu_{\mathrm{Lgap}} = 3.5\,\mathrm{nm}$. (c) Histogram of $\mathrm{L_{GAP}}$distribution of simulated CFs.

reported. No correction code or smart programming algorithms [183] (allowing for higher programming windows), have been used. In the memory array, the READ operation is performed with a digital controller by sensing the read current of each device and comparing it to multiple thresholds. This is why discrete steps in the experimental distributions (symbols) are obtained. The experimental distribution is cut at $\approx 1\,\mathrm{M\Omega}$ because of the lower limit in current sensing of the digital controller.

The device-to-device variability is higher with respect to the cycle-to-cycle variability on single devices. Therefore, to model the device-to-device variability, a dispersion of the $\mathrm{L_{GAP}}$ value has been introduced (as shown in Fig 3.26c), following a normal distribution with $\mu_{\mathrm{Lgap}} = 3.5\,\mathrm{nm}$ and $\sigma_{\mathrm{Lgap}} = 3\,\mathrm{Å}$. This source of variability is added to the intrinsic variability due to random placement of traps at each programming cycle described in Section 3.7.1, and it can capture spatial variations such as process related variations in local thickness of the deposited HfO_2 layer. Fig 3.26b illustrates the lateral view of a CF with $\mathrm{L_{GAP}} = \mu_{\mathrm{Lgap}}$. As shown in Fig 3.26a, the physical model well reproduces the device-to-device experimental LRS and HRS distributions.

3.8 Conclusion

In this chapter, a 3D model able to reproduce the resistance variability behavior of OxRAM devices, for a wide range of resistance values, has been presented. The change of resistance state and the associated variability is interpreted as a consequence of the modulation of the oxygen vacancies V_O concentration profile along the conductive filament. The model well reproduces the variability trend from HRS to LRS and is able to explain the variability of 1 kb OxRAM array. According to the results obtained with this model, it could be possible to reduce the variability of OxRAM resistance by controlling the gradient profile of traps at the edge of the CF. This understanding could be used for the technological optimization of future OxRAM devices. Thanks to this model, variability of OxRAM devices can be quantitatively estimated and be taken into account in simulations of neuromorphic systems based on OxRAM-based synapses.

Chapter 4

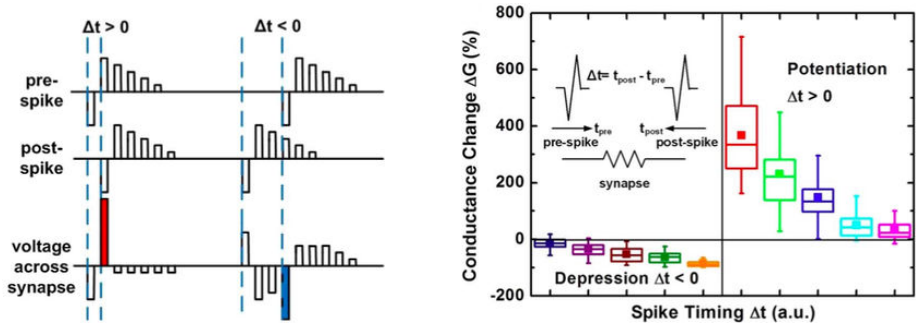
OxRAM devices as artificial synapses for convolutional neural networks

In this chapter, we study the use of OxRAM devices as artificial synapses in neuromorphic systems. We first review the results already reported in the literature on the use of OxRAM synapses in neuromorphic systems, examining advantages and disadvantages of both multilevel and binary approach. We then propose an OxRAM-based synapse design that combines together the advantages of multilevel and binary approaches. Based on the proposed synapses, we propose a hardware implementation of a convolutional neural network (CNN) for complex visual applications such as handwritten digits and traffic signs recognition. Based on OxRAM electrical characterization results and thanks to the understanding and modeling of device variability that we achieved in Chapter 3, we study in simulation the impact of OxRAM programming conditions on the network performance. We then explore the possibility of unsupervised learning. Finally, we investigate the tolerance of the proposed network to both temporal and spatial synaptic variability.

4.1 Introduction

Among the advantages of OxRAM technology, good scalability, write speed and low switching energy [46], [184]–[187] are attractive not only for conventional memory applications, but also for the implementation of artificial synapses in neuromorphic systems. Thanks to these appealing properties, OxRAM devices have been indicated in literature as good candidates to emulate biological synaptic plasticity in artificial neural networks. Two main approaches of OxRAM synapse implementation have been demonstrated by multiple research groups [92], [93], [96], [99], [125], [188], [189]: multilevel (or analog) and binary approach. In this section, we will review the main advantages and disadvantages of both implementations.

In the multilevel approach, a single device is used to implement an artificial



(a) (b)

Figure 4.1. (a) Multilevel 1R OxRAM programming scheme developed with pulse amplitude modulation. (b) STDP-like curve calculated on OxRAM devices employing the signal schemes in (a). Adapted from [96].

synapse. Multiple Low-Resistance State (LRS) resistance levels for emulating Long-Term Potentiation (LTP) and multiple High Resistance State (HRS) resistance levels for Long-Term Depression (LTD) are adopted. In unsupervised learning, a deterministic STDP rule can be used for the learning phase. Figure 4.1a illustrates schematically the programming scheme initially proposed by Yu *et al.* [96] for HfO_x-based synapses, where the use of multiple spikes with varying amplitudes are adopted. The programming scheme is proposed for 1R devices. In this programming scheme, pre- and post-synaptic spikes overlap and, according to their relative timing, the voltage drop across the synapse is either positive or negative. Since the considered OxRAM technology is bipolar (see Section 1.2.4), this results in either a SET or a RESET operation, respectively. The amplitude of the resulting SET or RESET programming pulse also varies according to the relative spike timing. Tuning the applied SET voltage amplitude results allows modulating the current flowing through the 1R device, and so multiple LRS states can be achieved. Tuning the RESET voltage allows to modulate the HRS states, leading to the STDP rule shown in Fig. 4.1b. Nevertheless this approach leads to different drawbacks. It implies that each neuron must generate pulses with varying amplitudes, thus leading to additional overhead in the neuron circuitry. Furthermore, this approach requires, for every spike event, the generation of unnecessary programming pulses that are not actually used to program the synapse. This, as previously discussed in Section 2.1, leads to excessive power consumption due to charging long interconnect metal lines in the case of large synaptic arrays. The lack of a selector transistor, associated to the use of unnecessary programming pulses, can also lead to programming disturb to the other synapses of the array. An alternative multilevel solution, that partially addresses these limitations, by avoiding the use of unnecessary programming pulses, has been proposed by Ambrogio *et al.* in [188]. A schematic illustration of this solution is shown in Fig. 4.2. It relies on a 1T1R synapse structure and two distinct

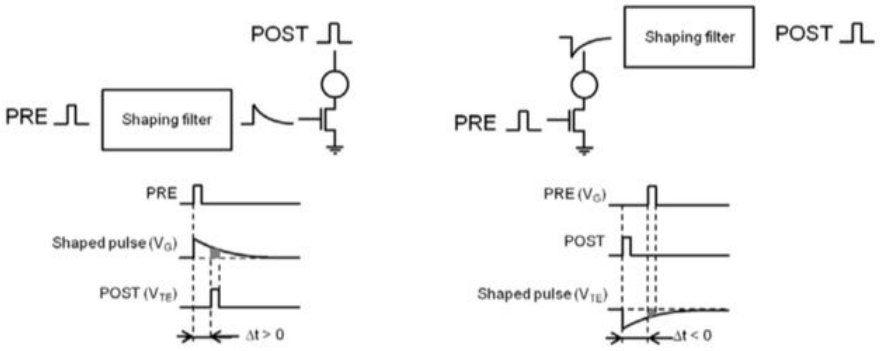


Figure 4.2. Multilevel programming scheme for 1T1R synapse in the LTP regime (left) and the LTD regime (right). Adapted from [188].

spike configurations are implemented for LTP and LTD. In LTP configuration, the pre-synaptic spike consists of an exponentially decreasing positive voltage spike applied to the gate of the transistor. The post-synaptic spike is a square pulse applied on the top electrode terminal of the 1T1R device. The effective voltage on the gate of the transistor during the resulting SET programming pulse is a function of the time difference between post and pre spike. This allows to modulate the compliance current during the SET operation, thus achieving multilevel LRS. In LTD configuration, the pre-synaptic pulse consists of a square positive pulse, applied on the gate of the transistor. The post-synaptic spike is a negative voltage pulse, exponentially decreasing in absolute value. According to the time difference between pre- and post-synaptic spike, the voltage drop across the device during the resulting RESET operation is modulated, thus achieving multiple HRS states. The main disadvantage of this approach lies in the fact that the choice between the LTP and the LTD configuration is done *a priori*, before the timing difference between pre- and post-spike are computed. Even if partial LTP-only or LTD-only learning rules have been proposed for simple visual applications [99], in most systems the possibility of dynamically switch between LTP and LTD in a full STDP learning rule is required. For this reason, the attractiveness of this synaptic design is limited. A multilevel solution that overcomes this limitation has been proposed by Wang *et al.* [189] from the same research group. The schematic of the proposed synapse is shown in Fig. 4.3 and it features a two-transistor-one-resistor (2T1R) structure. Thanks to the use of two transistors (communication gate and fire gate) it is possible to achieve LTP and LTD according to the timing of the pre-synaptic (V_{TE}) and post-synaptic (V_{FG}) spikes. This implementation comes at the cost of a lower integration density, due to the adoption of two transistors per synapse, and a relatively complex neuron spike shape in order to achieve bio-realistic STDP learning rule. It has been shown that using simpler waveforms with rectangular shape, it is possible to obtain a simplified binary STDP learning rule [189]. Always in the framework of the multilevel analog

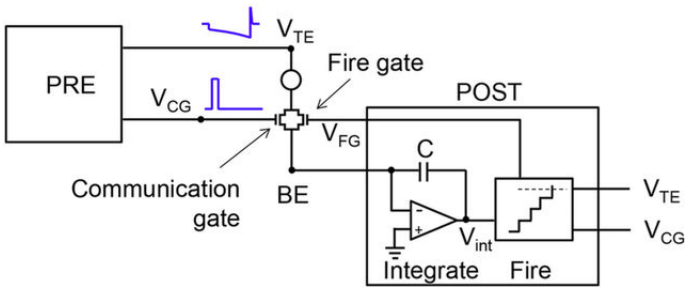


Figure 4.3. Multilevel programming scheme for 2T1R OxRAM synapse. Adapted from [189].

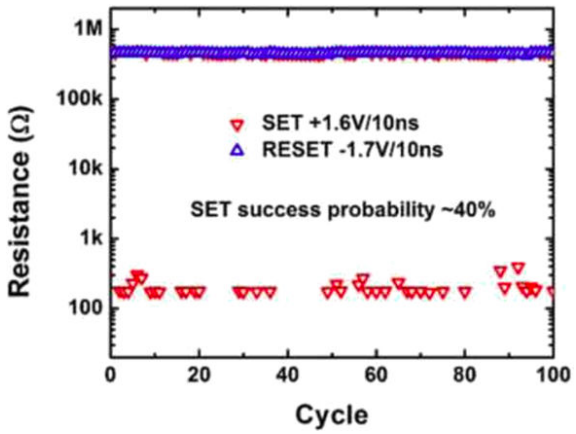


Figure 4.4. Example of intrinsic stochastic SET transition under weak programming condition for binary OxRAM synapse. Adapted from [125].

approach, it has been demonstrated in literature that it is possible to achieve gradual modulation of the conductance of OxRAM structures by applying identical LTP and LTD pulses in the case of bilayer TiO_x/TiO_y structures (10 ms programming pulse width) [92] or perovskite $Pr_{0.7}Ca_{0.3}MnO_3$ (PCMO) (pulse width down to 10 µs) [93].

An alternative strategy to the multilevel OxRAM programming scheme is the binary approach. In this programming methodology, only two distinct resistive states of the OxRAM device, LRS and HRS, are exploited. For unsupervised learning, this approach is associated to a probabilistic STDP learning rule, as explained in Section 1.3.2. This programming strategy has been adopted for OxRAM devices by Yu *et al.* in [125]. Figure 4.4 shows an example of stochastic switching in OxRAM devices, where switching probability $p_{SET} < 1$ is obtained by exploiting the intrinsic stochasticity of the device when operated using weak programming conditions. Alternatively, since controlling the intrinsic switching probability is not a trivial task, since it strongly depends on the technology and the programming conditions, an external source of stochasticity can be implemented using a Pseudo-Random Number Generator (PRNG) circuit. The use of a PRNG circuit to implement extrinsic

stochasticity offers the advantage of a full control on the switching probabilities for both LTP and LTD, independently on the OxRAM technology and the selected programming conditions. This advantage comes at the cost of additional design complexity and on-chip area consumption. It should be noted however that the PRNG is a resource shared at system level, and a single PRNG block can be used to implement the required switching probability for all synapses of a neuromorphic system. Compared to the multilevel synapse approach, the binary approach offers the great advantage of relying on simple SET and RESET programming pulses. These programming pulses are optimized for speed and low power consumption, and are thus ideal for the implementation of an energy efficient hardware implementation of an artificial neural network. However, the use of only two resistance levels per synapse, with respect to the multi-level approach, can be insufficient to achieve good performances in neuromorphic systems designed for some complex applications, as for example image recognition [190].

4.2 Multilevel synapse with binary OxRAMs in parallel

In the previous section, we introduced the advantages and disadvantages of implementing artificial synapses with OxRAM devices using multi-level and binary approaches, using one device per synapse. In this section, we propose a solution based on a "hybrid" approach, which tries to unify the advantages of both multi-level and binary approach. In this solution, a single synapse is composed of n multiple binary OxRAM cells operating in parallel. The model which we refer to is schematically represented in Fig. 4.5: all the devices on the same row, connected in parallel, build-up an equivalent synapse which connects a pre-synaptic neuron (neuron A) to a post-synaptic neuron (neuron B). Since parallel conductance sum up, the conductance of the equivalent synapses ranges from the sum of the n conductances in the HRS to the sum of all the n conductances in the LRS. This strategy provides the opportunity to build an analog-like conductance behavior for a binary device, at the cost of an increased number of devices needed to build a synapse. This approach offers the advantage of a simple programming methodology for the OxRAM devices, in which standard SET and RESET pulses, optimized for high endurance and low-power consumption, are used to switch the device resistance from LRS to HRS and vice versa. A similar concept, based on the use of multiple binary devices in parallel to obtain multilevel behavior, was independently developed by Bill and Legenstein from Graz University of Technology, Austria. Their work on the *compound memristive synapse model*, was published in *Frontiers in Neuroscience* on December 16, 2014 [190]. Our work was originally presented at the 2014 International Electron Devices Meeting (IEDM), held in San Francisco on December 15–17, 2014 [59].

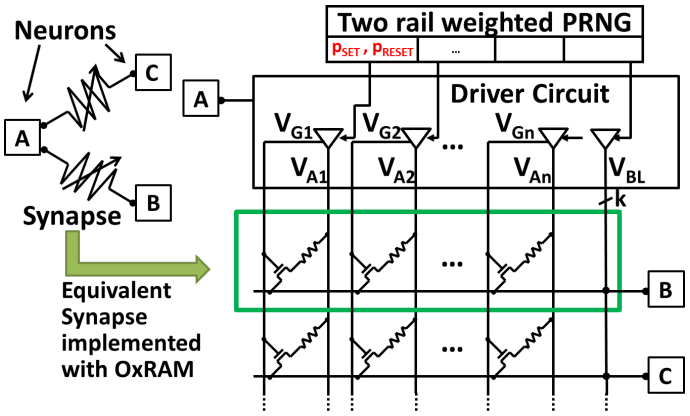


Figure 4.5. Schematic of OxRAM-based synapses used for convolution in CNN. All the OxRAM devices on the same row build one equivalent synapse. Driver circuit is used to individually program OxRAM devices and propagate spikes to next neuron layer. The weighted PRNG is used for on-line learning, to implement extrinsic stochasticity in probabilistic STDP learning rule.

In order to define the resistance state (LRS or HRS) of each OxRAM device needed to obtain the desired equivalent synaptic conductance, two alternative approaches can be used: supervised or unsupervised learning. Supervised learning is obtained using backpropagation algorithm [191], where the LRS/HRS status of each OxRAM device is determined with computer simulations (off-line learning), then discretized and imported in the memory array with a one-time programming operation. In unsupervised learning, the LRS/HRS status of the devices is learned in-situ (on-line learning) with the stochastic STDP learning rule shown in Fig. 4.6a. According to the difference Δt of the spiking time of the post-neuron (t_s) and the pre-neuron (t_x), a Long Term Potentiation (LTP) or a Long Term Depression (LTD) operation is carried out. An LTP (LTD) operation consists in applying to each device of the equivalent synapse a SET (RESET) operation with a probability p_{SET} (p_{RESET}). In the considered range of programming conditions, the studied devices do not show *intrinsic stochasticity*, i.e. switching probability is equal to 1. *Extrinsic stochasticity* is thus obtained using an external Pseudo Random Number Generator (PRNG) circuit block, which provides tunable switching probabilities p_{SET} and p_{RESET}. *Intrinsic stochasticity* can be envisioned by using weaker programming conditions [105], [125]. The driver circuit block can be used to individually program the OxRAM devices.

4.2.1 LTP and LTD curves on OxRAM synapses

In order to validate the functionality of the proposed synapse design, we carried out simulations of LTP and LTD operations on OxRAM synapses composed by a variable number of devices connected in parallel. Figs. 4.6 (b), (c), (d) and (e) show the evolution of the conductance corresponding to 100 LTP followed by 100 LTD

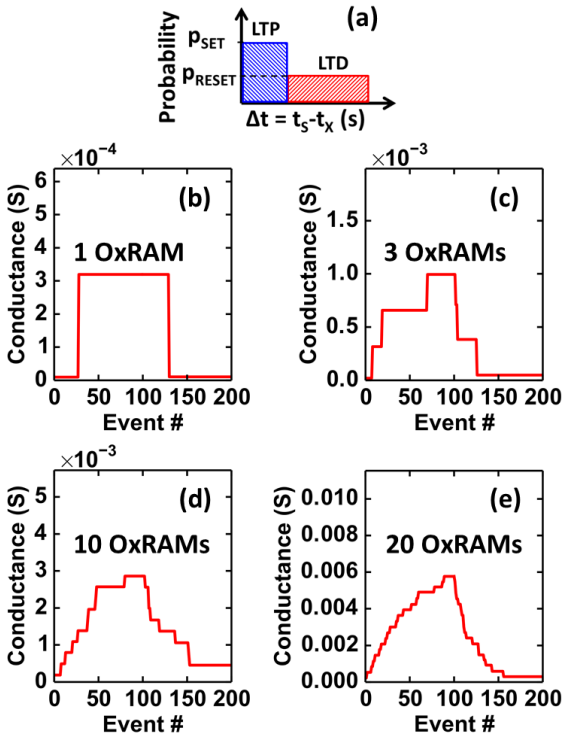


Figure 4.6. (a) Probabilistic STDP learning rule. 100 consecutive LTP and LTD events, with $p_{\mathrm{SET}} = 0.02$ and $p_{\mathrm{RESET}} = 0.04$ on a synapse composed of b) 1 OxRAM device, c) 3 OxRAM devices, d) 10 OxRAM devices and e) 20 OxRAM devices connected in parallel. The use of multiple devices allows to implement a multilevel equivalent synapse, and increasing the number of devices connected in parallel increases the number of intermediate conductance levels. It should be noted that the vertical axis scale is not constant.

operations for a synapse composed of b) $n = 1$ OxRAM device, c) $n = 3$ OxRAM devices, d) $n = 10$ OxRAM devices and e) $n = 20$ OxRAM devices connected in parallel, using a binary probabilistic approach with $p_{\mathrm{SET}} = 0.02$ and $p_{\mathrm{RESET}} = 0.04$.

In the case of (b) a single OxRAM device, obviously, only two conductance levels can be achieved. Using multiple OxRAM devices (c, d, e) allows to obtain a gradual modulation of conductance, with a behavior which is similar to an analog approach. Increasing the number of devices connected in parallel increases the number of intermediate conductance levels. Note that, for the same number of intermediate conductance levels, using multiple OxRAM devices does not necessarily introduce a penalty in power consumption with respect to a single analog device. In fact the number of switching events needed to program the synaptic weight is the same in the case of a single analog synapse and multiple binary OxRAMs in parallel (it is n switching events times 1 device and 1 switching event times n devices for the analog and binary approaches respectively). Achieving multiple conductance levels

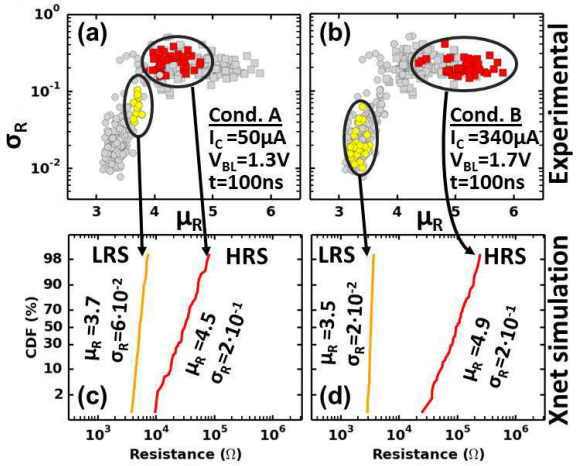


Figure 4.7. Experimental resistance levels and associated variability for (a) weak and (b) strong programming conditions. (c-d) Corresponding simulated synaptic distributions introduced in Xnet.

with multiple devices in parallel has the advantage of enabling a multilevel behavior in a way which is independent on technology: the fabrication of reliable nanoscale synaptic devices, featuring continuous conductance changes, has turned out to be a challenging task [190].

Thanks to the understanding of the origin of OxRAM devices resistance variability explained in Chapter 3, it is possible to associate to each OxRAM programming condition a mean value of resistance and its corresponding variability. As shown in Fig. 4.7, weak programming conditions result in smaller programming window (i.e. smaller separation between the distributions of HRS and LRS) and larger variability (Fig. 4.7 a and c). Stronger programming conditions, on the other hand, result in larger programming window and tighter distributions showing better variability for LRS (Fig. 4.7 b and d). Figure 4.8 shows the impact of the choice of the programming condition on the conductance evolution of the synapses. Light grey curves are the conductance response of 25 synapses composed of 20 OxRAM devices each, when 100 LTP and 100 LTD operations are performed consecutively with $p_{\mathrm{SET}} = 0.02$ and $p_{\mathrm{RESET}} = 0.04$. Red curves are the mean conductance over 25 synapses. When stronger programming conditions are used (condition B), the associated larger programming window allows achieving a wider range of conductance values with respect to weaker programming conditions (condition A). The quantities G_{max}, i.e. the average conductance after 100 LTP events, and ΔG, i.e. the difference between the maximum and minimum conductances on a set of 25 synapses, have been extracted for the two conditions. Due to the fact that a probabilistic learning rule is used, the impact of the device variability on the synaptic conductance response plays a secondary role with respect to the stochasticity introduced by the probabilistic STDP learning rule. In fact a ratio $\Delta G/G_{\mathrm{max}} \approx 32\%$ is obtained for both programming

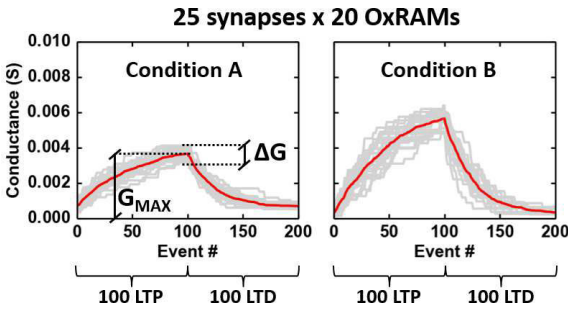


Figure 4.8. Conductance evolution corresponding to 100 consecutive LTP and LTD events. Grey lines are representative of 25 synapses composed of 20 OxRAMs each, programmed with (a) weak programming conditions (Condition A Fig. 4.7) and (b) strong programming conditions (Condition B Fig. 4.7).

conditions.

4.3 Convolutional Neural Network architecture

In the previous section, we proposed a solution for the implementation of an OxRAM-based synapse, which allows to obtain a multilevel-like behavior using devices operated in binary mode and connected in parallel. In this section, we will show how such OxRAM synapses can be used for the hardware implementation of spiking Convolutional Neural Networks (CNNs). As discussed in Section 1.3.4, software implementations of CNNs are currently the best solutions for complex visual tasks such as traffic sign people's faces recognition [108], [121]. The use of NVM synapses, and in particular OxRAM devices, would open the way to the efficient implementation of hardware CNNs thanks to their good properties of endurance, speed and low-power consumption [46].

The architecture that we propose is presented in Fig. 4.9, and it is designed following the software design guidelines for CNNs provided by Simard *et al.* [191]. Our test-bench application is the recognition of handwritten digits of the Mixed National Institute of Standards and Technology (MNIST) database [192], which is commonly used [112]. The database contains 60 000 training images and 10 000 testing images. The proposed architecture is composed of a feature extraction module, made of two cascaded convolutional layers, and a classification module, made of two fully connected layers. While in the fully connected classification module the neurons of a given layer are connected to every neuron of the previous layer by a large number of synapses, in convolutional layers a small set of synapses (constituting several kernels) is shared among different neurons to connect layer N and $N+1$ through a convolutional operation, as described in Section 1.3.4. A convolutional layer is composed of several feature maps, each of them being connected to the feature maps of the previous layer through a convolution kernel. The kernel corresponds to a

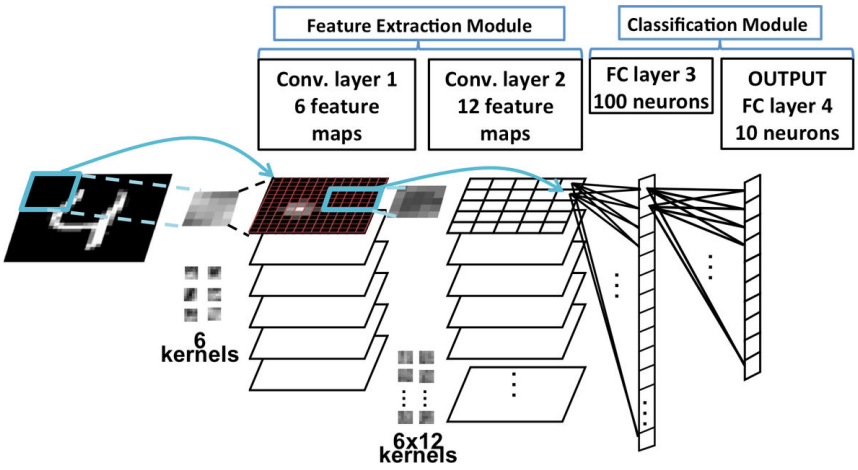


Figure 4.9. CNN architecture for handwritten digits recognition.

feature that has to be localized in the input image. In a layer, each feature map contains the results of the convolution of the input maps (which are the output feature maps of the previous layer), each of them with a different convolution kernel. It contains information about the locations where the kernel features are present in the input map. The feature extraction module therefore transforms the input image into a simpler set of feature maps. The classification module connects the obtained set of feature maps to the output layer. The first convolutional layer is composed of 6 feature maps of size 13×13 (169 neurons). The second convolutional layer is composed of 12 feature maps with size 5×5 (25 neurons). The third layer of the network, with fully connected topology, is composed of 100 neurons. The output layer is composed of 10 neurons, where each neuron is associated to one of the 10 digit categories. For the first convolutional layer, 6 kernels composed of 5×5 synapses are used to carry out the convolution operation

The designed architecture has a structure equivalent to the structures used for software implementations of CNNs. The main difference resides in the way the convolution operation is carried out. Mathematically, a discrete convolution operation, for a kernel with size $k \times k$, consists of a series of multiply-accumulate (MAC) operations shown schematically in Fig. 4.10, and described with the following equation:

$$F_{i,j} = \sum_{p=0}^{k-1} \sum_{q=0}^{k-1} V_{i+p,j+q} \cdot K_{p,q} \tag{4.1}$$

where $F_{i,j}$ is the brightness value at coordinates i, j in the feature map, $V_{i+p,j+q}$ is the brightness values at coordinates $i+p, j+q$ in the receptive field of the input image, $K_{p,q}$ is the kernel coefficient at coordinates p, q. p and q vary between 0 and $k-1$. $k = 5$ is the size of the kernel in our case.

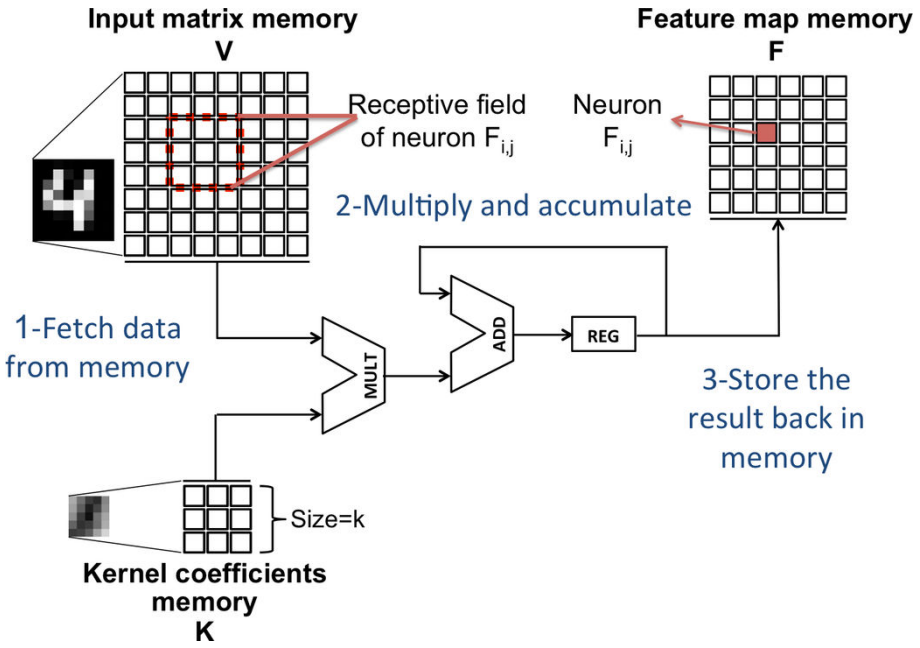


Figure 4.10. Schematic representation of the multiply accumulate circuit block for the implementation of convolution operation, using a Von Neumann approach [193].

In a conventional Von Neumann architecture, the convolution operation would be carried out using digital multipliers, adders and registers. In Fig. 4.10 the building blocks of a possible digital circuit for the convolution operation are shown. The operands V and K are stored in memory as numbers in digital format. At every clock cycle, these data have to be retrieved from the system memory and stored back in memory after computation. This process has to be repeated N_c times, according to the following equation:

$$N_c = k^2 \cdot f^2 \cdot N_K \cdot N_F \cdot N_{cl} \tag{4.2}$$

where k is the size of one kernel, f is the size of one feature map, N_k is the number of kernels in one convolutional layer, N_F is the number of feature maps in one convolutional layer, and N_{cl} is the number of convolutional layers in the network. Figure 4.11 reports the estimated number of clock cycles associated to the retrieval and storage of data in memory for MAC operations in a Von Neumann CNN, as a function of the number of kernels and feature maps, for $k = 5$, $f = 20$, $N_{cl} = 2$. We can observe that the number N_c can rapidly increase for large networks. This leads to an unwanted latency in computation, i.e. the so called *memory bottleneck*, peculiar to the Von Neumann architectures, as described in Chapter 1. In the case of a state-of-the-art convolutional neural network for the recognition of traffic signs with 8 bit synapses [194], for example, a Von Neumann implementation would require $\approx$ 125 million clock cycles for the recognition of one image. This would

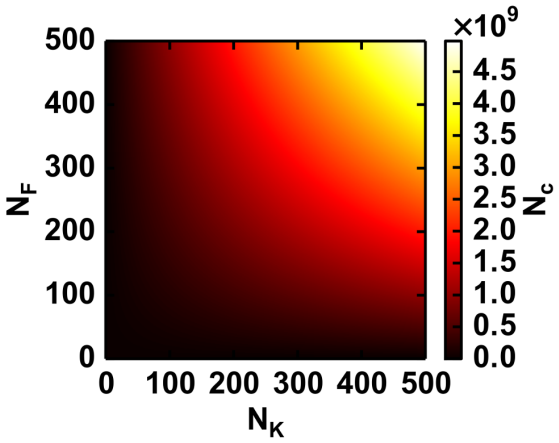


Figure 4.11. Estimated number of clock cycles associated to the retrieval and storage of data in memory for MAC operations in a Von Neumann CNN, as a function of the number of kernels N_{K} and feature maps N_{F}.

correspond to a latency of 625 ms assuming an operating frequency of 200 MHz. Possible solutions to mitigate the memory bottleneck, always in the framework of Von Neumann architectures, are the following:

- increasing the data level parallelism using Single Instruction, Multiple Data (SIMD) instructions [195];
- increasing the number of processing cores, assuming distributed memory [193].

Increasing data parallelism by 16 times, using 128 bit memory bus, and using 16 processing cores in parallel, would reduce the required number of cycles to $\approx 500\,000$ which, at 200 MHz frequency, corresponds to a latency equal to 2.5 ms per image recognition [193]. Smaller latencies could be achieved using higher frequencies in Graphics Processing Units (GPUs) that allow even larger parallelism, but these solutions are not viable for implementation in embedded systems. In fact power consumption of high-end GPUs is in the order of 250 W [193].

In our solution, thanks to the use of OxRAM synapses in a spiking neural network, the MAC operations required for convolutions are performed directly in memory, in a fully parallel and distributed approach. Specifically, the multiplication is carried out using the simple Ohm's law:

$$I_{\mathrm{OUTPUT}} = V_{\mathrm{SPIKE}} \cdot G_{\mathrm{kernel}} \tag{4.3}$$

where V_{INPUT} is a voltage that, using a proper encoding, represents the input image. G_{kernel} is the conductance of an OxRAM synapse which represents the kernel feature and I_{OUTPUT} is the current that has to be accumulated at the output feature map neuron.

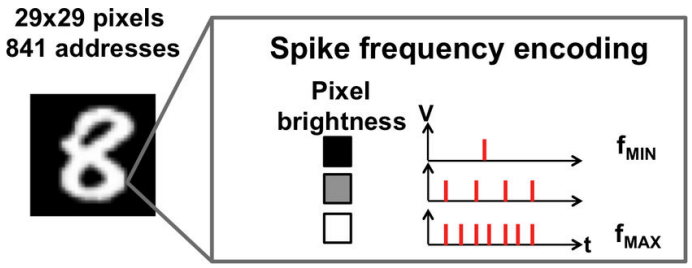


Figure 4.12. Schematic illustration of the spike encoding rule adopted to convert a static input image to AER representation.

We focus here on the first factor V_{SPIKE} of Eq. (4.3), i.e. the voltage encoding of the input image. Since the task of the network is to recognize static images, we have to first convert the static images into voltage spikes that can be fed to the artificial neural network. Figure 4.12 illustrates schematically the conversion algorithm that we adopted. The input images are composed of 29×29 pixels. Each pixel's brightness is converted into a voltage spike train with a given frequency, during a time slot $t = 1\,\mu\mathrm{s}$. The lowest pixel brightness (i.e. black pixel) is converted to the lowest spiking frequency $f_{\mathrm{MIN}} = 1\,\mathrm{MHz}$. The highest pixel brightness (i.e. white pixel) is converted to the highest spiking frequency $f_{\mathrm{MAX}} = 8\,\mathrm{MHz}$. All the gray-scale, intermediate pixel brightness values are linearly converted into spiking frequency between f_{MIN} and f_{MAX}. The static input image is thus converted into an Address-Event Representation (AER) format, where each pixel is associated to a neuron address (i, j coordinates with i, j varying from 0 to 28 or equivalently, a sequential address from 0 to 840) and a list of events in time (voltage spike train). This approach has the advantage of being compatible with bio-inspired sensors such artificial retinas [106] studied in Chapter 2.

Now that the input image has been converted into AER format, let's focus on the second factor G_{kernel} of Eq. (4.3). The kernel is a collection of $k \times k$ synaptic weights, representing a feature to be convoluted with the input image. We propose to implement the kernel in hardware using the OxRAM array presented in Fig. 4.13. In this array, each row represents one of the $k \times k$ synaptic weights of the kernel and, at each row, an OxRAM-based synapse composed of n devices connected in parallel is implemented. In the case of a 5×5 kernel, 25 rows are needed. As we have seen in Section 4.2.1, using n devices in parallel allows us to obtain a multilevel analog behavior, with an equivalent conductance that is tunable between G_{MIN} and G_{MAX}, for a total number of conductance levels equal to n.

We have examined the factors V_{SPIKE} and G_{kernel} of Eq. (4.3). Figure 4.14 explains how the resulting current I_{OUTPUT} is obtained. When a spike V_{SPIKE} occurs at coordinates x, y in the input image, an address decoder is used to dynamically map the kernel synapses to the feature map neurons that have the input neuron x, y in their receptive field. The spike is then propagated through the synapses of the

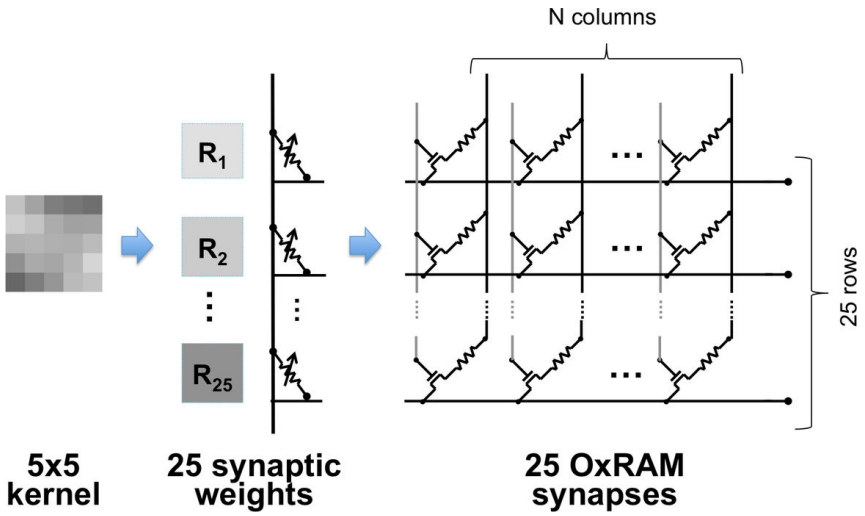


Figure 4.13. Proposed hardware implementation of convolutional kernel using OxRAM synapses.

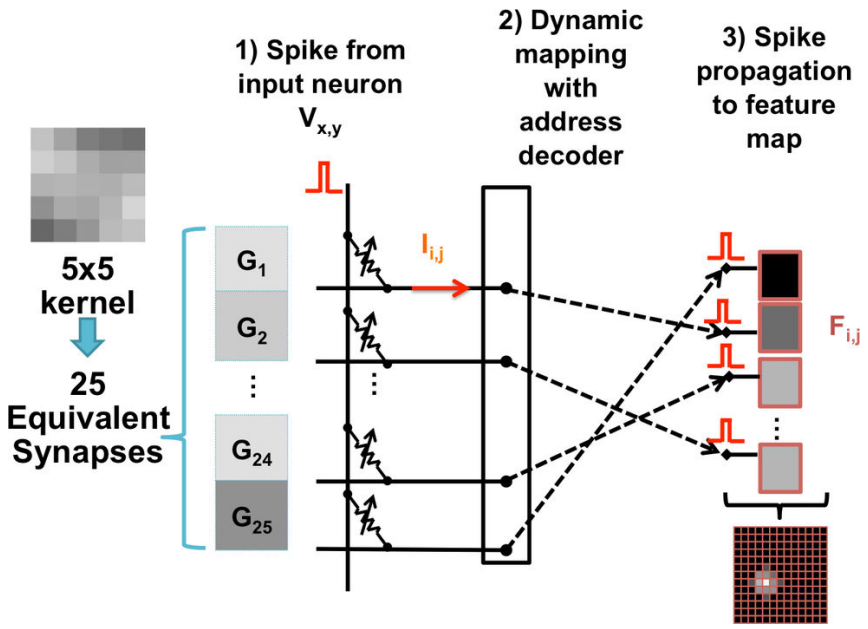


Figure 4.14. Spike propagation through synaptic kernel. The address decoder is used to dynamically map the kernel synapses to the feature map neurons that have the input neuron x, y in their receptive field.

kernel to the mapped Integrate and Fire (IF) neurons. The IF neurons accumulate (integrate) the incoming current over time, and will fire when a given threshold is reached. The spiking frequency $S_{\mathrm{F}i,j}$ of the feature map neuron $F_{i,j}$ at coordinates

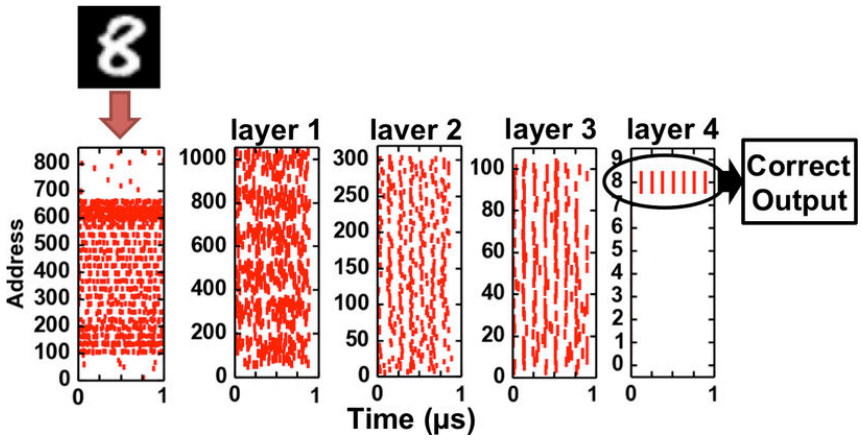


Figure 4.15. Propagation of spiking activity through CNN neuron layers.

i, j will thus be given by the following formula:

$$S_{\mathrm{F}i,j} \propto I_{\mathrm{OUTPUT}i,j} = \sum_{p=0}^{k-1} \sum_{q=0}^{k-1} S_{V\,i+p,j+q} \cdot G_{K\,p,q} \tag{4.4}$$

where $I_{\mathrm{OUTPUT}i,j}$ is the accumulated current at the node i, j. $S_{V\,i+p,j+q}$ are the spiking frequencies of the input neurons in the receptive field of the FM neuron i, j. $G_{K\,p,q}$ are the conductance values of the synaptic kernel, with p, q varying between 0 and $k-1$. We can thus observe that there exists an equivalence between the discrete formula of the convolution Eq. (4.2) and the spiking implementation in Eq. (4.4).

In order to validate the functionality of our design for the recognition task of the MNIST database, we performed simulations using the special-purpose spiking neural network simulator Xnet [165], [166], using synapses composed of $n = 20$ OxRAM devices connected in parallel. In order to define the resistance state of each OxRAM device we used the supervised backpropagation learning algorithm. Figure 4.15 shows an example of the propagation of the spikes through the layers of the CNN, from the input to the output layer, when a test image representing the handwritten digit "8" is presented to the network. At the input layer, the static image is converted in AER format, with neurons spiking at different frequencies according to the brightness of the corresponding image pixel. The signal are propagated through the network until the output layer, where the neuron with the highest spiking frequency (neuron number 8 in this specific case) indicates the category in which the input image has been categorized by the network.

4.3.1 Impact of OxRAM programming conditions

In order to study the impact of OxRAM programming conditions on the network performance, we performed a thorough simulation work. As demonstrated in Fig. 4.7

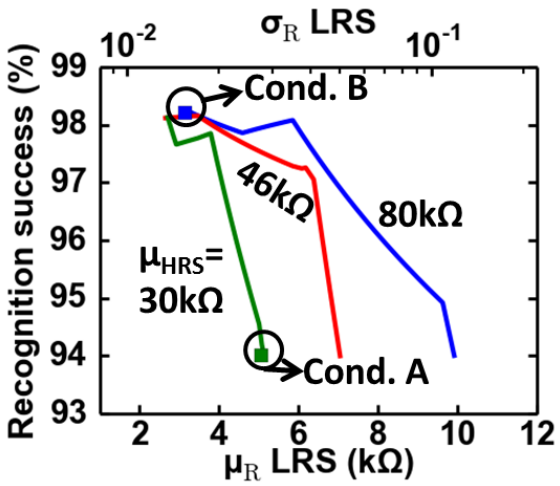


Figure 4.16. Recognition success for the network operated in read-mode as a function of LRS for different HRS. The kernels are defined using backpropagation algorithm. Both LRS and HRS variability are taken into account. Highlighted points correspond to weak and strong programming conditions of Fig. 4.7 a and b.

the device variability strongly depends on the programming conditions: the weaker the programming conditions, the larger the variability. Fig. 4.16 reports the accuracy of the CNN network in the recognition of the 10,000 MNIST handwritten digits as a function of the LRS mean value (bottom axis) and the associated device variability (top axis) for different HRS values. For this study, synapses composed of 20 OxRAM devices have been used and the kernels have been defined using backpropagation algorithm. The recognition rate slightly improves by decreasing the LRS value and the associated variability as well as by increasing the HRS. A recognition rate higher than 94% is achieved for all the studied programming conditions. Table 4.1 reports a summary of the performance of the CNN for the programming conditions highlighted in Fig. 4.7. For strong programming conditions (Fig. 4.7 b), a performance of 98.3% correctly recognized digits is achieved. Using weak programming conditions (Fig 4.7 a) the network performance is degraded (94% recognized digits), but the switching energy is reduced from 60 pJ to $<$10 pJ.

4.4 Unsupervised learning

In the previous section, we proposed a novel CNN architecture based on OxRAM synapses and we validated the functionality of the proposed design in Xnet simulations. In order to define the synaptic weights and therefore the resistance state of each OxRAM device, we used supervised learning with backpropagation algorithm.

	Cond. A Fig. 4.7 a	Cond. B Fig. 4.7 b
SET energy / dev.	5pJ	34 pJ
RESET energy / dev.	9pJ	58 pJ
Recognition success	94.0%	98.3%

Table 4.1. Summary of programming energy and network accuracy for weak and strong OxRAM programming conditions

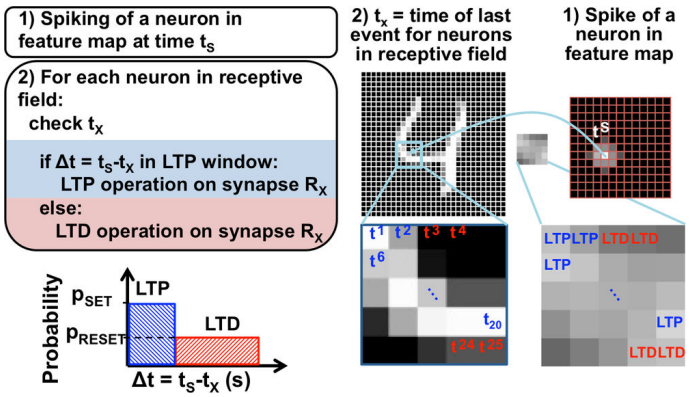


Figure 4.17. Proposed implementation of STDP for on-line, on-field learning. Synaptic weights are changed with SET/RESET pulses applied on OxRAM devices.

In this section, we explore the possibility of implementing unsupervised STDP learning on the studied CNN architecture. Figure 4.17 explains the proposed STDP learning rule that we adopted. The learning performances obtained in simulation are, as expected [108], worse than the performances obtained with supervised learning. Even if the recognition rate is low ($< 80\%$), the network is functional and learning of kernel features can occur in unsupervised way. We have also evaluated the proposed OxRAM based synapse on a Fully Connected binary probabilistic Neural Network for visual pattern recognition, having a similar number of connections of our CNN but a larger number of synapses due to the fully connected topology. Table 4.2 reports the learning statistics comparison between CNN and Fully Connected Neural Network approaches. Statistics were obtained for unsupervised STDP learning over a database of 60,000 patterns, with a spike-frequency encoding of each pattern on a time-frame of 1 µs, for a total learning time of 60 ms. Given the same number of connections in the network, the amount of programming events per device is up to 3 orders of magnitude higher for CNN, due to shared weights [59]. Device endurance, discussed in Section 3.3.1, becomes therefore a critical factor for spike-based learning in CNN, with an estimated endurance requirement of $> 10^5$ for a relatively small database like MNIST. The CNN approach allows reducing the number of synapses

	Learning phase duration: 60 ms 60 000 patterns of 1 µs	
	Fully Connected NN	Convolutional NN
Nb. of connections	$9.5 \cdot 10^4$	$8.6 \cdot 10^4$
Nb. of synapses	$9.5 \cdot 10^4$	$8.8 \cdot 10^3$ (shared)
Average SETs / synapse	$2.4 \cdot 10^2$	$4.2 \cdot 10^5$
Average RESETs / synapse	$5.0 \cdot 10^2$	$2.2 \cdot 10^5$

Table 4.2. Comparison of learning statistics for CNN and Fully Connected Network for STDP learning.

(memory array size), resulting in smaller neuron fan-out and parasitic capacitance, which implies thus easier hardware implementation. The estimated memory array size needed to implement in hardware the proposed CNN architecture for MNIST database recognition, using 10 OxRAM per synapse, is in the order of 600 kb. This is an attainable goal for current 1T1R technology capabilities, with state-of-the-art demonstrators with size in the order of 16 Gb [196]. However, for more complex applications, a larger number of synapses is needed, as we will describe in Section 4.5.

4.5 Synaptic weight resolution

In Section 4.2.1 we have demonstrated that using more OxRAM cells per synapse increases the synaptic weight resolution (Fig. 4.6), but comes at the cost of more complex process integration. We therefore study here the impact of the number n of OxRAM devices per synapse on the performance of the CNN presented in Section 4.3. A parametric simulation has been performed, varying the parameter n and keeping all the other parameters of the network constant. Both weak and strong programming conditions corresponding to Condition A and B in Fig. 4.7 are used. Figure 4.18 reports the recognition success of 10 000 handwritten digits after learning with backpropagation algorithm [191] as a function of the number n of OxRAM devices used to implement each synapse in the network. The recognition success improves as n increases for both programming conditions and for n higher than 12 the maximum network performance greater than 97% is reached. For weak programming conditions (red curve), the recognition error is slightly larger than the one obtained for strong programming conditions (blue curve) for all values of n. However, the difference is lower than 7% for n larger than 12.

We have also evaluated how the synaptic resolution requirements change for more complex applications than MNIST, i.e. the recognition of traffic signs of the German

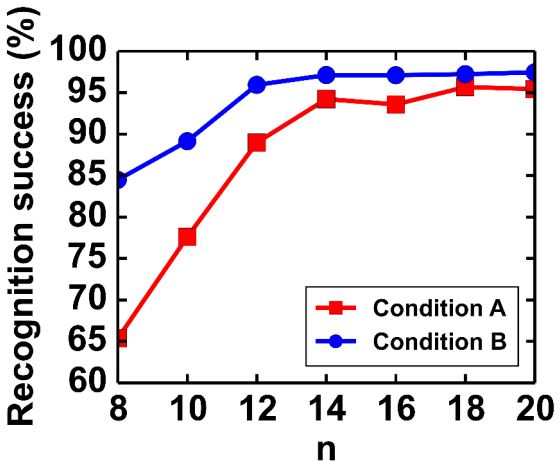
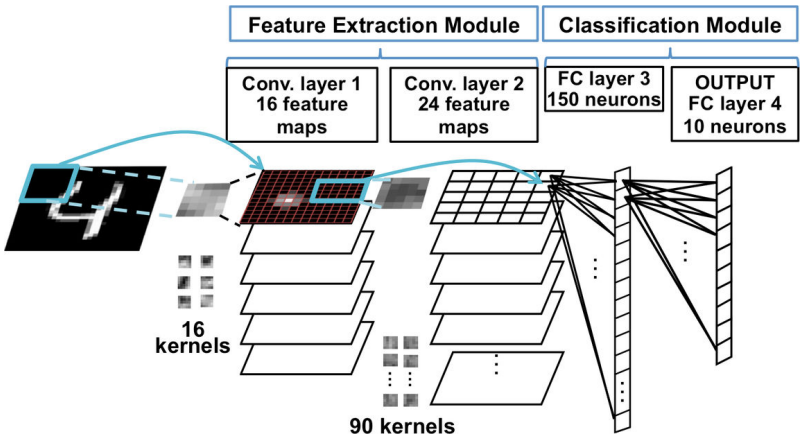
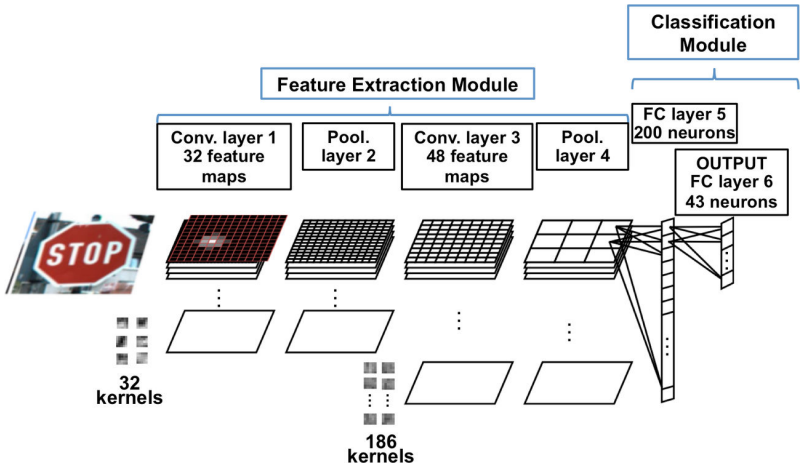


Figure 4.18. Recognition success over 10 000 handwritten digits database as a function of the number n of OxRAM devices used to implement each synapse in the network.

Traffic Sign Recognition Benchmark (GTSRB) database [197]. To do so, we have compared two state-of-the-art CNNs for the recognition of MNIST and GTSRB databases, adapted from the software implementation presented by Ciresan *et al.* in [194]. The proposed architectures are shown in Fig. 4.19a and Fig. 4.19b. They are both composed of a feature extraction module followed by a classification module. The feature extraction module is made of two cascaded convolutional layers, each of them followed by a max-pooling -i.e. subsampling- layer [112] in the case of the GTSRB network. The max-pooling layers reduce the size of the feature maps by a factor 2, thus reducing the complexity of the network. The classification module is made of two fully connected layers. For the MNIST applications 16 (size: 4×4) and 90 (size: 5×5) shared kernels are used in the first and second CNN layers respectively, while for the more complex GTSRB applications 32 (size: 4×4) and 186 (size: 5×5) shared kernels are implemented. In order to define the optimal value of the synaptic weight resolution, i.e. the number (n) of OxRAM devices needed per synapse, simulations have been performed on both CNN architectures. Figure 4.20 shows the simulation results in terms of recognition success as a function of the value n. It appears that more complex application tasks, such as the GTSRB database recognition, are indeed more demanding in terms of number of devices. In fact, in the case of handwritten digits recognition (MNIST), 11 OxRAM devices per synapse are enough to achieve a recognition performance equivalent to the reference recognition success rate obtained with the formal CNN model with floating-point precision synapses. In the case of the more complex recognition task of traffic signs, 20 OxRAM devices per synapse are necessary to achieve a recognition rate equivalent to the reference one. Using $n = 11$ devices per synapse, the estimated size of



(a) MNIST



(b) GTSRB

Figure 4.19. CNN architecture for (a) handwritten digits recognition (MNIST database) and (b) traffic signs recognition (GTSRB database).

the OxRAM array needed to implement the CNNs is 600 kb for MNIST and 1 Mb for GTSRB, respectively. Vertical Resistive Memories (VRAM), which consists of ReRAM cells integrated in multi-layered VNAND-like structure, is a simple and cost effective 3D processes to achieve high memory density [198]. The CNN architecture proposed in this chapter can be adapted using a 1T-nR structure, where 1 access transistor is used to access n OxRAM devices vertically stacked in the back-end of line process using a VRAM integration scheme [199].

4.5.1 Analog vs. digital integration neuron

We have discussed until now about the implementation of the OxRAM synapses. In this section, we focus on the neuron implementation. We consider in this work the

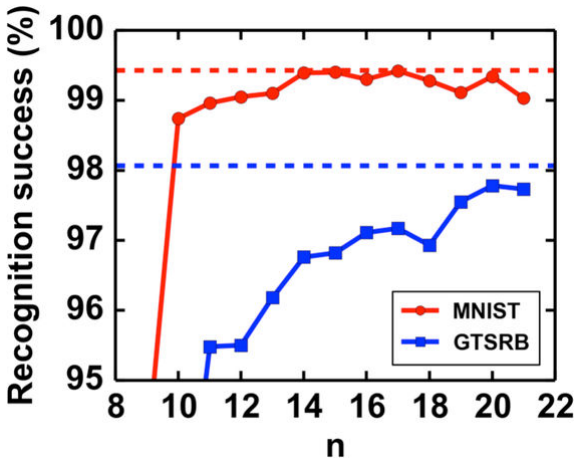


Figure 4.20. Recognition success as a function of the number n of parallel OxRAM devices used to implement an equivalent synapse, using analog neuron model and taking into account c2c and d2d variability. Dashed lines: reference recognition success rate obtained on the testing dataset with the formal CNN model and floating-point precision synapses.

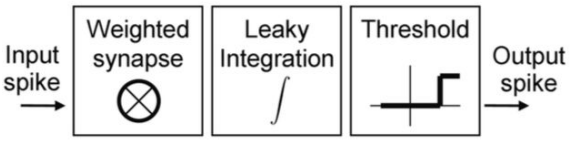
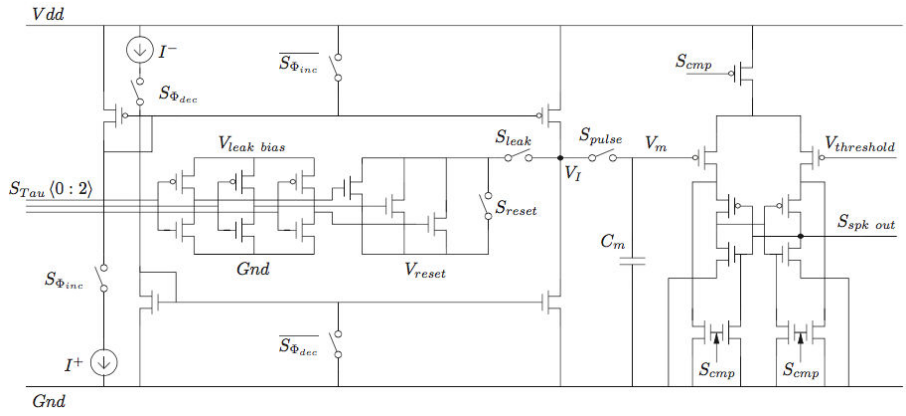


Figure 4.21. Block diagram of leaky-integrate-and-fire neuron [200].

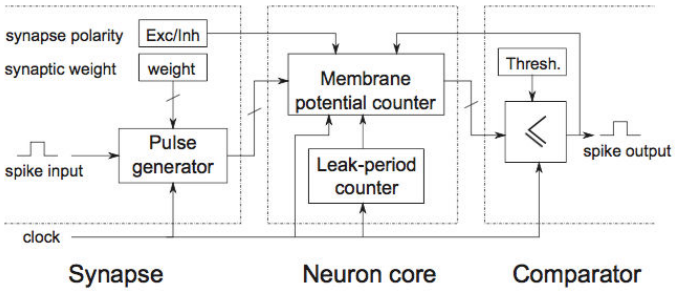
Leaky Integrate-and-Fire (LIF) model. Its main building blocks of a are reported in Fig. 4.21: after propagating through the synapses, the incoming spikes are integrated at the neuron and, when a given integration threshold is reached, an output spike is produced. Two design solutions can be adopted for the implementation of a LIF neuron: analog and digital. Figures 4.22a and 4.22b report the schematic implementations of analog and digital neuron, respectively [200].

- analog integration: a spike is propagated through every OxRAM devices of a synapse in a single step. Devices are selected at the same time and the current of the equivalent synapse is read-out using an analog integration neuron Fig. 4.22a. This solution allows faster spike propagation (because all devices are read at the same time), but it is potentially more sensitive to device variability. Nevertheless, we assumed here to be in an ideal case, thus neglecting CMOS variability and noise.

- digital integration: the spike signal is propagated through the synapse by performing a digital read operation, i.e. a binary sampling of each device and digital integration is performed Fig. 4.22b. This approach is in principle more



(a)



(b)

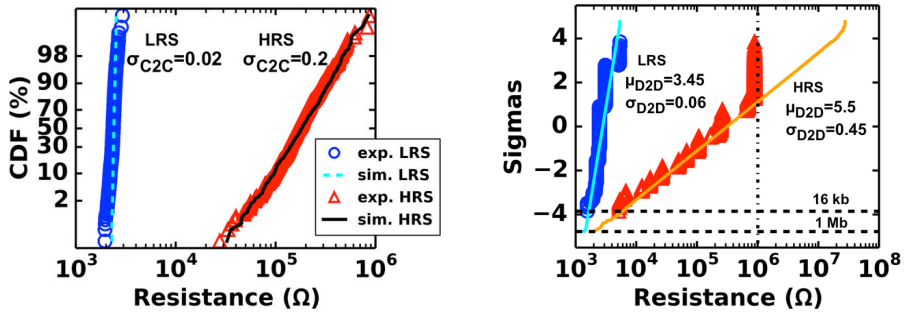
Figure 4.22. (a) Analog and (b) digital LIF neuron schematic [200].

robust to variability. Two digital read-out methods can be used: I/ Sequential, where one device at a time is read sequentially. The LRS/HRS state of each OxRAM is determined by comparing the read-out current to a single threshold. II/ "One-step", where the synaptic weight of the synapse is digitally read-out in a single step by comparing the read-out current to multiple thresholds. The latter solution allows for a faster operation but comes at the cost of a larger reading circuit.

The choice between analog and digital integration neuron schemes is discussed in the next section, in terms of variability, power and area consumption.

4.6 Tolerance to variability

In this section, we focus on the impact of the variability on the large network performance. To do so, we have used the results of both cycle-to-cycle (c2c) and device-to-device (d2d) variability extracted from 16 kb OxRAM array. Simulations have been carried out using the simplified trap-assisted-tunneling model presented in



(a) (b)

Figure 4.23. Experimental and simulated results of (a) cycle-to-cycle and (b) device-to-device synaptic variability.

	c2c variability		d2d variability	
	statistic estimator $[\mathrm{Log}_{10}(R)]$	CF property	statistic estimator $[\mathrm{Log}_{10}(R)]$	CF property
LRS	$\sigma_{\mathrm{C2C}} = 0.02$	Fixed CF shape, random V_O position	$\mu_{\mathrm{D2D}} = 3.45$ $\sigma_{\mathrm{D2D}} = 0.06$	$\mu_{\mathrm{Lgap}} = 1.6\,\mathrm{nm}$ $\sigma_{\mathrm{Lgap}} = 1\,\mathrm{Å}$
HRS	$\sigma_{\mathrm{C2C}} = 0.2$	Fixed CF shape, random V_O position	$\mu_{\mathrm{D2D}} = 5.5$ $\sigma_{\mathrm{D2D}} = 0.45$	$\mu_{\mathrm{Lgap}} = 3.5\,\mathrm{nm}$ $\sigma_{\mathrm{Lgap}} = 3\,\mathrm{Å}$

Table 4.3. Summary of variability statistic estimators and corresponding conductive filament properties.

Chapter 3. The experimental and simulation results for c2c and d2d variability are shown in Fig. 4.23a and Fig. 4.23b, respectively. Figure 4.23a reports the resistance distributions of a single device cycled 10^4 times. Figure 4.23b reports the resistance distributions corresponding to a single programming cycle for a 16 kb population. It is worth noticing that the TAT model allows us to simulate the resistance distributions of large memory arrays of up to 1 Mb. The experimental distribution of HRS is cut at $\approx 1\,\mathrm{M\Omega}$ because of the lower limit in current sensing of the digital controller. A summary of the extracted statistic estimators and the simulated OxRAM conductive filament properties used to model c2c and d2d variability is provided in Table 4.3.

The statistic estimators presented in Table 4.3 have been used to simulate in Xnet simulation the resistance distributions of the 600 kb array needed for the MNIST CNN. The resistance of the OxRAM device numbered i in the network is defined

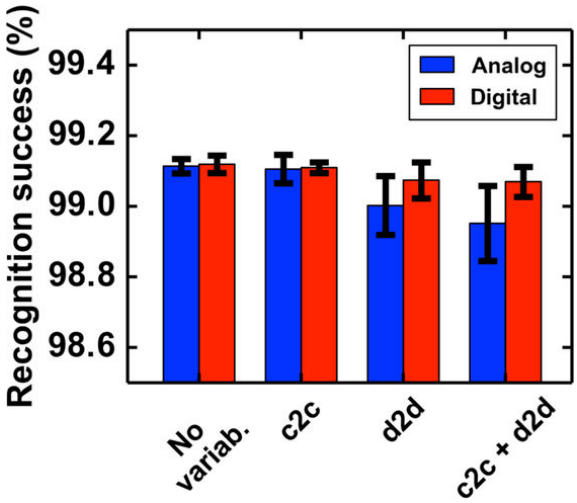


Figure 4.24. Impact of temporal (c2c) and spatial (d2d) variability on recognition success of OxRAM based CNN using analog and digital integration neurons.

using the following relations:

$$\mu_{i,\mathrm{C2C}} = \mathrm{lognorm}\left(\mu_{\mathrm{D2D}}, \sigma_{\mathrm{D2D}}\right) \tag{4.5}$$

$$R_i = \mathrm{lognorm}\left(\mu_{i,\mathrm{C2C}}, \sigma_{\mathrm{C2C}}\right) \tag{4.6}$$

where $\mathrm{lognorm}\left(\mu, \sigma\right)$ is a function that draws a random sample from the base-10 lognormal distribution with parameters μ and σ. Values of statistic estimators μ_{D2D}, σ_{D2D}, σ_{C2C} are extracted from experimental results (Table 4.3). Relations 4.5 and 4.6 allow considering the contributions of d2d and c2c variability in an independent way. In fact, it is possible to suppress c2c (d2d) variability by setting $\sigma_{\mathrm{C2C}} = 0$ ($\sigma_{\mathrm{D2D}} = 0$). By suppressing the effect of both c2c and d2d it is possible to simulate an ideal synapse with neither temporal nor spatial variability.

Fig 4.24 presents the performance of the simulated network on the MNIST database recognition task comparing analog and digital integration neuron models. In this simulation, $n = 11$ OxRAM devices have been used to implement each synapse. $n = 11$ has been chosen based on the results of Fig. 4.20, because increasing the number of devices over $n = 11$ does not lead to better performance, but increases the system complexity. Recognition success is defined as the number of well-recognized digits over the tested 10 000 handwritten digits. Results show that the d2d and the c2c variability do not impact significantly the application performance. Even in the worst (but realistic) case where both d2d and c2c variability are considered, the recognition success is still larger than 98.9%. Indeed, these results indicate the strong tolerance of the CNN to variability, with both digital and analog integration. With digital integration, the effect of variability is suppressed even further and no significant difference is observed between the studied scenarios. It should be observed that a digital integration neuron design offers slightly better variability immunity,

however, digital neurons typically consume more power and need for larger silicon area: using data from [200], we have estimated 41 pJ/spike and 538 µm^2 for the digital neuron versus 2 pJ/spike and 120 µm^2 for the analog neuron.

4.7 Conclusion

In this chapter we have presented a novel design of an OxRAM-based synapse, offering multilevel capabilities using multiple binary HfO_2 devices connected in parallel. Electrical characterization, physical modeling and simulations suggest that OxRAM technology is a good candidate for the hardware implementation of artificial synapses in neuromorphic systems. Using the proposed OxRAM synapses, we have presented for the first time a hardware implementation of a convolutional neural network where the convolution operation is performed directly in memory, overcoming the memory bottleneck of the Von Neumann implementations. A thorough analysis of both cycle-to-cycle and device-to-device variability of OxRAM synapses, extracted from a 28 nm-CMOS OxRAM array data, has been carried out. The impact of device variability on CNN performance has been studied, evaluating both analog and digital integration neurons. Results show that the proposed CNN architecture is highly tolerant to variability with no need of program-and-verify algorithms. Recognition success rates higher than 99% and 97% have been demonstrated for the MNIST and GTSRB networks, respectively, which are similar to the state-of-the-art recognition success rates obtained with formal CNN models, implemented with floating-point precision synapses. Furthermore, the proposed architecture allows to reduce the estimated time required for the recognition of each pattern, considering similar operating frequency. For instance, in the case of the MNIST recognition application, a latency equal to 1 µs per image is estimated for our proposed OxRAM-based CNN, with spike encoding frequency $f_{MAX} = 8$ MHz. The estimated latency per image estimated for a full CMOS, Von Neuman architecture is 2.5 ms, using 16 parallel processing cores with a clock frequency equal to $f_{clock} = 200$ MHz. The obtained results confirm that OxRAM technology is a promising candidate for hardware implementation of spiking, resistive memory-based CNNs.

Chapter 5

Conclusions

In this Ph.D. thesis, we have explored the use of PCM and OxRAM devices as artificial synapses for neuromorphic systems. Firstly, we focused on PCM technology, being the most mature among the emerging non-volatile memory technologies. PCM devices offer the possibility of multilevel programming by gradually changing the size of the crystalline portion of the active phase-change material. This property closely resembles the plasticity of real synapses, so PCM devices were among the first emerging NVMs to be investigated as nanoscale artificial synapse [101], [153]. We analyzed the drawbacks related to the use of the multilevel PCM synapse approach, i.e. the generation of programming pulses with varying amplitude and complex refresh schemes. Therefore, driven by the motivation to overcome the limitations associated to the multilevel programming, we have explored the use of PCM synapses in binary mode [201]. We have investigated PCM as synapses through simulations of fully connected artificial neural network for the detection of cars in a video. We have provided PCM programming schemes for synaptic architectures with- or without-selector transistor. The proposed synaptic programming schemes avoid the use of complex refresh operations and unnecessary programming pulses required by multilevel PCM synapses. Starting from the results obtained from electrical characterization, we have carried out simulations of a large scale artificial neural network for complex visual application. We have demonstrated that, by tuning the resistance levels of the SET and RESET states according to the selected programming conditions, it is possible to tune the power consumption of the system. Specifically, simulation results show that the learning mode power consumption can be dramatically reduced if the RESET state of the PCM devices is tuned to a relatively low resistance. Read-mode power consumption, on the contrary, can be minimized by increasing the resistance values for both SET and RESET states of the PCM devices. These considerations give additional degrees of freedom to system designers, who can properly select PCM programming conditions based on whether the designed system is used mostly in read mode after an initial programming phase, or a continuous learning is required over the life-span of the system, in order to adapt

for example to an input whose nature changes over time. Furthermore, we have investigated the issue of PCM resistance drift and we have proposed a strategy to mitigate this problem. We have also observed that, using scaled devices, it is possible to dramatically reduce the power consumption thanks to the smaller programming current. Summing up, we have successfully demonstrated the interest of using PCM devices in binary mode in neuromorphic systems for visual applications.

Secondly we have considered OxRAM technology, which is a very promising emerging NVM in terms of scalability, low power consumption and speed. Since variability is the main drawback of OxRAM technology, we have carried out an extensive work of electrical characterization on single bitcells and on 16 kb memory array, in order to understand the source of variability. Starting from the electrical characterization results, we have developed a simplified trap-assisted tunneling model able to reproduce the OxRAM variability from low (LRS) to high resistance state (HRS), highlighting the continuity of the mechanisms involved in the variability [202], [203]. We have carried out the analysis of OxRAM variability with a dual goal. On one hand, the developed model provides an insight on the source of variability of OxRAM, providing technology guidelines for the improvement of reliability. Activity is on-going to physically characterize the properties of the conductive filament(s) using *in situ* Transmission Electron Microscopy (TEM) to observe the dynamic switching of OxRAM devices [54]. The engineering of the memory stack can also improve the reliability of OxRAM devices, adopting for example bilayer solutions [174]. On the other hand, the computational efficiency of the developed model allows to simulate large memory arrays and take into account the synaptic variability corresponding to a wide range of programming conditions in neuromorphic systems simulations.

Thanks to the results obtained through device electrical characterization and simulation, we have proposed a novel design of an OxRAM-based synapse, offering multilevel capabilities using multiple binary HfO_2 devices connected in parallel. Using such OxRAM-based synapses, we have proposed for the first time a hardware implementation of a convolutional neural network where the convolution operation is performed directly in memory, overcoming the memory bottleneck of the Von Neumann implementations. CNNs are the state-of-the-art architectures for image recognition applications, used in commercial software implementations such as Facebook [121], for the recognition of people faces. A thorough analysis of both cycle-to-cycle and device-to-device variability of OxRAM synapses on system-level performance has been carried out. The impact of device variability on CNN performance has thus been studied, evaluating both analog and digital integration neuron models. Simulation results show that the proposed CNN architecture is highly tolerant to variability. Recognition success rates higher than 99% and 97% have been demonstrated for the handwritten digits and traffic sign recognition, respectively. These results, obtained simulating OxRAM synapses and taking into account the

device variability, are equivalent to those obtained in software implementation of CNN using floating point precision synaptic weights. Furthermore, the time required for the recognition of each image has been reduced, with an estimated latency reduced from 2.5 ms per image to 1 µs per image in the case of handwritten digits recognition. Such results confirm that OxRAM technology is a promising candidate for hardware implementation of spiking, resistive memory-based neuromorphic systems.

5.1 Future perspectives

In this work, we started from electrical results obtained from the characterization of memory bitcells and 16 kb array, and then moved to the evaluation of the proposed neuromorphic systems by simulation. The next step requires imperatively the fabrication of a hardware demonstrator, where CMOS neurons and RRAM synapses will be co-integrated.

For this purpose, the convolutional neural network architecture is a very promising candidate for a possible hardware implementation, due to its modular architecture and because it relies on the use of relatively small synaptic kernels to perform the operation of convolution. For example, a kernel with typical size $n = 5 \times 5$ would require an array of size 25×20 NVM devices, if 20 NVM parallel devices per synapse are adopted. This is definitely an attainable goal for current technology. Furthermore, due to the small size of synaptic array, problems due to parasitic capacitance and charging of long metal lines are easily avoided.

A first step towards the realization of a full network would thus be the realization of a single NVM kernel array, with its corresponding output neurons organized in a feature map. This kernel can be used to perform the convolution of one feature with a given input image at a time. Thanks to the use of AER representation encoding, it is a simple task to test a variety of input patterns, visual or auditory. Since the NVM devices are re-programmable many times, the extraction of multiple features can be tested, using the same synaptic array.

Once the design of a single synaptic kernel and the corresponding feature map has been validated, the next step would be the implementation of a whole convolutional layer, where multiple convolution operations are carried out in parallel using multiple kernels and feature maps. After completing the design of a convolutional layer, the next task is to cascade multiple convolutional layers one after the other, to achieve a complete feature extraction module. Finally, a classification module has to be implemented. This is the building block which is most demanding in terms of memory size, due to its fully connected topology. In the case of the network for the recognition of handwritten digits from the MNIST database, a memory array of 600 kb would be required. In the case of traffic sign recognition, we estimated an array size of 1 Mb. For more complex applications, a larger memory array might be necessary. The possibility of vertical integration of resistive memory devices [199],

[204] is a promising solution to increase synaptic density.

The use of binary devices for the implementation of artificial synapses has the advantage of being a very flexible approach, in the sense that it is not tightly bound to a specific memory technology, but it can be easily adapted to alternative options. In fact, the binary approach relies on non-volatile memory devices with only two programmable resistance levels. This feature is common to any emerging NVM technology: not only PCM and OxRAM, but also CBRAM and STT-MRAM. Neuromorphic systems designers can easily take advantage of the progresses of the research and development of emerging NVM devices for conventional memory applications, in terms of device reliability, low-power operation, uniformity, manufacturability, using simple programming schemes based on standard SET/RESET operations. The systems designer is thus relieved from the task of coming up with complex programming schemes aimed at obtaining multilevel programming, that need to be change every time a new technology is adopted.

An aspect that has to be taken into account is the fact that, in this work, we focused in this work on supervised learning, using backpropagation algorithm. In addition to this, it is of primary importance a deeper investigation of unsupervised learning with CNNs, which is particularly useful when a training data set is not available.

In conclusion, the fabrication of a fully integrated, emerging NVM-based neuromorphic hardware demonstrator seems, to the author of this thesis, a task that can be accomplished in the near future.

Appendix A

The Xnet simulator

In this thesis, neuromorphic system simulations have been performed using Xnet, an event-driven simulator for spiking neuromorphic architectures developed by Bichler *et al.* at CEA-LIST [165], [166]. Xnet is currently being used in the framework of the collaboration between LETI and LIST for the study of the use of emerging memory devices as artificial synapses in neuromorphic systems. The simulator is a special-purpose software that has been designed to provide an intermediate modeling layer for neuromorphic hardware, closing the gap between hardware description languages, such as VHDL [205], Verilog [206] or SystemC [207], and neural network simulators used by the neuroscience community, such as Neuron [208], Brian [209] or NEST [210]. Xnet is characterized by a high computational efficiency, thanks to the use C++ programming language standard libraries. This allows for fast and efficient architectural exploration, in terms of network topology, neuron parameters and learning rule parameters. At the same time, natively takes into account experimental variability and stochasticity in both synaptic and neural models, allowing for a feedback exchange between technology and architecture levels.

Figure A.1 indicates the typical Xnet simulation flow. Input data are first converted into spiking activity that can be fed to the spiking neural network. In the case of video streams in Address Event Representation (AER) format, the input data can be used directly as the input of the system. This is the case, for example, of the car video presented in this thesis in Chapter 2. This direct compatibility is particularly useful because AER format is a standard asynchronous communication protocol, used in many bio-inpired sensors such as artificial retinas, and widely used in the neuromorphic community [106], [211], [212]. In the case of static input images, as in the case of handwritten digits (MNIST) or traffic signs (GTSRB) images, the pre-porcessing phase consists in scaling, filtering and conversion of each pixel brightness into spike encoding for the input neurons, with multiple available conversion algorithms. In the case of auditive data, the audio stream is fed into a band-pass filter bank, and each filter is associated to an input neuron.

After the pre-processing phase, spikes are elaborated by the event processing

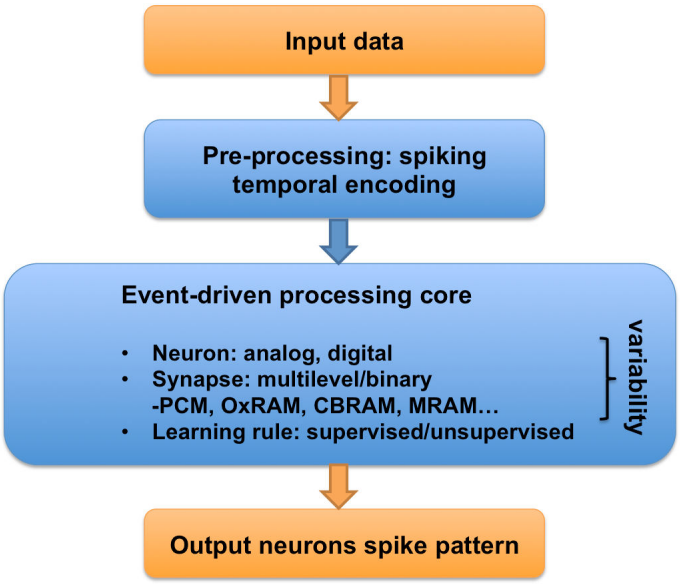


Figure A.1. Diagram of Xnet simulation flow.

engine, which is organized around an event queue, implemented with a priority queue of the C++ standard library. Such engine comprises functional models of LIF neurons, with both analog and digital integration implementations. For the synapses, the model is based on real device characterization data, with the possibility of simulating both binary and multilevel devices, with tunable synaptic redundancy. Variability characteristics are extrapolated with statistic methods from electrical characterization data. Neurons and synapses can be arranged in multiple neural network topologies. In this thesis, we focused on multilayer fully connnected neural networks and convolutional neural networks. The learning rule associated to the simulated neuromorphic system can be both supervised and unsupervised (STDP).

In order to obtain the output of the network, the output neuron layer spiking activity is monitored. It is thus compared to a reference spiking activity (expected output) in order to compute the performance of the network.

In conclusion, Xnet is a powerful tool that allows for computationally efficient simulations of neural networks, taking into account electrical characteristics of real nanodevices.

Appendix B

Author's publications

Journals

[1] D. Garbin, E. Vianello, O. Bichler, Q. Rafhay, C. Gamrat, G. Ghibaudo, B. DeSalvo, and L. Perniola, "HfO_2-based OxRAM devices as synapses for convolutional neural networks", *Electron Devices, IEEE Transactions on*, vol. 6, no. 8, pp. 2494–2501, 2015. DOI: 10.1109/TED.2015.2440102.

[2] D. Garbin, E. Vianello, Q. Rafhay, M. Azzaz, S. Jeannot, P. Candelier, B. DeSalvo, G. Ghibaudo, and L. Perniola, "Resistive memory variability: a simplified trap-assisted tunneling model", *Solid-State Electronics*, vol. 115, pp. 126–132, 2016. DOI: 10.1016/j.sse.2015.09.004.

[3] S. Raoux, A. K. König, H.-Y. Cheng, D. Garbin, R. W. Cheek, J. L. Jordan-Sweet, and M. Wuttig, "Phase transitions in Ga–Sb phase change alloys", *Physica status solidi (b)*, vol. 249, no. 10, pp. 1999–2004, 2012. DOI: 10.1002/pssb.201200370.

International conferences

[1] D. Garbin, O. Bichler, E. Vianello, Q. Rafhay, C. Gamrat, L. Perniola, G. Ghibaudo, and B. DeSalvo, "Variability-tolerant convolutional neural network for pattern recognition applications based on OxRAM synapses", in *Electron Devices Meeting (IEDM), 2014 IEEE International*, IEEE, 2014, pp. 28–4.

[2] D. Garbin, E. Vianello, O. Bichler, M. Azzaz, Q. Rafhay, P. Candelier, C. Gamrat, G. Ghibaudo, B. DeSalvo, and L. Perniola, "On the impact of OxRAM-based synapses variability on convolutional neural networks performance", in *Symposium on Nanoscale Architecture (NANOARCH), 2015 IEEE/ACM International*, IEEE/ACM, 2015, pp. 193–198.

[3] D. Garbin, Q. Rafhay, E. Vianello, S. Jeannot, P. Candelier, B. DeSalvo, G. Ghibaudo, and L. Perniola, "Modeling of OxRAM variability from low to high resistance state using a stochastic trap assisted tunneling-based resistor network", *2015 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, pp. 125–128, 2015.

[4] D. Garbin, M. Suri, O. Bichler, D. Querlioz, C. Gamrat, and B. DeSalvo, "Probabilistic neuromorphic system using binary phase-change memory (pcm) synapses: detailed power consumption analysis", in *Nanotechnology (IEEE-NANO), 2013 13th IEEE Conference on*, IEEE, 2013, pp. 91–94.

[5] B. DeSalvo, E. Vianello, D. Garbin, O. Bichler, and L. Perniola, "From memory in our brain to emerging resistive memories in neuromorphic systems", in *Memory Workshop (IMW), 2015 IEEE International*, May 2015, pp. 1–4. DOI: 10.1109/IMW.2015.7150286.

[6] G. Piccolboni, G. Molas, J. Portal, R. Coquand, M. Bocquet, D. Garbin, E. Vianello, C. Carabasse, V. Delaye, C. Pellissier, T. Magis, C. Cagli, M. Gely, O. Cueto, D. Deletuyelle, G. Ghibaudo, B. De Salvo, and L. Perniola, "Investigation of the potentialities of vertical resistive RAM (VRAM) for neuromorphic applications", in *Electron Devices Meeting (IEDM), 2015 IEEE International*, IEEE, 2015, pp. 447–450.

[7] E. Vianello, D. Garbin, N. Jovanovic, O. Bichler, O. Thomas, B. de Salvo, and L. Perniola, "(keynote) oxide based resistive memories for low power embedded applications and neuromorphic systems", 3, vol. 69, The Electrochemical Society, 2015, pp. 3–10.

[8] M. Azzaz, A. Benoist, E. Vianello, D. Garbin, E. Jalaguier, C. Cagli, C. Charpin, S. Bernasconi, S. Jeannot, T. Dewolf, G. Audoit, C. Guedj, S. Denorme, P. Candelier, C. Fenouillet-Beranger, and L. Perniola, "Benefit of Al2O3/HfO2 bilayer for BEOL RRAM integration through 16 kb memory cut characterization", in *Solid State Device Research Conference (ESSDERC), 2015 45th European*, IEEE, 2015, pp. 266–269.

[9] E. Vianello, O. Thomas, G. Molas, O. Turkyilmaz, N. Jovanović, D. Garbin, G. Palma, M. Alayan, C. Nguyen, J. Coignus, *et al.*, "Resistive memories for ultra-low-power embedded computing design", in *Electron Devices Meeting (IEDM), 2014 IEEE International*, IEEE, 2014, pp. 6–3.

[10] M. Mongillo, D. Garbin, G. Navarro, E. Vianello, M. Coue, B. Mayall, and D. Cooper, "In-situ biasing and switching of electronic devices into a TEM", in *18th International Microscopy Congress (IMC 2014)*, 2014, IT–7.

[11] M. Suri, D. Garbin, O. Bichler, D. Querlioz, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Impact of pcm resistance-drift in neuromorphic systems and drift-mitigation strategy", in *Nanoscale Architectures (NANOARCH), 2013 IEEE/ACM International Symposium on*, IEEE, 2013, pp. 140–145.

[12] S. Raoux, H.-Y. Cheng, D. Garbin, R. Cheek, A. Koenig, and M. Wuttig, "Fast turn-around materials characterization for phase change memory application using a static laser tester", in *2013 MRS Spring Meeting & Exhibit. Symposium EE: Phase-Change Materials for Memory, Reconfigurable Electronics, and Cognitive Applications*, 2013, EE1–04.

[13] S. Raoux, H.-Y. Cheng, J. Jordan-Sweet, T. Monin, F. Xiong, A. König, D. Garbin, R. Cheek, E. Pop, and M. Wuttig, "Crystallization properties of ga-sb phase change alloys", in *European Phase Change and Ovonic Science Symposium (E\PCOS), Berlin, Germany, September 2013*, 2013.

[14] H. Cheng, J. Wu, R. Cheek, S. Raoux, M. BrightSky, D. Garbin, S. Kim, T. Hsu, Y. Zhu, E. Lai, *et al.*, "A thermally robust phase change memory by engineering the Ge/N concentration in $(Ge, N)_xSb_yTe_z$ phase change material", in *Electron Devices Meeting (IEDM), 2012 IEEE International*, IEEE, 2012, pp. 31–1.

[15] S. Raoux, D. Garbin, C.-I. Wu, H.-Y. Cheng, R. Cheek, A. König, M. Wuttig, M. J. BrightSky, H.-L. Lung, and C. H. Lam, "Comparison of data retention measured by static laser testing and in pcram devices", in *European Phase Change and Ovonic Science Symposium (E\PCOS), Tampere, Finland, September 2012*, 2012.

Patents

[1] E. Vianello and D. Garbin, "Method of programming a resistive random access memory", U.S. application no. 14/956,838, Dec. 2, 2015.

Appendix C

Résumé en français

Introduction

Le cerveau humain est composé d'un grand nombre de réseaux interconnectés, dont les neurones et les synapses en sont les briques constitutives. Caractérisé par une faible consommation de puissance, de quelques Watts seulement, le cerveau humain est capable d'accomplir des tâches qui sont inaccessibles aux systèmes de calcul actuels, basés sur une architecture de type Von Neumann. La conception de systèmes neuromorphiques vise à réaliser une nouvelle génération de systèmes de calcul qui ne soit pas de type Von Neumann. L'utilisation de mémoires non-volatile innovantes en tant que synapses artificielles, pour application aux systèmes neuromorphiques, est donc étudiée dans cette thèse. Deux types de technologies de mémoires sont examinés : les mémoires à changement de phase (Phase-Change Memory, PCM) et les mémoires résistives à base d'oxyde (Oxide-based resistive Random Access Memory, OxRAM).

C.1 Mémoires non-volatiles émergentes et systèmes neuromorphiques

Dans le chapitre 1, nous introduisons le contexte et la motivation derrière la recherche menée au cours de la préparation de cette thèse. D'une part, les dispositifs mémoire non volatile (NVM) émergents sont étudiés ayant à l'esprit le rôle central qu'ils vont jouer dans les architectures de mémoire du future. D'autre part, une nouvelle application de dispositifs NVM, qui a attiré un grand intérêt au cours des dernières années, est étudiée : la réalisation de synapses artificiels dans les architectures de calcul inspirées du cerveau humain. Compte tenu de l'interdisciplinarité necessaire dans ce projet, ce chapitre décrit en profondeur les concepts de bases pour contextualiser cette recherche.

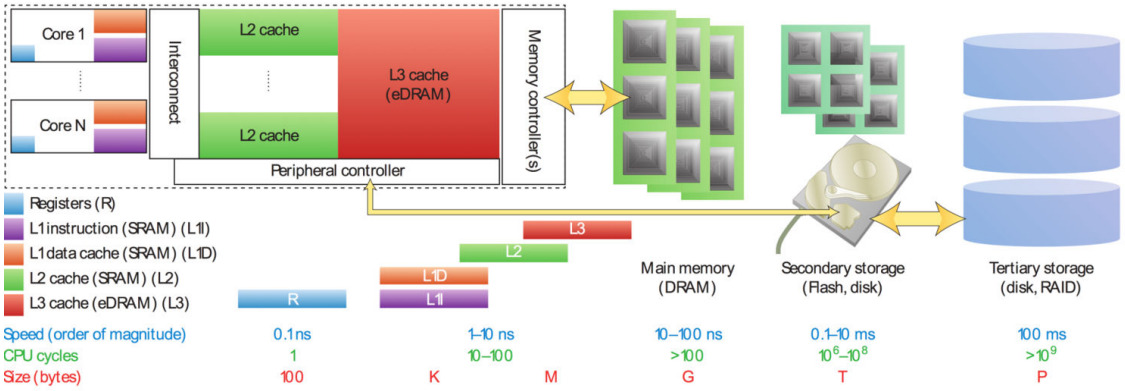


FIGURE C.1. La hiérarchie de la mémoire dans les ordinateurs. Une petite quantité de mémoire volatile à haute performance et coûteuse est à proximité du CPU. De grandes quantités de mémoires non volatiles plus lentes et des unités de stockage à faible coût sont loin du CPU, en bas de la hiérarchie. Source : [4].

C.1.1 Technologies de mémoire non volatile émergentes

La conception des systèmes informatiques est actuellement basée sur l'architecture de Von Neumann [1]. Dans cette architecture, une distinction marquée existe entre le rôle de l'unité centrale de traitement (*Central Processing Unit*, CPU) et l'unité de mémoire (*Memory Unit*, MU). Il existe un écart entre le processeur et la mémoire en termes de performances : les performances de calcul sont généralement limitées par la vitesse avec laquelle les données peuvent être récupérées dans la mémoire. Le temps de latence et la bande passante sont donc les principaux facteurs de limitation des performances [2].

La mémoire est généralement structurée comme une hiérarchie de dispositifs de mémoires volatile et non volatile, afin de parvenir à un compromis entre coût et performances optimales. Le but de cette hiérarchie, comme montré sur la Fig. C.1, est de combler l'écart de performance entre le processeur qui est rapide et les technologies de mémoire et de stockage plus lentes, tant en gardant le coût du système le plus faible possible.

Depuis l'apparition et la forte croissance des appareils portables tels que les lecteurs de musique et les téléphones cellulaires, la mémoire flash s'est imposée dans la hiérarchie de stockage de l'information, entre la RAM et le disque dur, comme solution de stockage non volatile. Comme le montre la figure C.2, la croissance de la technologie Flash a explosé au cours des dernières années, et elle est devenue la technologie de stockage de données dominante pour les applications mobiles.

Cependant, la technologie Flash est confrontée à de nombreux défis pour le développement des futurs noeuds technologiques, en raison de limitations physiques intrinsèques. Pour cette raison, des efforts de recherche sont actuellement en cours afin de trouver de nouvelles technologies de mémoire non volatile, avec une meilleure

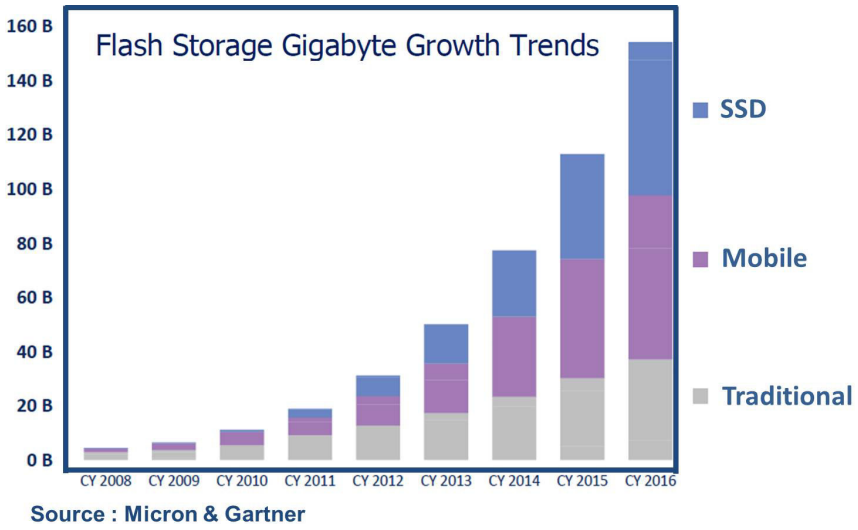


FIGURE C.2. Tendances de croissance du gigaoctet de stockage Flash, source : [12].

capacité d'évolution par rapport à la Flash.

Dans la recherche de solutions innovantes de mémoire non-volatile, différentes technologies sont ainsi apparues au cours des 15 dernières années [19], [20].

Les principales technologies de mémoire non volatile émergentes sont les suivantes :

- Les mémoires à changement de phase (*Phase-Change Random Access Memory*, PCRAM ou PCM) ;
- Les mémoires magnetiques (*Spin-Transfer-Torque Magnetic Random Access Memory*, STT-MRAM) ;
- Les mémoires à pont conducteur (*Conductive-Bridging Random Access Memory*, CBRAM) ;
- Les mémoires à oxyde métallique (*metal Oxide resistive Random Access Memory*, OxRAM).

Le tableau C.1 présente les performances actuellement atteintes par ces technologies.

C.1.2 Systèmes neuromorphiques

En plus d'un changement radical dans l'organisation de la hiérarchie de la mémoire dans les architectures de calcul de type Von Neumann, les mémoires non-volatiles innovantes ont été identifiées comme des acteurs clés dans un possible changement du paradigme de calcul, au-delà de l'architecture traditionnelle de Von Neumann,

		PCRAM	STT-MRAM	CBRAM	OxRAM
Feature Size F (nm)	Demonstrated	45	65	20	5
	Projected	8	16	5	<5
Cell Area	Demonstrated	$4F^2$	$20F^2$	$4F^2$	$4F^2$
	Projected	$4F^2$	$8F^2$	$4F^2$	$4F^2$
Programming Voltage (V)	Demonstrated	3	1.8	0.6	1
	Projected	<3	<1	<0.5	<1
Programming Time (ns)	Demonstrated	100	35	<1	<1
	Projected	<50	<1	<1	<1
Programming Energy (J/bit)	Demonstrated	$6 \cdot 10^{-12}$	$2.5 \cdot 10^{-12}$	$8 \cdot 10^{-12}$	$< 1 \cdot 10^{-12}$
	Projected	$1 \cdot 10^{-15}$	$1.5 \cdot 10^{-13}$	N.A.	$1 \cdot 10^{-16}$
Read Voltage (V)	Demonstrated	1.2	1.8	0.2	0.1
	Projected	<1	<1	<0.2	0.1
Retention Time	Demonstrated	>10yr	>10yr	>10yr	>10yr
	Projected	>10yr	>10yr	>10yr	>10yr
Endurance (nb. cycles)	Demonstrated	10^9	$> 10^{12}$	10^{10}	10^{12}
	Projected	10^9	$> 10^{15}$	$> 10^{11}$	$> 10^{12}$

TABLE C.1. Comparaison des performances des différentes technologies de mémoire non volatile émergents selon la International Technology Roadmap for Semiconductor (ITRS) 2013 [63], avec des projections pour l'année 2026.

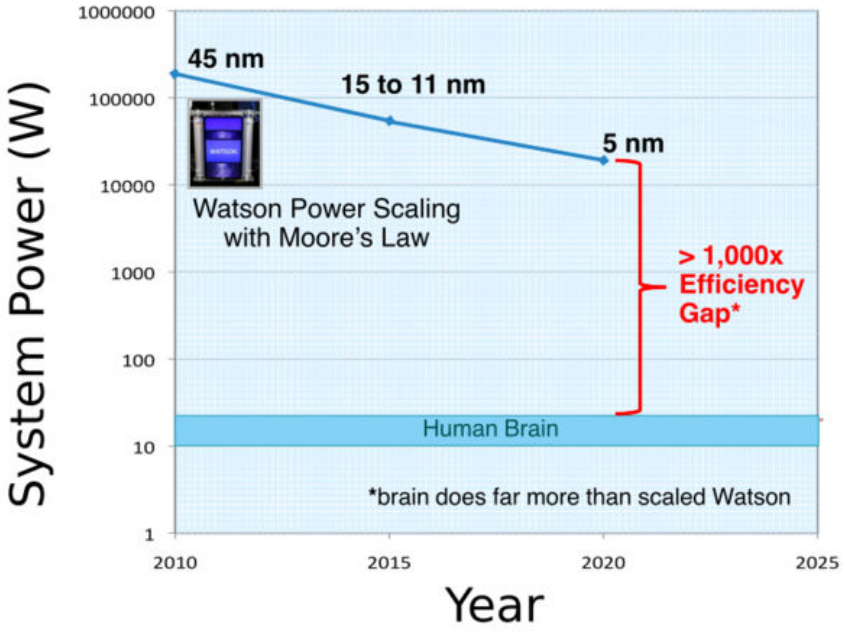


FIGURE C.3. Comparaison entre la consommation de puissance d'un superordinateur Watson d'IBM et le cerveau humain [68].

grâce à leur utilisation comme synapses artificiels dans les systèmes neuromorphiques [64].

Comme le montre la figure C.3, la consommation de puissance associée à l'architecture de Von Neumann est de plusieurs ordres de grandeurs supérieure par rapport à la puissance requise par le cerveau humain. L'invention de nouvelles architectures est donc nécessaire pour combler l'écart de l'efficacité qui existe entre les architectures de calcul classiques et le cerveau humain. Dans la recherche de solutions de calcul plus efficaces, les systèmes neuromorphiques ont ainsi été proposées comme une nouvelle génération de systèmes de calcul, avec un rôle complémentaire par rapport aux machines de Von Neumann [68].

On estime que dans le cerveau humain, il y a environ 10^{11} neurones, et 10^{15} synapses (Fig. C.4). En raison du nombre de synapses (environ 4 ordres de grandeur plus grand que le nombre de neurones) le défi est donc de trouver une conception efficace pour la synapse, afin d'être en mesure d'intégrer les réseaux de neurones à grande échelle sur une puce. Cette implémentation matérielle de synapses artificiels est discuté dans les chapitres 2 et 4 de cette thèse.

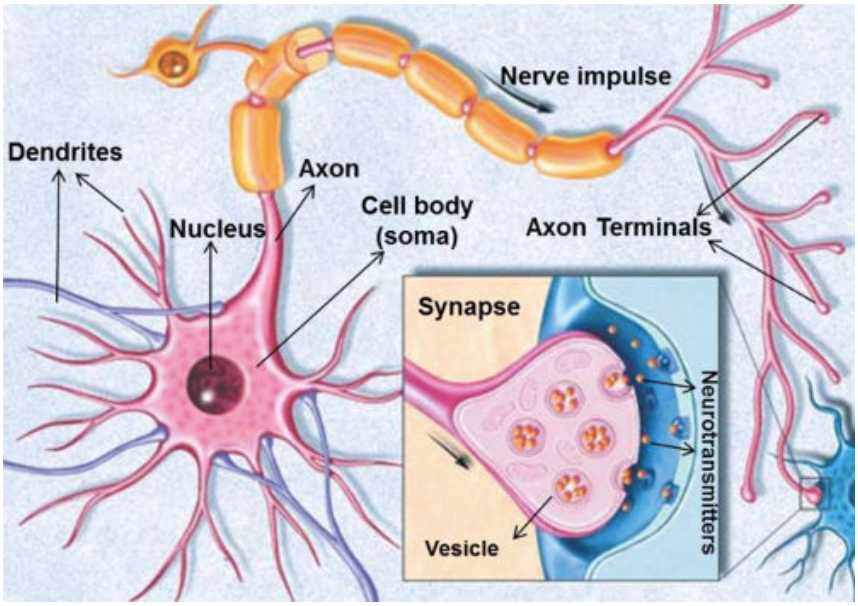


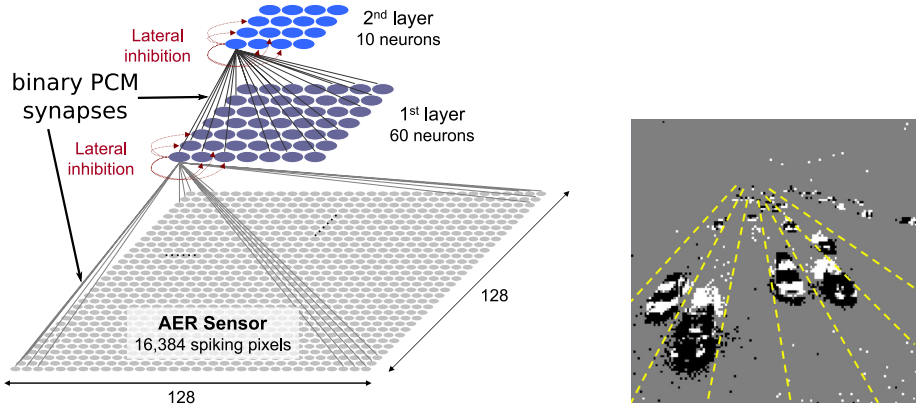
FIGURE C.4. Vue schématique de la structure de base d'une cellule neuronale. L'encart montre un zoom de la synapse biologique. Source : [85].

C.2 Systèmes neuromorphiques basés sur des synapses de type PCRAM

Dans le chapitre 2, nous avons tout d'abord étudié l'utilisation de dispositifs de type PCM en tant que synapses dans un réseau neuronal artificiel entièrement connecté (*fully connected*, Fig. C.5a). En ce qui concerne la réalisation de synapses artificielles, deux types d'approche existent [105]. Le prémier, dit multniveau, utilise plusieurs niveaux de résistance par dispositif synaptique. Le deuxieme approche, dit binaire, consiste à utiliser des dispositifs synaptiques avec deux états de résistance seulement.

Nous avons présenté les limitations associées à l'utilisation de l'approche multiniveaux. Par conséquent, afin de surmonter les limitations associées à l'approche multiniveaux, nous avons exploré l'utilisation des synapses PCM en mode binaire. Sur la base des résultats obtenus à partir des caractérisations électriques, nous avons effectué des simulations d'un réseau neuronal artificiel à grande échelle pour une application visuelle complexe (Fig. C.5b). Les niveaux des états de résistance SET et RESET ont donc été réglés en simulation en fonction des conditions de programmation expérimentales sélectionnées.

Deux systèmes de programmation, pour des architectures avec ou sans dispositif sélecteur, sont prévus (Figs. C.6a et C.6b). Les systèmes de programmation proposées évitent ainsi l'utilisation de systèmes de rafraîchissement complexes requis par des synapses PCM multiniveaux.



(a) (b)

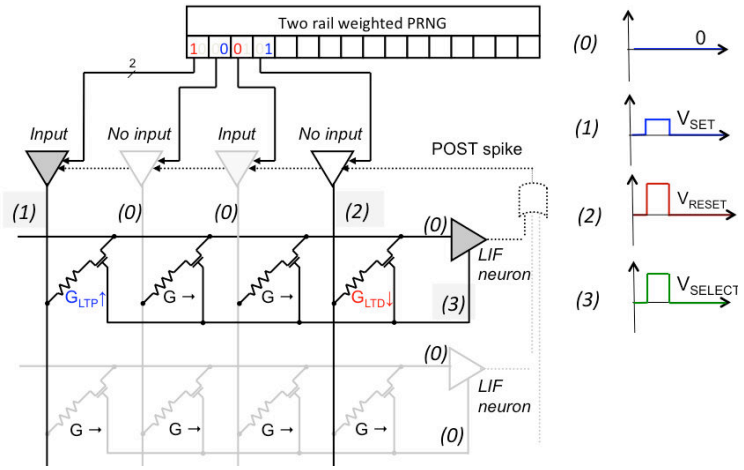
FIGURE C.5. (a) Schématique du système neuromorphic entièrement connecté étudié en simulation. (b) Un exemple de l'un des images de l'entrée vidéo, montrant des voitures qui passent sur plusieurs voies d'une autoroute. La séparation entre les voies (en jaune) a été ajoutée pour illustrer la distinction entre les différentes voies et n'est pas présente dans la vidéo d'entrée d'origine.

Les résultats de simulation montrent de plus que la consommation de puissance en mode d'apprentissage associée au système neuromorphique étudié peut être considérablement réduit si l'état RESET des dispositifs PCM est fixé à résistance relativement faible. La consommation d'énergie en mode de lecture, d'autre part, peut être minimisée par l'augmentation des valeurs de résistance des états SET et RESET des dispositifs PCM. Nous avons également étudié la question de la dérive de la résistance de PCM dans le temps, et nous avons proposé une stratégie pour atténuer ce problème. Nous avons observé qu'en utilisant des dispositifs de taille réduite (Figs. C.7a et C.7b), il est possible de réduire considérablement la consommation d'énergie grâce à un courant de programmation plus petit (Tableau C.2).

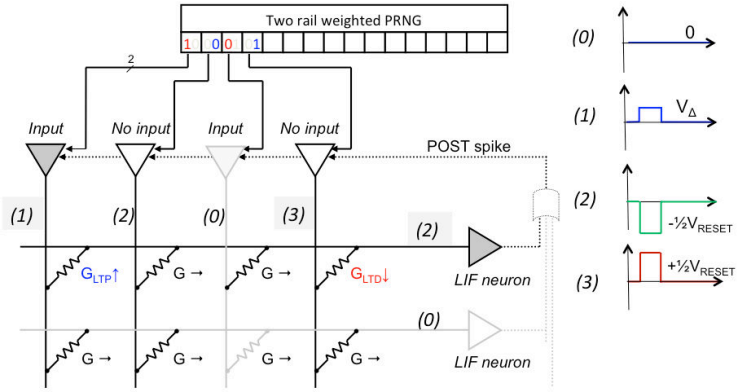
En conclusion, nous avons démontré avec succès l'intérêt de l'utilisation de dispositifs de PCM en mode binaire pour la réalisation d'un système neuromorphique dedié à des applications visuelles complexes.

C.3 Technologie OxRAM : mécanismes de défauts et variabilité

Dans le chapitre 3, les principales caractéristiques de la technologie OxRAM à base de HfO_2, l'une des technologies de mémoire non volatile émergentes les plus prometteuses, ont été présentées. Les dispositifs étudiés dans ce chapitre sont des structures Métal-Isolant-Métal (MIM) composées d'une couche HfO_2 entre une électrode supérieure TiN/Ti et une électrode inférieure de TiN (Fig. C.8). Une cellule



(a) Architecture avec dispositif sélecteur

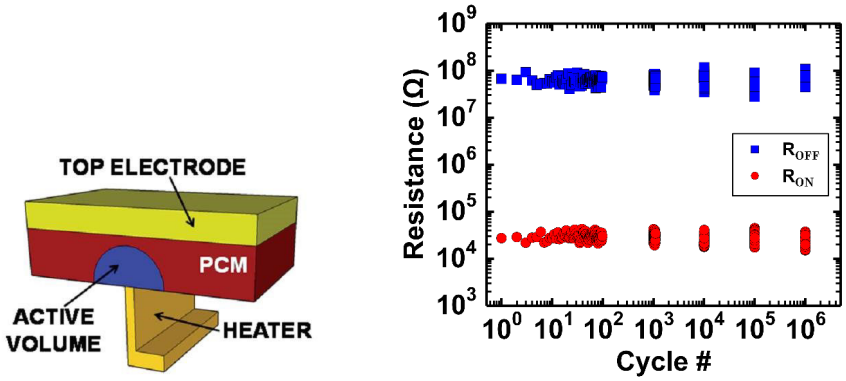


(b) Architecture sans dispositif sélecteur

FIGURE C.6. (a) Architecture avec dispositif sélecteur et (b) sans dispositif sélecteur.

mémoire est composée d'une structure 1 Transistor − 1 Résistance (1T1R), où le transistor d'accès est utilisé pour sélectionner la cellule lorsqu'elle est intégrée dans une matrice et pour limiter le courant circulant à travers le dispositif pendant la programmation. La caractérisation électrique a été effectué sur des cellules 1T1R individuelles et sur une puce CMOS 28 nm de test numérique qui contient 16 circuits sous test (CUTs) de 1 kb de mémoire OxRAM chacun, plus un contrôleur numérique (Fig. C.9a,b), fabriquée dans le cadre d'un projet de R&D entre STMicroelectronics et le CEA-LETI [172]–[174]. Les caractéristiques I−V typiques et le comportement de commutation d'un cellule d'information 1T1R sont présentés dans la figure C.10.

Les résultats expérimentaux de l'étude des mécanismes de défaillance et d'endurance ont été discutées et une méthodologie de programmation pour améliorer l'endurance à faible courant de fonctionnement a été proposée. Une opération de « formation » appropriée a été suggérée dans cette thèse (Fig. C.11). Il prévoit un



(a) (b)

FIGURE C.7. (a) Représentation schématique de la structure de dispositif PCM à taille réduite [25]. (b) Expérience pour 10^6 cycles SET-RESET pour un dispositif PCM à taille réduite.

Quantity	Large heater synapses	Wall storage synapses
p_{LTP}	0.14	0.14
p_{LTD}	0.11	0.21
t_{LTP}	7.6 ms	13.4 ms
Nb. SET pulses	$4.5 \cdot 10^5$	$8.9 \cdot 10^5$
Nb. RESET pulses	$1.6 \cdot 10^7$	$4.7 \cdot 10^7$
Nb. Read pulses	$2.48 \cdot 10^9$	$2.48 \cdot 10^9$
Energy associated to SET events	0.4 mJ	0.2 mJ
Energy associated to RESET events	47.3 mJ	4.3 mJ
Total energy (SET + RESET)	47.7 mJ	4.5 mJ
Total power (SET + RESET)	70.1 µW	6.6 µW
Read energy	43 µJ	0.3 µJ
Read power	64 nW	0.5 nW

TABLE C.2. Comparaison des statistiques d'apprentissage PCM obtenues pour les dispositifs de grande et petite taille.

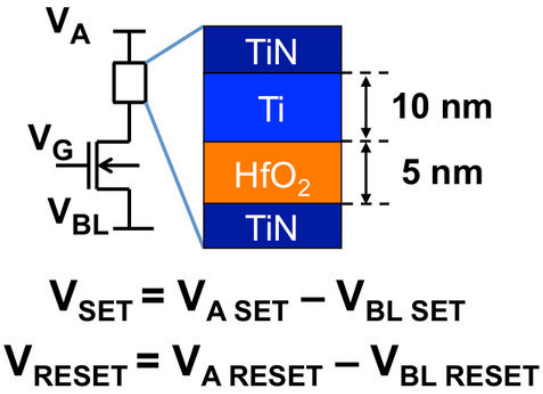


FIGURE C.8. Schéma du dispositif 1T-R.

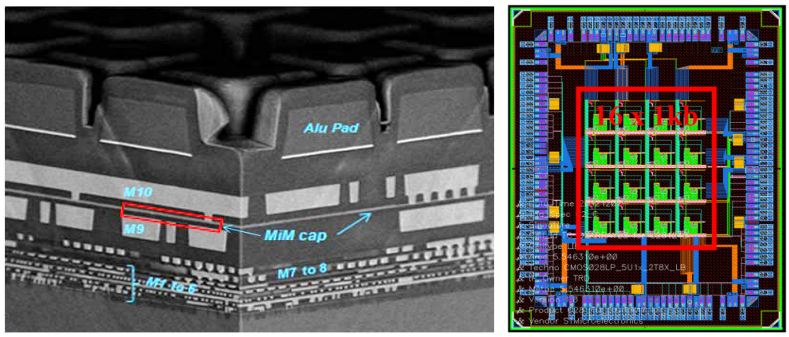


FIGURE C.9. (a) Coupe SEM du stack CMOS 28 nm, y compris le dispositif MIM, (b) layout du circuit demonstrateur 16 kb [172].

courant de formation ($I_{C\ FORMING}$) supérieur au courant utilisé lors des opérations de SET ultérieurs ($I_{C\ SET}$), afin d'augmenter la taille du réservoir de lacunes d'oxygène générées pendant l'opération de formation. Une valeur optimale expérimentale de $I_{C\ FORMING}/I_{C\ SET}$ d'environ $2,7$ est observée.

Un modèle physique capable d'expliquer à la fois la variabilité des OxRAM à l'état faiblement résistif (*low resistance state*, LRS) et à l'état hautement résistif (*high resistance state*, HRS) a été présenté. La figure C.12 montre un bon accord entre les résultats expérimentaux et les simulations.

L'étude de variabilité a été réalisée avec un double objectif. Du point de vue des applications de mémoire classique, la variabilité est en effet l'un des facteurs limitants l'adoption de la technologie OxRAM dans des produits commerciaux. La compréhension de la source des variations de résistance des OxRAM peut donc donner des lignes directrices pour résoudre ce problème. Du point de vue de l'informatique neuromorphique, les dispositifs OxRAM sont des candidats idéaux pour la réalisation des synapses artificielles. Le développement d'un modèle capable de reproduire la

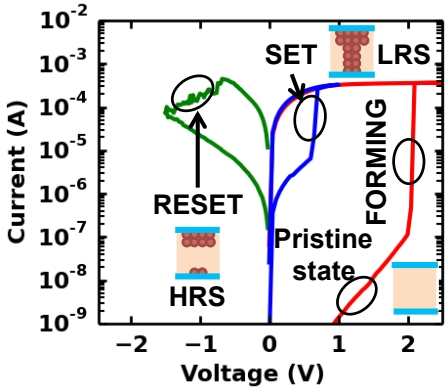


FIGURE C.10. Caractéristiques de courant-tension OxRAM typiques. Les opérations de FORMING, SET et RESET sont mises en évidence.

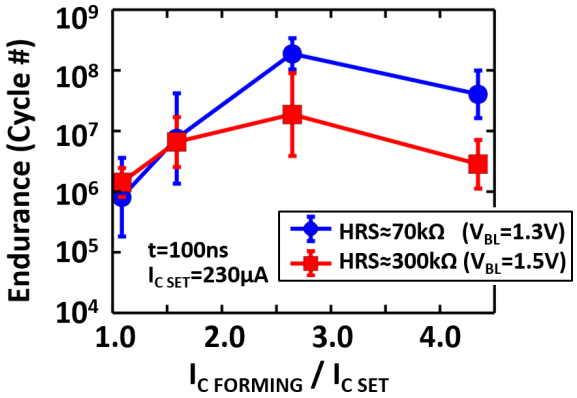


FIGURE C.11. Endurance en fonction du ratio $I_{C\ FORMING}/I_{C\ SET}$, obtenu pour différentes valeurs de V_{RESET} avec temps de programmation $t = 100\,ns$. Chaque point correspond à la valeur moyenne sur environ 4 cellules. Une valeur optimale expérimentale du rapport est observée $I_{C\ FORMING}/I_{C\ SET} \approx 2,7$.

variabilité de l'appareil pour une large gamme de conditions de programmation peut être utilisé pour étudier l'impact de la variabilité synaptique au niveau du système.

C.4 Dispositifs OxRAM en tant que synapses pour des réseaux de neurones convolutifs

Dans le chapitre 4, nous avons présenté une nouvelle conception de synapse artificielle à base de dispositifs OxRAM, offrant des capacités à plusieurs niveaux en utilisant de multiples dispositifs binaires connectés en parallèle (Fig. C.13). La caractérisation électrique, la modélisation physique et les simulations suggèrent que la technologie OxRAM est un bon candidat pour la réalisation de synapses artificielles dans les

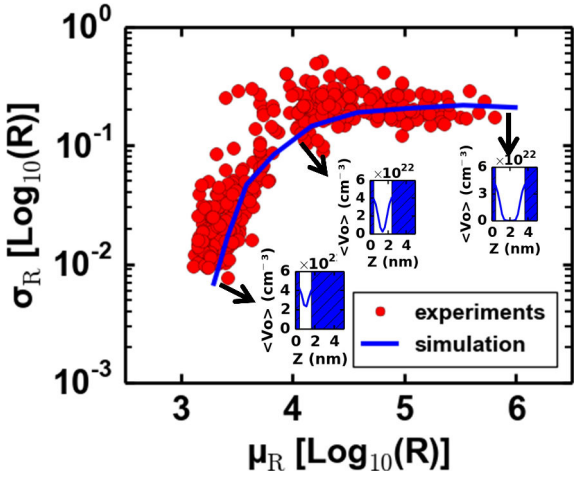


FIGURE C.12. Écart-type de la résistance expérimental et obtenu en simulation en fonction de la résistance moyenne.

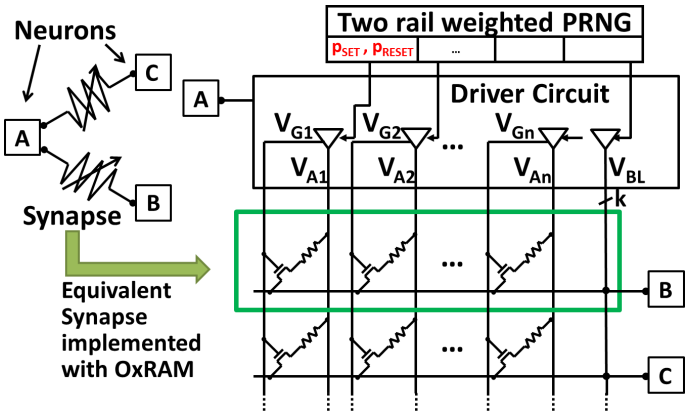
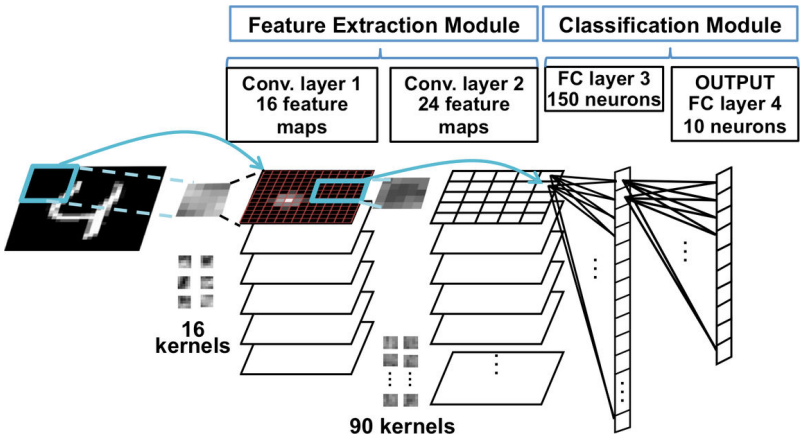


FIGURE C.13. Schéma de synapses à base de OxRAM utilisés pour l'opération de convolution dans l'architecture CNN. Tous les dispositifs de OxRAM sur la même ligne construisent une synapse équivalente. Le circuit de pilotage est utilisé pour programmer individuellement les dispositifs OxRAM et propager les impulsions entre les couches de neurones. Le circuit *Pseudo-Random Number Generator* PRNG est utilisé pour l'apprentissage en ligne, pour obtenir la stochasticité extrinsèque dans une règle d'apprentissage probabiliste *Spike-Timing Dependent Plasticity*, STDP.

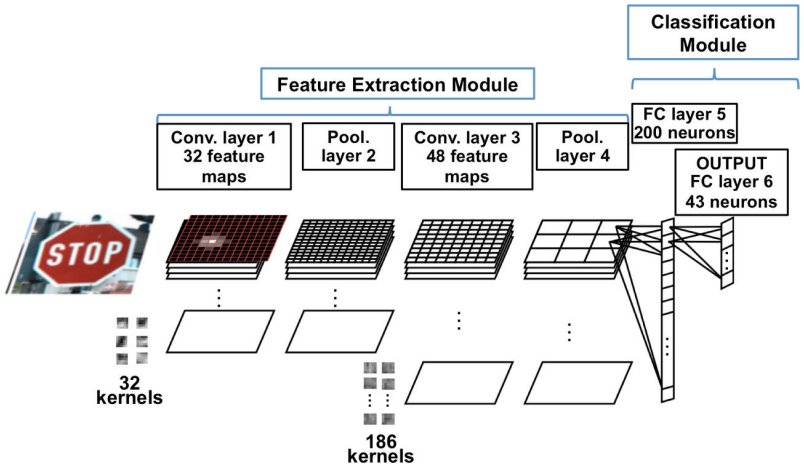
systèmes neuromorphiques.

En utilisant les synapses OxRAM proposées, nous avons présenté pour la première fois une réalisation d'un réseau neuronal convolutif (*Convolutional Neural Network*, CNN) où l'opération de convolution est effectuée directement dans la mémoire (Fig. C.14a et C.14b).

Une analyse approfondie de la variabilité cycle à cycle et dispositif à dispositif des synapses OxRAM, extraite d'une matrice OxRAM, a été réalisée. L'impact de la



(a) MNIST



(b) GTSRB

FIGURE C.14. Architecture CNN pour (a) reconnaissance de chiffres manuscrits (MNIST database) et (b) reconnaissance de panneaux de signalisation routière (GTSRB database).

variabilité des dispositifs sur la performance du réseau convolutif a été étudié. Les résultats montrent que l'architecture proposée CNN est très tolérante aux variations, sans qu'il soit nécessaire d'utiliser un algorithme de correction.

Des taux de reconnaissance supérieurs à 99% et 97% ont été respectivement démontrés pour les réseaux de reconnaissance de chiffres manuscrits et de panneaux de signalisation routière. Le taux sont proches de l'état de l'art des taux de reconnaissance obtenus avec des modèles formels de CNN, réalisés en utilisant des synapses avec précision à virgule flottante. En outre, l'architecture proposée permet de réduire le temps nécessaire à la reconnaissance de chaque image par rapport à une architecture de type Von Neumann, compte tenu d'une fréquence de fonctionnement similaire.

C.5 Conclusions

Dans cette thèse, nous avons exploré l'utilisation de dispositifs PCM et OxRAM en tant que synapses artificiels pour les systèmes neuromorphiques. Les résultats obtenus confirment que les deux technologies sont des candidats prometteurs pour la réalisation des systèmes neuromorphiques à base de mémoires résistives, tant en termes d'efficacité énergétique que de bonnes performancese. En perspective, la prochaine étape nécessitera la fabrication d'un démonstrateur matériel, où les neurones CMOS et les synapses RRAM seront co-intégrés.

Bibliography

[1] J. Von Neumann, "First draft of a report on the EDVAC", *IEEE Annals of the History of Computing*, no. 4, pp. 27–75, 1993.

[2] A. Macii, L. Benini, and M. Poncino, *Memory Design Techniques for Low Energy Embedded Systems*. Springer Science & Business Media, 2002.

[3] G. W. Burr, M. J. Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson, B. Kurdi, C. Lam, L. A. Lastras, A. Padilla, *et al.*, "Phase change memory technology", *Journal of Vacuum Science & Technology B*, vol. 28, no. 2, pp. 223–262, 2010.

[4] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing", *Nature nanotechnology*, vol. 10, no. 3, pp. 191–194, 2015.

[5] K. Itoh, *VLSI memory chip design*. Springer Science & Business Media, 2013.

[6] S. Borkar and A. A. Chien, "The future of microprocessors", *Communications of the ACM*, vol. 54, no. 5, pp. 67–77, 2011.

[7] IBM Corporation. (2006). IBM details next generation of storage innovation, [Online]. Available: `http://www-03.ibm.com/press/us/en/pressrelease/20209.wss` (visited on Aug. 7, 2015).

[8] J. Brewer and M. Gill, *Nonvolatile Memory Technologies with Emphasis on Flash: A Comprehensive Guide to Understanding and Using Flash Memory Devices*. John Wiley & Sons, 2011, vol. 8.

[9] R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti, "Introduction to flash memory", *Proceedings of the IEEE*, vol. 91, no. 4, pp. 489–502, 2003.

[10] HGST, a Western Digital Company, "Solid state drives for enterprise data center environments", Oct. 2012. [Online]. Available: `http://www.hgst.com/sites/default/files/resources/SSD_techbrief.pdf`.

[11] Micron Technology, Inc. and Intel Corporation. (2015). Micron and Intel unveil new 3D NAND Flash memory, [Online]. Available: `http://investors.micron.com/releasedetail.cfm?ReleaseID=903522` (visited on Aug. 31, 2015).

[12] G. Hawk, "Now is the time for flash storage", in *Flash Memory Summit 2012 Proceedings*, 2012.

[13] K. Kim and S. Lee, "Memory technology in the future", *Microelectronic Engineering*, vol. 84, no. 9, pp. 1976–1981, 2007.

[14] K. Prall, "Scaling non-volatile memory below 30nm", in *2007 22nd IEEE Non-Volatile Semiconductor Memory Workshop*, 2007.

[15] M. H. Kryder and C. S. Kim, "After hard drives - what comes next?", *Magnetics, IEEE Transactions on*, vol. 45, no. 10, pp. 3406–3413, 2009.

[16] B. De Salvo, *Silicon non-volatile memories: paths of innovation*. John Wiley & Sons, 2009.

[17] J. Childress, "Long live data: opportunities and challenges for emerging NVM", in *Leti Memory Workshop 2015*, Jun. 2015.

[18] Y. de Charantenay, "Emerging non volatile memory (NVM) market trends - technical choices are about to be made by key players - STTMRAM or RRAM?", in *Leti Memory Workshop 2015*, Jun. 2015.

[19] S. S. Parkin, "Spintronic materials and devices: past, present and future", in *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International*, Dec. 2004, pp. 903–906. DOI: 10.1109/IEDM.2004.1419328.

[20] M. K. Qureshi, V. Srinivasan, and J. A. Rivers, "Scalable high performance main memory system using phase-change memory technology", *SIGARCH Comput. Archit. News*, vol. 37, no. 3, pp. 24–33, Jun. 2009, ISSN: 0163-5964. DOI: 10.1145/1555815.1555760. [Online]. Available: http://doi.acm.org/10.1145/1555815.1555760.

[21] S. Raoux, G. W. Burr, M. J. Breitwisch, C. T. Rettner, Y.-C. Chen, R. M. Shelby, M. Salinga, D. Krebs, S.-H. Chen, H.-L. Lung, *et al.*, "Phase-change random access memory: a scalable technology", *IBM Journal of Research and Development*, vol. 52, no. 4.5, pp. 465–479, 2008.

[22] H. Cheng, J. Wu, R. Cheek, S. Raoux, M. BrightSky, D. Garbin, S. Kim, T. Hsu, Y. Zhu, E. Lai, *et al.*, "A thermally robust phase change memory by engineering the Ge/N concentration in $(Ge, N)_xSb_yTe_z$ phase change material", in *Electron Devices Meeting (IEDM), 2012 IEEE International*, IEEE, 2012, pp. 31–1.

[23] Q. Hubert, C. Jahan, A. Toffoli, G. Navarro, S. Chandrashekar, P. Noé, D. Blachier, V. Sousa, L. Perniola, J.-F. Nodin, *et al.*, "Lowering the reset current and power consumption of phase-change memories with carbon-doped ge2sb2te5", in *Memory Workshop (IMW), 2012 4th IEEE International*, IEEE, 2012, pp. 1–4.

[24] G. Navarro, M. Coue, A. Kiouseloglou, P. Noe, F. Fillot, V. Delaye, A. Persico, A. Roule, M. Bernard, C. Sabbione, *et al.*, "Trade-off between set and data retention performance thanks to innovative materials for phase-change memory", in *Electron Devices Meeting (IEDM), 2013 IEEE International*, IEEE, 2013, pp. 21–5.

[25] V. Sousa, G. Navarro, N. Castellani, M. Coue, O. Cueto, C. Sabbione, P. Noe, L. Perniola, S. Blonkowski, P. Zuliani, and R. Annunziata, "Operation fundamentals in 12Mb phase change memory based on innovative ge-rich gst materials featuring high reliability performance", in *VLSI Technology (VLSIT), 2015 Symposium on*, Jun. 2015.

[26] S. Raoux, A. K. König, H.-Y. Cheng, D. Garbin, R. W. Cheek, J. L. Jordan-Sweet, and M. Wuttig, "Phase transitions in Ga–Sb phase change alloys", *Physica status solidi (b)*, vol. 249, no. 10, pp. 1999–2004, 2012. DOI: 10.1002/pssb.201200370.

[27] Y. V. Pershin and M. Di Ventra, "Memory effects in complex materials and nanoscale systems", *Advances in Physics*, vol. 60, no. 2, pp. 145–227, 2011.

[28] H. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Phase change memory", *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2201–2227, 2010.

[29] J. Liang, R. G. D. Jeyasingh, H.-Y. Chen, and H.-S. P. Wong, "An ultra-low reset current cross-point phase change memory with carbon nanotube electrodes", *Electron Devices, IEEE Transactions on*, vol. 59, no. 4, pp. 1155–1163, 2012.

[30] Q. Hubert, C. Jahan, A. Toffoli, V. Delaye, D. Lafond, H. Grampeix, and B. De Salvo, "Detailed analysis of the role of thin-interfacial layer in-based pcm", *Electron Devices, IEEE Transactions on*, vol. 60, no. 7, pp. 2268–2275, 2013.

[31] M. Stanisavljevic, A. Athmanathan, N. Papandreou, H. Pozidis, and E. Eleftheriou, "Phase-change memory: feasibility of reliable multilevel-cell storage and retention at elevated temperatures", in *Reliability Physics Symposium (IRPS), 2015 IEEE International*, IEEE, 2015, 5B–6.

[32] D. Ielmini, S. Lavizzari, D. Sharma, and A. L. Lacaita, "Physical interpretation, modeling and impact on phase change memory (PCM) reliability of resistance drift due to chalcogenide structural relaxation", in *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, IEEE, 2007, pp. 939–942.

[33] A. Khvalkovskiy, D. Apalkov, S. Watts, R. Chepulskii, R. Beach, A. Ong, X. Tang, A. Driskill-Smith, W. Butler, P. Visscher, *et al.*, "Basic principles of stt-mram cell operation in memory arrays", *Journal of Physics D: Applied Physics*, vol. 46, no. 7, pp. 74 001–74 020, 2013.

[34] M. Julliere, "Tunneling between ferromagnetic films", *Physics letters A*, vol. 54, no. 3, pp. 225–226, 1975.

[35] J. C. Slonczewski, "Conductance and exchange coupling of two ferromagnets separated by a tunneling barrier", *Physical Review B*, vol. 39, no. 10, p. 6995, 1989.

[36] ——, "Current-driven excitation of magnetic multilayers", *Journal of Magnetism and Magnetic Materials*, vol. 159, no. 1, pp. L1–L7, 1996.

[37] L. Berger, "Emission of spin waves by a magnetic multilayer traversed by a current", *Physical Review B*, vol. 54, no. 13, p. 9353, 1996.

[38] A. D. Kent and D. C. Worledge, "A new spin on magnetic memories", *Nature nanotechnology*, vol. 10, no. 3, pp. 187–191, 2015.

[39] T. Min, Q. Chen, R. Beach, G. Jan, C. Horng, W. Kula, T. Torng, R. Tong, T. Zhong, D. Tang, *et al.*, "A study of write margin of spin torque transfer magnetic random access memory technology", *Magnetics, IEEE Transactions on*, vol. 46, no. 6, pp. 2322–2327, 2010.

[40] E. Vianello, G. Molas, F. Longnos, P. Blaise, E. Souchier, C. Cagli, G. Palma, J. Guy, M. Bernard, M. Reyboz, *et al.*, "Sb-doped ges 2 as performance and reliability booster in conductive bridge ram", in *Electron Devices Meeting (IEDM), 2012 IEEE International*, IEEE, 2012, pp. 31–5.

[41] J. Guy, G. Molas, E. Vianello, F. Longnos, S. Blanc, C. Carabasse, M. Bernard, J. Nodin, A. Toffoli, J. Cluzel, *et al.*, "Investigation of the physical mechanisms governing data-retention in down to 10nm nano-trench al 2 o 3/cutege conductive bridge ram (cbram)", in *Electron Devices Meeting (IEDM), 2013 IEEE International*, IEEE, 2013, pp. 30–2.

[42] M. Barci, J. Guy, G. Molas, E. Vianello, A. Toffoli, J. Cluzel, A. Roule, M. Bernard, C. Sabbione, L. Perniola, *et al.*, "Impact of set and reset conditions on cbram high temperature data retention", in *Reliability Physics Symposium, 2014 IEEE International*, IEEE, 2014, 5E–3.

[43] M. Kund, G. Beitel, C.-U. Pinnow, T. Rohr, J. Schumann, R. Symanczyk, K.-D. Ufert, and G. Muller, "Conductive bridging ram (cbram): an emerging non-volatile memory technology scalable to sub 20nm", in *IEEE International-Electron Devices Meeting, 2005. IEDM Technical Digest.*.

[44] R. Waser and M. Aono, "Nanoionics-based resistive switching memories", *Nature materials*, vol. 6, no. 11, pp. 833–840, 2007.

[45] G. Palma, E. Vianello, O. Thomas, M. Suri, S. Onkaraiah, A. Toffoli, C. Carabasse, M. Bernard, A. Roule, O. Pirrotta, *et al.*, "Interface engineering of ag–based conductive bridge ram for reconfigurable logic applications", *Electron Devices, IEEE Transactions on*, vol. 61, no. 3, pp. 793–800, 2014.

[46] E. Vianello, O. Thomas, G. Molas, O. Turkyilmaz, N. Jovanović, D. Garbin, G. Palma, M. Alayan, C. Nguyen, J. Coignus, *et al.*, "Resistive memories for ultra-low-power embedded computing design", in *Electron Devices Meeting (IEDM), 2014 IEEE International*, IEEE, 2014, pp. 6–3.

[47] G. Palma, E. Vianello, C. Cagli, G. Molas, M. Reyboz, P. Blaise, B. De Salvo, F. Longnos, and F. Dahmani, "Experimental investigation and empirical modeling of the set and reset kinetics of Ag-GeS2 conductive bridging memories", in *Memory Workshop (IMW), 2012 4th IEEE International*, IEEE, 2012, pp. 1–4.

[48] H.-S. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. Chen, and M.-J. Tsai, "Metal-oxide RRAM", *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, 2012, ISSN: 0018-9219.

[49] M. Fujimoto, H. Koyama, M. Konagai, Y. Hosoi, K. Ishihara, S. Ohnishi, and N. Awaya, "Tio2 anatase nanolayer on tin thin film exhibiting high-speed bipolar resistive switching", *Applied Physics Letters*, vol. 89, no. 22, p. 223 509, 2006.

[50] H. D. Lee, B. Magyari-Köpe, and Y. Nishi, "Model of metallic filament formation and rupture in nio for unipolar switching", *Physical Review B*, vol. 81, no. 19, p. 193 202, 2010.

[51] Y. Li, S. Long, H. Lv, Q. Liu, M. Wang, H. Xie, K. Zhang, X. Yang, and M. Liu, "Novel self-compliance bipolar 1d1r memory device for high-density rram application", in *Memory Workshop (IMW), 2013 5th IEEE International*, IEEE, 2013, pp. 184–187.

[52] Y. Yang, P. Gao, S. Gaba, T. Chang, X. Pan, and W. Lu, "Observation of conducting filament growth in nanoscale resistive memories", *Nature communications*, vol. 3, p. 732, 2012.

[53] Q. Liu, J. Sun, H. Lv, S. Long, K. Yin, N. Wan, Y. Li, L. Sun, and M. Liu, "Real-time observation on dynamic growth/dissolution of conductive filaments in oxide-electrolyte-based reram", *Advanced Materials*, vol. 24, no. 14, pp. 1844–1849, 2012.

[54] M. Mongillo, D. Garbin, G. Navarro, E. Vianello, M. Coue, B. Mayall, and D. Cooper, "In-situ biasing and switching of electronic devices into a TEM", in *18th International Microscopy Congress (IMC 2014)*, 2014, IT–7.

[55] L. Larcher, F. Puglisi, P. Pavan, A. Padovani, L. Vandelli, and G. Bersuker, "A compact model of program window in hfo x rram devices for conductive filament characteristics analysis", *Electron Devices, IEEE Transactions on*, vol. 61, no. 8, pp. 2668–2673, 2014.

[56] X. Guan, S. Yu, and H.-S. Wong, "On the switching parameter variation of metal-oxide RRAM-Part I: physical modeling and simulation methodology", *Electron Devices, IEEE Transactions on*, vol. 59, no. 4, pp. 1172–1182, 2012.

[57] A. Fantini, L. Goux, R. Degraeve, D. Wouters, N. Raghavan, G. Kar, A. Belmonte, Y.-Y. Chen, B. Govoreanu, and M. Jurczak, "Intrinsic switching variability in HfO2 RRAM", in *Memory Workshop (IMW), 2013 5th IEEE International*, IEEE, 2013, pp. 30–33.

[58] S. Long, X. Lian, C. Cagli, X. Cartoixa, R. Rurali, E. Miranda, D. Jiménez, L. Perniola, M. Liu, and J. Suñé, "Quantum-size effects in hafnium-oxide resistive switching", *Applied Physics Letters*, vol. 102, no. 18, p. 183 505, 2013.

[59] D. Garbin, O. Bichler, E. Vianello, Q. Rafhay, C. Gamrat, L. Perniola, G. Ghibaudo, and B. DeSalvo, "Variability-tolerant convolutional neural network for pattern recognition applications based on OxRAM synapses", in *Electron Devices Meeting (IEDM), 2014 IEEE International*, IEEE, 2014, pp. 28–4.

[60] S. Lai and T. Lowrey, "Oum-a 180 nm nonvolatile memory cell element technology for stand alone and embedded applications", in *Electron Devices Meeting, 2001. IEDM'01. Technical Digest. International*, IEEE, 2001, pp. 36–5.

[61] A. Bette, J. DeBrosse, D. Gogl, H. Hoenigschmid, R. Robertazzi, C. Arndt, D. Braun, D. Casarotto, R. Havreluk, S. Lammers, *et al.*, "A high-speed 128 kbit mram core for future universal memory applications", in *VLSI Circuits, 2003. Digest of Technical Papers. 2003 Symposium on*, IEEE, 2003, pp. 217–220.

[62] J. Åkerman, "Toward a universal memory", *Science*, vol. 308, no. 5721, pp. 508–510, 2005.

[63] International Technology Roadmap for Semiconductors (ITRS), *Emerging research devices*, 2013.

[64] D. Kuzum, S. Yu, and H. P. Wong, "Synaptic electronics: materials, devices and applications", *Nanotechnology*, vol. 24, no. 38, p. 382 001, 2013.

[65] M. Suri, "Technologies émergentes de mémoire résistive pour les systèmes et application neuromorphique", PhD thesis, Université de Grenoble, 2013.

[66] G. Indiveri and S.-C. Liu, "Memory and information processing in neuromorphic systems", *Proceedings of the IEEE*, vol. 103, no. 8, pp. 1379–1397, 2015.

[67] B. DeSalvo, E. Vianello, D. Garbin, O. Bichler, and L. Perniola, "From memory in our brain to emerging resistive memories in neuromorphic systems", in *Memory Workshop (IMW), 2015 IEEE International*, May 2015, pp. 1–4. DOI: 10.1109/IMW.2015.7150286.

[68] M. Ritter, "Cognitive computing: new ways of thinking", in *IBM Research Colloquia*, 2012.

[69] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity", *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[70] D. O. Hebb, *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[71] A. Turing, *Intelligent machinery. report for national physical laboratory. reprinted in ince, dc (editor). 1992. mechanical intelligence: collected works of am turing*, 1948.

[72] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.", *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[73] M. L. Minsky and S. A. Papert, *Perceptrons - Expanded Edition: An Introduction to Computational Geometry*. MIT press Boston, MA: 1987.

[74] P. WERBOS, "Beyond regression: new fools for prediction and analysis in the behavioral sciences", *PhD thesis, Harvard University*, 1974.

[75] D. E. Rumelhart, J. L. McClelland, P. R. Group, *et al.*, *Parallel distributed processing*. IEEE, 1988, vol. 1.

[76] C. Mead, "Neuromorphic electronic systems", *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1629–1636, 1990.

[77] J.-Y. Boulet, D. Louis, C. Godefroy, A. Steimle, P. Tannhof, and G. Paillet, *Neuron circuit*, US Patent 5,621,863, Apr. 1997.

[78] M. Holler, S. Tam, H. Castro, and R. Benson, "An electrically trainable artificial neural network (etann) with 10240'floating gate'synapses", in *Neural Networks, 1989. IJCNN., International Joint Conference on*, IEEE, pp. 191–196.

[79] U. Ramacher, W. Raab, N. Bruls, M. Wesseling, E. Sicheneder, J. Glass, A. Wurz, and R. Manner, "Synapse-1: a high-speed general purpose parallel neurocomputer system", in *Parallel Processing Symposium, 1995. Proceedings., 9th International*, IEEE, 1995, pp. 774–781.

[80] C. Gamrat, A. Mougin, P. Peretto, and O. Ulrich, "The architecture of mind neurocomputers", in *MicroNeuro Int. Conf. on Microelectronics for Neural Networks, Munich, Germany*, 1991, pp. 463–469.

[81] G.-q. Bi and M.-m. Poo, "Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type", *The Journal of neuroscience*, vol. 18, no. 24, pp. 10 464–10 472, 1998.

[82] ——, "Synaptic modification by correlated activity: hebb's postulate revisited", *Annual review of neuroscience*, vol. 24, no. 1, pp. 139–166, 2001.

[83] S. Thorpe, D. Fize, C. Marlot, *et al.*, "Speed of processing in the human visual system", *Nature*, vol. 381, no. 6582, pp. 520–522, 1996.

[84] J. White, E. Southgate, J. Thomson, and S. Brenner, "The structure of the nervous system of the nematode caenorhabditis elegans: the mind of a worm", *Phil. Trans. R. Soc. Lond*, vol. 314, pp. 1–340, 1986.

[85] J. Karey, L. Ariniello, and M. McComb, *Brain facts: a primer on the brain and nervous system*. Washington, D.C: Society for Neuroscience, 2002, ISBN: 0916110001.

[86] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. Van Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Häfliger, S. Renaud, *et al.*, "Neuromorphic silicon neuron circuits", *Frontiers in neuroscience*, vol. 5, 2011.

[87] G. Palma, M. Suri, D. Querlioz, E. Vianello, and B. De Salvo, "Stochastic neuron design using conductive bridge ram", in *Nanoscale Architectures (NANOARCH), 2013 IEEE/ACM International Symposium on*, IEEE, 2013, pp. 95–100.

[88] J. B. Lont and W. Guggenbühl, "Analog CMOS implementation of a multilayer perceptron with nonlinear synapses", *Neural Networks, IEEE Transactions on*, vol. 3, no. 3, pp. 457–465, 1992.

[89] B. W. Lee and B. J. Sheu, "General-purpose neural chips with electrically programmable synapses and gain-adjustable neurons", *IEEE journal of solid-state circuits*, vol. 27, no. 9, pp. 1299–1302, 1992.

[90] S. Saïghi, C. G. Mayr, T. Serrano-Gotarredona, H. Schmidt, G. Lecerf, J. Tomas, J. Grollier, S. Boyn, A. F. Vincent, D. Querlioz, *et al.*, "Plasticity in memristive devices for spiking neural networks", *Frontiers in neuroscience*, vol. 9, 2015.

[91] H. Choi, H. Jung, J. Lee, J. Yoon, J. Park, D.-j. Seong, W. Lee, M. Hasan, G.-Y. Jung, and H. Hwang, "An electrically modifiable synapse array of resistive switching memory", *Nanotechnology*, vol. 20, no. 34, p. 345 201, 2009.

[92] K. Seo, I. Kim, S. Jung, M. Jo, S. Park, J. Park, J. Shin, K. P. Biju, J. Kong, K. Lee, *et al.*, "Analog memory and spike-timing-dependent plasticity characteristics of a nanoscale titanium oxide bilayer resistive switching device", *Nanotechnology*, vol. 22, no. 25, p. 254 023, 2011.

[93] S. Park, H. Kim, M. Choo, J. Noh, A. Sheri, S. Jung, K. Seo, J. Park, S. Kim, W. Lee, *et al.*, "Rram-based synapse for neuromorphic system with pattern recognition function", in *Electron Devices Meeting (IEDM)*, 2012, pp. 10–2.

[94] S. Park, J. Noh, M.-l. Choo, A. M. Sheri, M. Chang, Y.-B. Kim, C. J. Kim, M. Jeon, B.-G. Lee, B. H. Lee, *et al.*, "Nanoscale rram-based synaptic electronics: toward a neuromorphic computing device", *Nanotechnology*, vol. 24, no. 38, p. 384 009, 2013.

[95] S. Park, A. Sheri, J. Kim, J. Noh, J. Jang, M. Jeon, B. Lee, B. Lee, B. Lee, and H.-j. Hwang, "Neuromorphic speech systems using advanced reram-based synapse", *IEDM Tech Dig*, vol. 25, pp. 1–25, 2013.

[96] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, and H.-S. P. Wong, "An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation", *Electron Devices, IEEE Transactions on*, vol. 58, no. 8, pp. 2729–2737, 2011.

[97] Y. Wu, S. Yu, H.-S. P. Wong, Y.-S. Chen, H.-Y. Lee, S.-M. Wang, P.-Y. Gu, F. Chen, and M.-J. Tsai, "Alox-based resistive switching device with gradual resistance modulation for neuromorphic device application", in *Memory Workshop (IMW), 2012 4th IEEE International*, IEEE, 2012, pp. 1–4.

[98] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H.-S. P. Wong, "A neuromorphic visual system using rram synaptic devices with sub-pj energy and tolerance to variability: experimental characterization and large-scale modeling", in *Electron Devices Meeting (IEDM), 2012 IEEE International*, IEEE, 2012, pp. 10–4.

[99] ——, "A low energy oxide-based electronic synaptic device for neuromorphic visual systems with tolerance to device variation", *Advanced Materials*, vol. 25, no. 12, pp. 1774–1779, 2013.

[100] D. Kuzum, R. G. Jeyasingh, S. Yu, and H.-S. P. Wong, "Low-energy robust neuromorphic computation using synaptic devices", *Electron Devices, IEEE Transactions on*, vol. 59, no. 12, pp. 3489–3494, 2012.

[101] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Phase change memory as synapse for ultra-dense neuromorphic systems: application to complex visual pattern extraction", in *Electron Devices Meeting (IEDM), 2011 IEEE International*, IEEE, 2011, pp. 4–4.

[102] M. Suri, O. Bichler, D. Querlioz, B. Traoré, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Physical aspects of low power synapses based on phase change memory devices", *Journal of Applied Physics*, vol. 112, no. 5, p. 054 904, 2012.

[103] W. Lu, K.-H. Kim, T. Chang, and S. Gaba, "Two-terminal resistive switches (memristors) for memory and logic applications", in *Design Automation Conference (ASP-DAC), 2011 16th Asia and South Pacific*, IEEE, 2011, pp. 217–223.

[104] G. Burr, R. Shelby, C. di Nolfo, J. Jang, R. Shenoy, P. Narayanan, K. Virwani, E. Giacometti, B. Kurdi, and H. Hwang, "Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element", in *Electron Devices Meeting (IEDM), 2014 IEEE International*, IEEE, 2014, pp. 29–5.

[105] M. Suri, O. Bichler, D. Querlioz, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Cbram devices as binary synapses for low-power stochastic neuromorphic systems: auditory (cochlea) and visual (retina) cognitive processing applications", in *Electron Devices Meeting (IEDM), 2012 IEEE International*, IEEE, 2012, pp. 10–3.

[106] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128x128 120 db 15 us latency asynchronous temporal contrast vision sensor", *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 2, pp. 566–576, 2008.

[107] V. Chan, S.-C. Liu, and A. Van Schaik, "Aer ear: a matched silicon cochlea pair with address event representation interface", *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 54, no. 1, pp. 48–59, 2007.

[108] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[109] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex", *The Journal of physiology*, vol. 160, no. 1, p. 106, 1962.

[110] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex", *Cerebral cortex*, vol. 1, no. 1, pp. 1–47, 1991.

[111] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks", *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 3, pp. 328–339, 1989.

[112] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[113] D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification", *Neural Networks*, vol. 32, pp. 333–338, 2012.

[114] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P. E. Barbano, "Toward automatic phenotyping of developing embryos from videos", *Image Processing, IEEE Transactions on*, vol. 14, no. 9, pp. 1360–1371, 2005.

[115] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning", in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, IEEE, 2013, pp. 3626–3633.

[116] R. Vaillant, C. Monrocq, and Y. Le Cun, "Original approach for the localisation of objects in images", *IEE Proceedings-Vision, Image and Signal Processing*, vol. 141, no. 4, pp. 245–250, 1994.

[117] S. J. Nowlan and J. C. Platt, "A convolutional neural network hand tracker", *Advances in Neural Information Processing Systems*, pp. 901–908, 1995.

[118] C. Garcia and M. Delakis, "Convolutional face finder: a neural architecture for fast and robust face detection", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 11, pp. 1408–1423, 2004.

[119] M. Osadchy, Y. L. Cun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models", *The Journal of Machine Learning Research*, vol. 8, pp. 1197–1215, 2007.

[120] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks", *ArXiv preprint arXiv:1411.4280*, 2014.

[121] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: closing the gap to human-level performance in face verification", in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, IEEE, 2014, pp. 1701–1708.

[122] M. Bernacki and P. Włodarczyk. (2004). Principles of training multi-layer neural network using backpropagation, [Online]. Available: `http://home.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html` (visited on Sep. 2, 2015).

[123] H. Markram, W. Gerstner, and P. J. Sjöström, "A history of spike-timing-dependent plasticity", *Frontiers in synaptic neuroscience*, vol. 3, 2011.

[124] O. Kavehei and E. Skafidas, "Highly scalable neuromorphic hardware with 1-bit stochastic nano-synapses", in *Circuits and Systems (ISCAS), 2014 IEEE International Symposium on*, IEEE, 2014, pp. 1648–1651.

[125] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H.-S. P. Wong, "Stochastic learning in oxide binary synaptic device for neuromorphic computing", *Frontiers in neuroscience*, vol. 7, 2013.

[126] A. F. Vincent, J. Larroque, W. S. Zhao, N. Ben Romdhane, O. Bichler, C. Gamrat, J.-O. Klein, S. Galdin-Retailleau, and D. Querlioz, "Spin-transfer torque magnetic memory as a stochastic memristive synapse", in *Circuits and Systems (ISCAS), 2014 IEEE International Symposium on*, IEEE, 2014, pp. 1074–1077.

[127] D. H. Goldberg, G. Cauwenberghs, and A. G. Andreou, "Probabilistic synaptic weighting in a reconfigurable network of vlsi integrate-and-fire neurons", *Neural Networks*, vol. 14, no. 6, pp. 781–793, 2001.

[128] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[129] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, 2013.

[130] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation", in *Advances in Neural Information Processing Systems*, 2014, pp. 1799–1807.

[131] Q. V. Le, "Building high-level features using large scale unsupervised learning", in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, IEEE, 2013, pp. 8595–8598.

[132] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Černockỳ, "Strategies for training large scale neural network language models", in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, IEEE, 2011, pp. 196–201.

[133] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups", *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[134] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr", in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, IEEE, 2013, pp. 8614–8618.

[135] M. Paliwal and U. A. Kumar, "Neural networks and statistical techniques: a review of applications", *Expert systems with applications*, vol. 36, no. 1, pp. 2–17, 2009.

[136] T. W. Berger, M. Baudry, R. D. Brinton, J.-s. Liaw, V. Z. Marmarelis, A. Y. Park, B. J. Sheu, and A. R. Tanguay Jr, "Brain-implantable biomimetic electronics as the next era in neural prosthetics", *Proceedings of the IEEE*, vol. 89, no. 7, pp. 993–1012, 2001.

[137] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure-activity relationships", *Journal of chemical information and modeling*, vol. 55, no. 2, pp. 263–274, 2015.

[138] T. Ciodaro, D. Deva, J. De Seixas, and D. Damazio, "Online particle detection with neural networks based on topological calorimetry information", in *Journal of Physics: Conference Series*, IOP Publishing, vol. 368, 2012, p. 012 030.

[139] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning", *Nature communications*, vol. 5, 2014.

[140] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, and W. Denk, "Connectomic reconstruction of the inner plexiform layer in the mouse retina", *Nature*, vol. 500, no. 7461, pp. 168–174, 2013.

[141] M. K. Leung, H. Y. Xiong, L. J. Lee, and B. J. Frey, "Deep learning of the tissue-regulated splicing code", *Bioinformatics*, vol. 30, no. 12, pp. i121–i129, 2014.

[142] H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, *et al.*, "The human splicing code reveals new insights into the genetic determinants of disease", *Science*, vol. 347, no. 6218, p. 1 254 806, 2015.

[143] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch", *The Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[144] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings", *ArXiv preprint arXiv:1406.3676*, 2014.

[145] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On using very large target vocabulary for neural machine translation", *ArXiv preprint arXiv:1412.2007*, 2014.

[146] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks", in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[147] A. Muthuramalingam, S. Himavathi, and E. Srinivasan, "Neural network implementation using fpga: issues and application", *International journal of information technology*, vol. 4, no. 2, pp. 86–92, 2008.

[148] S. Lai, "Current status of the phase change memory and its future", in *Electron Devices Meeting, 2003. IEDM'03 Technical Digest. IEEE International*, IEEE, 2003, pp. 10–1.

[149] M. H. Lankhorst, B. W. Ketelaars, and R. Wolters, "Low-cost and nanoscale non-volatile memory concept for future silicon chips", *Nature materials*, vol. 4, no. 4, pp. 347–352, 2005.

[150] G. Servalli, "A 45nm generation phase change memory technology", in *2009 IEEE International Electron Devices Meeting (IEDM)*, 2009.

[151] R. Bez, "Chalcogenide pcm: a memory technology for next decade", in *Electron Devices Meeting (IEDM), 2009 IEEE International*, IEEE, 2009, pp. 1–4.

[152] J. Oh, J. H. Park, Y. Lim, H. Lim, Y. Oh, J. S. Kim, J. Shin, Y. J. Song, K. Ryoo, D. Lim, *et al.*, "Full integration of highly manufacturable 512mb pram based on 90nm technology", in *Electron Devices Meeting, 2006. IEDM'06. International*, IEEE, 2006, pp. 1–4.

[153] D. Kuzum, R. G. Jeyasingh, B. Lee, and H.-S. P. Wong, "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing", *Nano letters*, vol. 12, no. 5, pp. 2179–2186, 2011.

[154] S. B. Eryilmaz, D. Kuzum, R. Jeyasingh, S. Kim, M. BrightSky, C. Lam, and H.-S. P. Wong, "Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array", *Frontiers in neuroscience*, vol. 8, 2014.

[155] C. D. Wright, Y. Liu, K. I. Kohary, M. M. Aziz, and R. J. Hicken, "Arithmetic and biologically-inspired computing using phase-change materials", *Advanced Materials*, vol. 23, no. 30, pp. 3408–3413, 2011.

[156] B. L. Jackson, B. Rajendran, G. S. Corrado, M. Breitwisch, G. W. Burr, R. Cheek, K. Gopalakrishnan, S. Raoux, C. T. Rettner, A. Padilla, *et al.*, "Nanoscale electronic synapses using phase change devices", *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 9, no. 2, p. 12, 2013.

[157] O. Bichler, M. Suri, D. Querlioz, D. Vuillaume, B. DeSalvo, and C. Gamrat, "Visual pattern extraction using energy-efficient "2-pcm synapse" neuromorphic architecture", *Electron Devices, IEEE Transactions on*, vol. 59, no. 8, pp. 2206–2214, 2012.

[158] O. Bichler, D. Querlioz, S. J. Thorpe, J.-P. Bourgoin, and C. Gamrat, "Extraction of temporally correlated features from dynamic vision sensors with spike-timing-dependent plasticity", *Neural Networks*, vol. 32, pp. 339–348, 2012.

[159] Y. Kondo and Y. Sawada, "Functional abilities of a stochastic logic neural network", *Neural Networks, IEEE Transactions on*, vol. 3, no. 3, pp. 434–443, 1992.

[160] W. Senn and S. Fusi, "Convergence of stochastic learning in perceptrons with binary synapses", *Physical Review E*, vol. 71, no. 6, p. 061 907, 2005.

[161] J. H. Lee and K. K. Likharev, "Defect-tolerant nanoelectronic pattern classifiers", *International Journal of Circuit Theory and Applications*, vol. 35, no. 3, pp. 239–264, 2007.

[162] P. A. Appleby and T. Elliott, "Stable competitive dynamics emerge from multispike interactions in a stochastic model of spike-timing-dependent plasticity", *Neural computation*, vol. 18, no. 10, pp. 2414–2464, 2006.

[163] O. Bichler, D. Querlioz, S. J. Thorpe, J.-P. Bourgoin, and C. Gamrat, "Unsupervised features extraction from asynchronous silicon retina through spike-timing-dependent plasticity", in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, IEEE, 2011, pp. 859–866.

[164] C.-L. Lin, S.-C. Wu, C.-C. Tang, Y.-H. Lai, S.-R. Yang, and S.-C. Wu, "Unipolar resistive switching and retention of rta-treated zinc oxide (zno) resistive ram", in *Physical and Failure Analysis of Integrated Circuits (IPFA), 2011 18th IEEE International Symposium on the*, IEEE, 2011, pp. 1–5.

[165] O. Bichler, "Contribution à la conception d'architecture de calcul auto-adaptative intégrant des nanocomposants neuromorphiques et applications potentielles", PhD thesis, Université Paris Sud-Paris XI, 2012.

[166] O. Bichler, D. Roclin, C. Gamrat, and D. Querlioz, "Design exploration methodology for memristor-based spiking neuromorphic architectures with the xnet event-driven simulator", in *Nanoscale Architectures (NANOARCH), 2013 IEEE/ACM International Symposium on*, IEEE, 2013, pp. 7–12.

[167] D. Ielmini, D. Sharma, S. Lavizzari, and A. L. Lacaita, "Reliability impact of chalcogenide-structure relaxation in phase-change memory (PCM) cells-part i: experimental study", *Electron Devices, IEEE Transactions on*, vol. 56, no. 5, pp. 1070–1077, 2009.

[168] N. Papandreou, H. Pozidis, T. Mittelholzer, G. Close, M. Breitwisch, C. Lam, and E. Eleftheriou, "Drift-tolerant multilevel phase-change memory", in *Memory Workshop (IMW), 2011 3rd IEEE International*, IEEE, 2011, pp. 1–4.

[169] N. Papandreou, H. Pozidis, A. Pantazi, A. Sebastian, M. Breitwisch, C. Lam, and E. Eleftheriou, "Programming algorithms for multilevel phase-change memory", in *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, IEEE, 2011, pp. 329–332.

[170] M. Suri, D. Garbin, O. Bichler, D. Querlioz, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Impact of pcm resistance-drift in neuromorphic systems and drift-mitigation strategy", in *Nanoscale Architectures (NANOARCH), 2013 IEEE/ACM International Symposium on*, IEEE, 2013, pp. 140–145.

[171] T. Diokh, E. Le-Roux, S. Jeannot, M. Gros-Jean, P. Candelier, J. Nodin, V. Jousseaume, L. Perniola, H. Grampeix, T. Cabout, *et al.*, "Investigation of the impact of the oxide thickness and reset conditions on disturb in hfo 2-rram integrated in a 65nm cmos technology", in *Reliability Physics Symposium (IRPS), 2013 IEEE International*, IEEE, 2013, 5E–4.

[172] A. Benoist, S. Blonkowski, S. Jeannot, S. Denorme, J. Damiens, J. Berger, P. Candelier, E. Vianello, H. Grampeix, J. Nodin, *et al.*, "28nm advanced CMOS resistive RAM solution as embedded non-volatile memory", in *Reliability Physics Symposium, 2014 IEEE International*, IEEE, 2014, 2E–6.

[173] D. Garbin, E. Vianello, O. Bichler, M. Azzaz, Q. Rafhay, P. Candelier, C. Gamrat, G. Ghibaudo, B. DeSalvo, and L. Perniola, "On the impact of OxRAM-based synapses variability on convolutional neural networks performance", in *Symposium on Nanoscale Architecture (NANOARCH), 2015 IEEE/ACM International*, IEEE/ACM, 2015, pp. 193–198.

[174] M. Azzaz, A. Benoist, E. Vianello, D. Garbin, E. Jalaguier, C. Cagli, C. Charpin, S. Bernasconi, S. Jeannot, T. Dewolf, G. Audoit, C. Guedj, S. Denorme, P. Candelier, C. Fenouillet-Beranger, and L. Perniola, "Benefit of Al2O3/HfO2 bilayer for BEOL RRAM integration through 16 kb memory cut characterization", in *Solid State Device Research Conference (ESSDERC), 2015 45th European*, IEEE, 2015, pp. 266–269.

[175] Y. Yin Chen, S. Member, B. Govoreanu, S. Member, L. Goux, R. Degraeve, A. Fantini, G. Sankar Kar, D. J. Wouters, G. Groeseneken, J. A. Kittl, M. Jurczak, and L. Altimime, "Balancing SET/RESET pulse for $> 10^{10}$ endurance in HfO_2/Hf 1T1R bipolar RRAM", *Electron Devices, IEEE Transactions on*, vol. 59, no. 12, 2012. DOI: 10.1109/TED.2012.2218607.

[176] D. Ielmini, F. Nardi, and C. Cagli, "Universal reset characteristics of unipolar and bipolar metal-oxide rram", *Electron Devices, IEEE Transactions on*, vol. 58, no. 10, pp. 3246–3253, 2011.

[177] F. Nardi, S. Larentis, S. Balatti, D. C. Gilmer, and D. Ielmini, "Resistive switching by voltage-driven ion migration in bipolar RRAM-part i: experimental study", *Electron Devices, IEEE Transactions on*, vol. 59, no. 9, pp. 2461–2467, 2012.

[178] S. Larentis, F. Nardi, S. Balatti, D. C. Gilmer, and D. Ielmini, "Resistive switching by voltage-driven ion migration in bipolar RRAM-part ii: modeling", *Electron Devices, IEEE Transactions on*, vol. 59, no. 9, pp. 2468–2475, 2012.

[179] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Statistical fluctuations in hfo x resistive–witching memory: part i-set/reset variability", *Electron Devices, IEEE Transactions on*, vol. 61, no. 8, pp. 2912–2919, 2014.

[180] N. Raghavan, R. Degraeve, A. Fantini, L. Goux, D. Wouters, G. Groeseneken, and M. Jurczak, "Stochastic variability of vacancy filament configuration in ultra-thin dielectric RRAM and its impact on off-state reliability", in *IEEE international electron devices meeting (IEDM)*, 2013, pp. 554–557.

[181] F. M. Puglisi, P. Pavan, A. Padovani, and L. Larcher, "A compact model of hafnium-oxide-based resistive random access memory", in *IC Design & Technology (ICICDT), 2013 International Conference on*, 2013, pp. 85–88.

[182] L. Larcher, A. Padovani, O. Pirrotta, L. Vandelli, and G. Bersuker, "Microscopic understanding and modeling of HfO2 RRAM device physics", in *Electron Devices Meeting (IEDM), 2012 IEEE International*, IEEE, 2012, pp. 20–1.

[183] M. Barci, G. Molas, A. Toffoli, M. Bernard, A. Roule, C. Cagli, J. Cluzel, E. Vianello, B. De Salvo, and L. Perniola, "Bilayer metal-oxide CBRAM technology for improved window margin and reliability", in *Memory Workshop (IMW), 2015 7th IEEE International*, IEEE, 2015, pp. 1–4.

[184] L. Goux, A. Fantini, G. Kar, Y. Chen, N. Jossart, R. Degraeve, S. Clima, B. Govoreanu, G. Lorenzo, G. Pourtois, D. Wouters, J. Kittl, L. Altimime, and M. Jurczak, "Ultralow sub-500nA operating current high-performance TiNAl2O3HfO2HfTiN bipolar RRAM achieved through understanding-based stack-engineering", in *VLSI Technology (VLSIT), 2012 Symposium on*, Jun. 2012, pp. 159–160.

[185] Y.-B. Kim, S. R. Lee, D. Lee, C. B. Lee, M. Chang, J. H. Hur, M.-J. Lee, G.-S. Park, C. J. Kim, U. Chung, *et al.*, "Bi-layered RRAM with unlimited endurance and extremely uniform switching", in *VLSI Technology (VLSIT), 2011 Symposium on*, IEEE, 2011, pp. 52–53.

[186] B. Govoreanu, G. Kar, Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altimime, D. Wouters, J. Kittl, and M. Jurczak, "10x10nm2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation", in *Electron Devices Meeting (IEDM), 2011 IEEE International*, Dec. 2011, pp. 31.6.1–31.6.4.

[187] T.-y. Liu, T. H. Yan, R. Scheuerlein, Y. Chen, J. Lee, G. Balakrishnan, G. Yee, H. Zhang, A. Yap, J. Ouyang, *et al.*, "A 130.7-2-layer 32-Gb ReRAM memory device in 24-nm technology", *Solid-State Circuits, IEEE Journal of*, vol. 49, no. 1, pp. 140–153, 2014.

[188] S. Ambrogio, S. Balatti, F. Nardi, S. Facchinetti, and D. Ielmini, "Spike-timing dependent plasticity in a transistor-selected resistive switching memory", *Nanotechnology*, vol. 24, no. 38, p. 384 012, 2013.

[189] Z. Wang, S. Ambrogio, S. Balatti, and D. Ielmini, "A 2-transistor/1-resistor artificial synapse capable of communication and stochastic learning in neuromorphic systems", *Frontiers in neuroscience*, vol. 8, 2014.

[190] J. Bill and R. Legenstein, "A compound memristive synapse model for statistical learning through stdp in spiking neural networks", *Frontiers in neuroscience*, vol. 8, 2014.

[191] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis", in *Null*, IEEE, 2003, p. 958.

[192] Y. LeCun, C. Cortes, and C. J. Burges. (1998). The mnist database of handwritten digits, [Online]. Available: `http://yann.lecun.com/exdb/mnist/`.

[193] O. Bichler, D. Garbin, E. Vianello, L. Perniola, B. DeSalvo, and C. Gamrat. (2015). Implementing deep neural networks with non volatile memories, [Online]. Available: `http://www.gdr-isis.fr/neurostic/wp-content/uploads/2015/07/NeuroSTIC2015_O.Bichlet.pdf` (visited on Sep. 10, 2015).

[194] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification", in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, 2012, pp. 3642–3649.

[195] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on cpus", in *Proc. Deep Learning and Unsupervised Feature Learning NIPS Workshop*, vol. 1, 2011.

[196] R. Fackenthal, M. Kitagawa, W. Otsuka, K. Prall, D. Mills, K. Tsutsui, J. Javanifard, K. Tedrow, T. Tsushima, Y. Shibahara, *et al.*, "19.7 a 16Gb ReRAM with 200MB/s write and 1GB/s read in 27nm technology", in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International*, IEEE, 2014, pp. 338–339.

[197] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The german traffic sign recognition benchmark: a multi-class classification competition", in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, IEEE, 2011, pp. 1453–1460.

[198] I. Baek, C. Park, H. Ju, D. Seong, H. Ahn, J. Kim, M. Yang, S. Song, E. Kim, S. Park, *et al.*, "Realization of vertical resistive memory (vrram) using cost effective 3d process", in *Electron Devices Meeting (IEDM), 2011 IEEE International*, IEEE, 2011, pp. 31–8.

[199] G. Piccolboni, G. Molas, J. Portal, R. Coquand, M. Bocquet, D. Garbin, E. Vianello, C. Carabasse, V. Delaye, C. Pellissier, T. Magis, C. Cagli, M. Gely, O. Cueto, D. Deletuyelle, G. Ghibaudo, B. De Salvo, and L. Perniola, "Investigation of the potentialities of vertical resistive RAM (VRAM) for neuromorphic applications", in *Electron Devices Meeting (IEDM), 2015 IEEE International*, IEEE, 2015, pp. 447–450.

[200] A. Joubert, B. Belhadj, O. Temam, and R. Héliot, "Hardware spiking neurons design: analog or digital?", in *Neural Networks (IJCNN), The 2012 International Joint Conference on*, IEEE, 2012, pp. 1–5.

[201] D. Garbin, M. Suri, O. Bichler, D. Querlioz, C. Gamrat, and B. DeSalvo, "Probabilistic neuromorphic system using binary phase-change memory (pcm) synapses: detailed power consumption analysis", in *Nanotechnology (IEEE-NANO), 2013 13th IEEE Conference on*, IEEE, 2013, pp. 91–94.

[202] D. Garbin, Q. Rafhay, E. Vianello, S. Jeannot, P. Candelier, B. DeSalvo, G. Ghibaudo, and L. Perniola, "Modeling of OxRAM variability from low to high resistance state using a stochastic trap assisted tunneling-based resistor network", *2015 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, pp. 125–128, 2015.

[203] D. Garbin, E. Vianello, Q. Rafhay, M. Azzaz, S. Jeannot, P. Candelier, B. DeSalvo, G. Ghibaudo, and L. Perniola, "Resistive memory variability: a simplified trap-assisted tunneling model", *Solid-State Electronics*, vol. 115, pp. 126–132, 2016. DOI: 10.1016/j.sse.2015.09.004.

[204] I.-T. Wang, Y.-C. Lin, Y.-F. Wang, C.-W. Hsu, and T.-H. Hou, "3D synaptic architecture with ultralow sub-10 fJ energy per spike for neuromorphic computation", in *Electron Devices Meeting (IEDM), 2014 IEEE International*, IEEE, 2014, pp. 28–5.

[205] "IEEE standard VHDL language reference manual", *IEEE Std 1076-2008 (Revision of IEEE Std 1076-2002)*, pp. c1–626, Jan. 2009. DOI: 10.1109/IEEESTD.2009.4772740.

[206] "IEEE standard for Verilog hardware description language", *IEEE Std 1364-2005 (Revision of IEEE Std 1364-2001)*, pp. 1–560, 2006. DOI: 10.1109/IEEESTD.2006.99495.

[207] "IEEE standard for standard SystemC language reference manual", *IEEE Std 1666-2011 (Revision of IEEE Std 1666-2005*, pp. 1–638, Jan. 2012. DOI: 10.1109/IEEESTD.2012.6134619.

[208] N. T. Carnevale and M. L. Hines, *The NEURON book*. Cambridge University Press, 2006.

[209] D. Goodman and R. Brette, "Brian: a simulator for spiking neural networks in python", *Frontiers in neuroinformatics*, vol. 2, 2008.

[210] M.-O. Gewaltig and M. Diesmann, "Nest (neural simulation tool)", *Scholarpedia*, vol. 2, no. 4, p. 1430, 2007.

[211] G. Indiveri and T. K. Horiuchi, "Frontiers in neuromorphic engineering", *Frontiers in neuroscience*, vol. 5, 2011.

[212] M. Mahowald, *An analog VLSI system for stereoscopic vision*. Springer Science & Business Media, 1994.