Politecnico di Milano
Dipartimento di Elettronica, Informazione e Bioingegneria
Doctoral Programme In Information Technology

Modeling and simulation of spiking neural networks with resistive switching synapses

Doctoral Dissertation of:
Valerio Milo

Supervisor:
Prof. Daniele Ielmini

Tutor:
Prof. Carlo Ettore Fiorini

The Chair of the Doctoral Program:
Prof. Barbara Pernici

2018 – Cycle XXXI

Contents

Abstract

DURING the last five decades, the microelectronics industry has been steadily evolving thanks to the Moore's law predicting an exponential increase of the number of transistors on the chip, and the increase of clock frequency at each technology generation.

Currently, this scaling trend is coming to an end mainly due to the excessive power dissipation. In addition, performance gap between the central processing unit (CPU) and the off-chip working memory makes current digital processors based on conventional von Neumann architecture inefficient in terms of energy and latency especially for the implementation of emerging data-centric applications such as big data analytics and machine learning tasks.

To face these challenges, emerging memory devices, also known as memristors, such as resistive random access memory (RRAM), phase change memory (PCM) and spin-transfer torque magnetic random access memory (STT-MRAM) have recently gained significant interest for their non-volatility, scalability, low current operation, and compatibility with complementary metal-oxide-semiconductor (CMOS) process.

Moreover, novel approaches aiming to radically subvert von Neumann architecture blurring the distinction between computation and memory have also been subject of intensive research. Among these novel approaches, neuromorphic computing has rapidly attracted considerable attention for its ambitious objective to emulate the brain ability to carry out extremely complex cognitive functions such as learning, recognition, inference, and decision making with an unrivaled energy efficiency due to its event-driven information processing.

To achieve this goal, RRAM can play a key role enabling to replicate synaptic plasticity rules believed to be the origin of learning such as spike-timing dependent plasticity (STDP) and spike-rate dependent plasticity (SRDP) at device level thanks to its tunable resistance. Also, nanoscale size of RRAM devices offers the great opportunity to achieve high-density integration of resistive devices, thus paving the way for the hardware implementation of high-density spiking neural networks with resistive synapses capable of brain-inspired computing.

This doctoral dissertation covers modeling and simulation of spiking neural networks with hybrid CMOS/RRAM synapses capable of bio-realistic learning rules for implementation of brain-inspired cognitive tasks such as unsupervised learning of visual patterns and associative learning.

Chapter 1 first provides an overview of fundamental issues currently challenging the performance improvement of today's digital computing systems based on standard CMOS technology. Moreover, physical mechanisms and key characteristics of the main emerging non-volatile memory technologies, and the novel computing approaches proposed to overcome the current technology paradigm are extensively described.

Chapter 2 focuses on physics-based modeling of HfO_2 RRAM devices. First, a previous numerical model of HfO_2 RRAM providing a detailed understanding of the resistive switching mechanism at device scale is reviewed. After, a previous analytical model of HfO_2 RRAM derived from numerical model is also reviewed. In addition, a stochastic model taking into account the statistical variability of set/reset processes in HfO_2 RRAM devices is described.

Chapter 3 presents two hybrid CMOS/RRAM synapse circuits developed to replicate two fundamental bio-realistic learning rules such as STDP and SRDP. First, the implementation of STDP rule by a hybrid CMOS/R-RAM synapse circuit with one-transistor/one-resistor (1T1R) structure is discussed by simulations and experiments. In addition, the implementation of SRDP rule by a hybrid CMOS/RRAM synapse circuit with 4-transistors/one-resistor (4T1R) structure is also discussed by simulations and experimental measurements.

Chapter 4 covers the implementation of unsupervised learning and recognition of visual patterns by 2-layer feedforward spiking neural networks with 1T1R RRAM synapses capable of STDP at simulation and experimental level. After discussing learning of a single pattern, on-line learning of sequential patterns and multiple patterns is also extensively addressed in simulation and experiments.

Chapter 5 covers the implementation of unsupervised learning of visual patterns by 2-layer feedforward spiking neural networks with 4T1R RRAM synapses capable of SRDP. After discussing learning of a single pattern for variable initial weight configuration, on-line learning of sequential visual patterns is also investigated by simulations at network level.

Finally, chapter 6 presents a circuit implementation of a Hopfield recurrent spiking neural network with excitatory and inhibitory 1T1R RRAM synapses capable of STDP. After discussing learning and recall of both a single attractor state and a sequence of two non-overlapping attractor states via simulations, RRAM-based Hopfield network is used to explore fundamental human brain primitives such as associative memory, pattern completion, and error correction.

CHAPTER 1

Memory and computing beyond Moore's law

1.1 Introduction

In recent years, the tremendous growth of computing devices has revolutionized our society paving the way to the era of Internet of Things (IoT) and Big Data.

In this scenario, new challenges are emerging as urgent priorities. First, to reduce the excessive energy and latency costs due to the intensive data movement towards the clouds for processing operations, the massive amount of data generated by smart IoT sensors should be elaborated in real-time, which would benefit especially some emerging applications such as active health monitoring and decision-making in driver-less cars or robots. Also, the today's dramatic growth of Internet data requires increasingly fast and scalable memory technologies to offer more storage capacity in portable computers and data centers.

To meet these challenges, the exploration of new technological solutions and novel paradigms thus plays a crucial role.

Since the 1960s, the extraordinary advances achieved in computing and

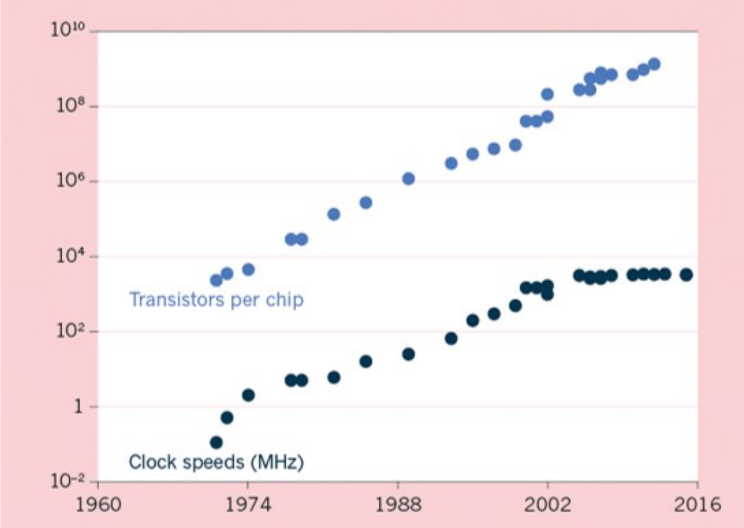


Figure 1.1: *Scaling trend of the number of transistors per chip and processor operating frequency over the past 50 years. According to Moore's law, the downscaling of transistor size led to an exponential increase of device density on the chip, and the increase of clock frequency generation after generation. However, in recent years these trends have slowed down abruptly mainly because of the excessive power dissipation caused by static and dynamic leakage processes. Reprinted from [2].*

information technology have been driven by the miniaturization of metal-oxide-semiconductor field-effect transistor (MOSFET) in step with the Moore's law predicting the doubling of number of integrated transistors in a microprocessor chip approximately every two years [1]. This has resulted in an exponential increase of device density on the chip, and an increase of operating frequency generation after generation, eventually leading to the development of today's digital complementary metal-oxide-semiconductor (CMOS) microprocessors.

However, as shown in Fig. 1.1, fundamental issues have recently slowed down these trends. First, increasing leakage currents do not allow to decrease the threshold voltage of MOSFET devices further, thus hindering the scaling-down of supply voltage and transistor size in digital circuits [3]. Also, the large power consumption achieved by today's CMOS-based microprocessors, which ranges from 50 to 100 W per cm^2 [3, 4], has put a hard limit on the maximum clock frequency because of the excessive heating on the chip referred to as heat wall [2]. As a result, maximum operating frequency has remained almost unchanged since early 2000s.

In addition to the heat wall, another hard barrier known as memory wall is also challenging the Moore's law [5]. In conventional processors, the central processing unit (CPU) carries out operations at a speed much higher than that needed to access the memory where the data are stored, which makes the rate of bus-limited data movement between CPU and memory a severe performance bottleneck [4,6]. The cause of this fundamental issue is the physical separation of CPU and memory in the von Neumann architecture of current digital computers, hence the name von Neumann bottleneck.

Recently, these concerns have been the subject of an intense research leading to the investigation of novel concepts at device, circuit, and system level [7–10]. On the one hand, to tackle scaling issues due to increasing sub-threshold leakage, transistors are being redesigned to improve their channel electrostatic control [11] and sub-threshold slope, possibly overcoming the Boltzmann barrier of 60 mV/dec [12, 13]. On the other hand, to continue improving performance according to Moore's law despite the power-limited operating frequency, more processor cores were integrated on the same chip [2, 3]. The great advantage due to this parallel computation approach is that although each core on the chip is operated at frequency lower than the maximum clock rate not to hit the power barrier, the use of multiple cores at the same time enables to increase the overall performances. However, multi-core processors have not removed the issue because of the difficulty in making many algorithms parallel [2] and achieving a significant energy gain by decreasing clock rate further [3]. Thus, the requirement to achieve a higher energy efficiency has encouraged the transition toward Systems on Chip (SoC) based on the co-integration of CPU, graphics processing unit (GPU), which typically uses hundreds of cores running in parallel with high memory bandwidths [6], and application-specific accelerators, *e.g.* image or video processing accelerators, to combine benefits resulting from specialization and extensive parallelism [3].

In addition to this, novel concepts have also been explored to overcome memory wall. Fig. 1.2 shows the memory hierarchy of conventional processors based on von Neumann architecture. In this hierarchy, cache, main memory, and storage memory are based on static random access memory (SRAM), dynamic random access memory (DRAM), and Flash memory, respectively. SRAM offers an extremely fast access time of about 1-10 ns, but it has the disadvantage to consume an area larger than 100 F^2, where F is the minimum feature size allowed by the lithography. As opposed to SRAM, DRAM is much less expensive in terms of area, only 6 F^2, but it offers a lower operation speed ranging from 10 to 100 ns, whereas Flash memory enables an ultra-high density storage up to Terabytes, albeit typical access time is of the order of hundreds of μs. Therefore, this means that

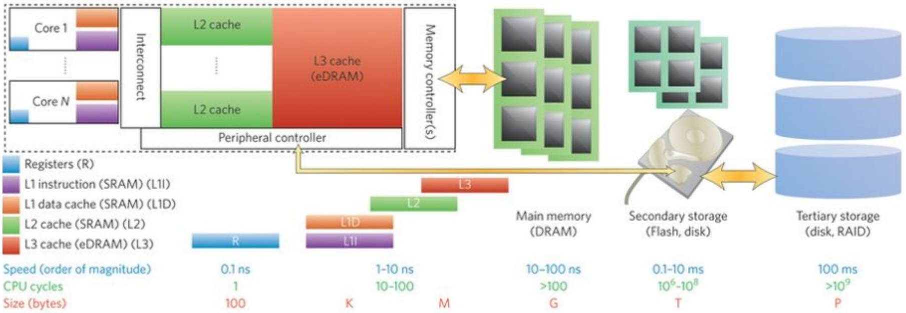


Figure 1.2: *Memory hierarchy in today's computers based on von Neumann architecture. All the memory levels namely SRAM, DRAM, Flash and Hard Disk Drive are shown and compared in terms of access time, number of CPU cycles to access the memory and size. Copyright 2015, Springer Nature. Reprinted, with permission, from [14].*

moving from cache to storage memory, write/read speed decreases whereas memory capacity increases [14, 15].

Based on this hierarchy memory, efficiency and latency burdens of memory wall can thus be removed by increasing the data locality, namely reducing the gap between memory and computing [14].

To this end, in last 15 years emerging memory technologies have been intensively investigated. These novel materials-based devices, also called memristors [16, 17], are non-volatile memory devices with a 2-terminal-based simple structure exhibiting high speed, high density, low power operation and high compatibility with CMOS technology. In particular, as opposed to CMOS-based memory elements such as SRAM, DRAM and Flash memory where information storage is based on charge, emerging memories exploit the physics of active materials to store information. However, although some emerging memories have led to the development of commercial technologies on the market [18–21], they have failed to match challenging performances in terms of operation speed, data bandwidth and cost essential to achieve conclusive removal of memory wall so far [22].

In parallel to the investigation of emerging memory devices, the major difficulties faced to overcome memory wall and the other fundamental barriers have also led to the exploration of novel approaches such as hybrid memory-logic integration, bioinspired computing and in-memory computing [23] (Fig. 1.3), which, benefiting from the progress of memristive technologies, do not aim to re-engineer current systems, but rather to radically subvert von Neumann architecture [22]. Furthermore, growing interest has also been attracted by another approach known as quantum computing [9],

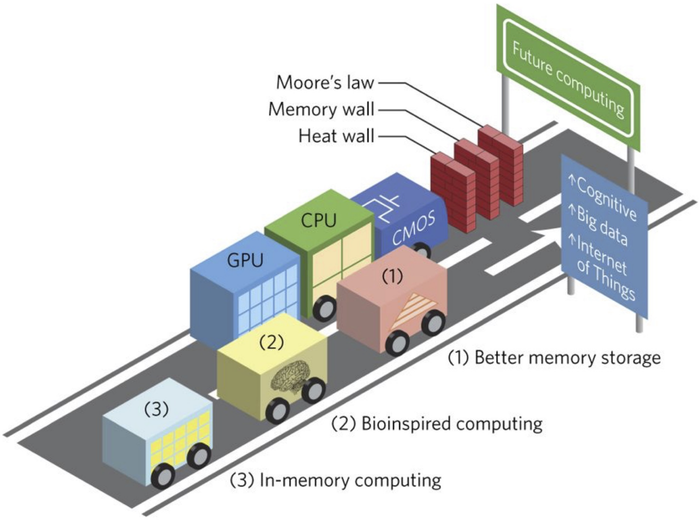


Figure 1.3: *Illustrative picture evidencing three hard barriers challenging today's conventional CMOS-based computing systems namely the end of Moore's law, the memory wall and the heat wall, and some promising approaches which, thanks to development of memristor technology, can play a key role to tackle current limitations and meet new computing challenges including cognitive computing, big data and Internet of Things. Copyright 2018, Springer Nature. Reprinted, with permission, from [23].*

which promises to solve untractable problems for current processors with a dramatic saving in terms of time and cost exploiting quantum mechanics.

Ultimately, based on recent developments at research and industry level, it can reasonably be expected that the achievement of improved devices combined with more efficient computing schemes will thus play a crucial role in coming years since not only it will contribute to hinder the looming end of Moore's law, but also will open the way to a completely new scenario for computing able to meet emerging technological challenges such as brain emulation, analysis of increasingly large databases and implementation of ultra-low power systems for IoT applications with performances we have never seen before.

1.2 Emerging non-volatile memory technologies

Emerging non-volatile memories are two-terminal memory devices based on the resistive switching phenomenon which can be achieved by various physical processes as shown in Fig. 1.4. In this wide range of resistive switching memory concepts, phase change memory (PCM), resistive

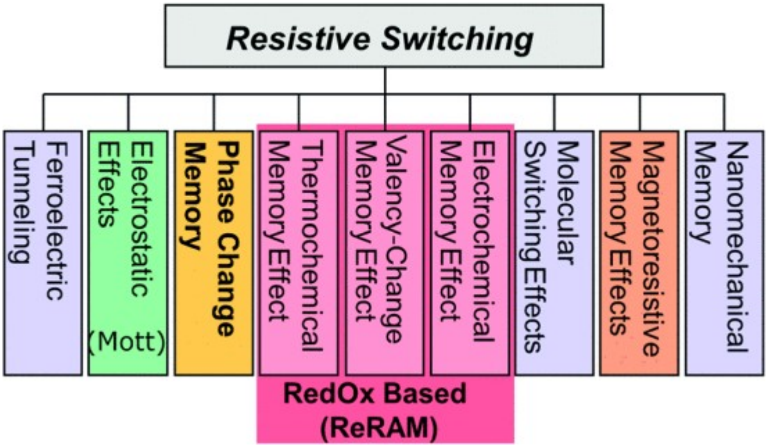

Resistive Switching
Ferroelectric Tunneling
Electrostatic Effects (Mott)
Phase Change Memory
Thermochemical Memory Effect
Valency-Change Memory Effect
Electrochemical Memory Effect
Molecular Switching Effects
Magnetoresistive Memory Effects
Nanomechanical Memory
RedOx Based (ReRAM)

Figure 1.4: *Taxonomy of all the emerging memory devices based on resistive-switching phenomenon. Copyright 2015 IEEE. Reprinted, with permission, from [24].*

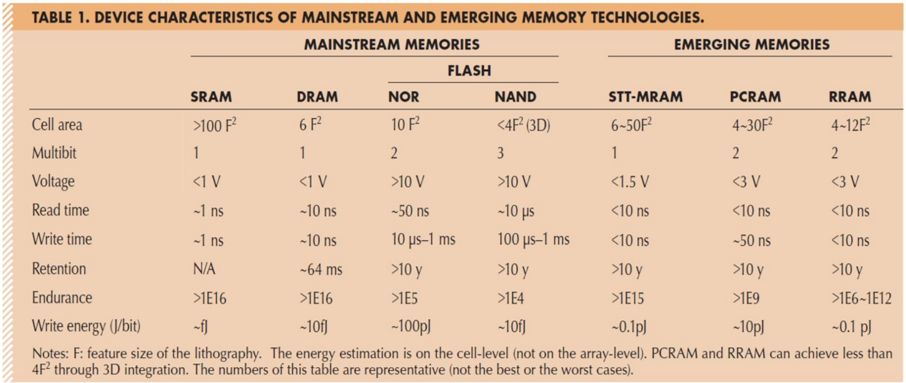


TABLE 1. DEVICE CHARACTERISTICS OF MAINSTREAM AND EMERGING MEMORY TECHNOLOGIES.

	MAINSTREAM MEMORIES				EMERGING MEMORIES		
			FLASH				
	SRAM	DRAM	NOR	NAND	STT-MRAM	PCRAM	RRAM
Cell area	>100 F^2	6 F^2	10 F^2	<4F^2 (3D)	6~50F^2	4~30F^2	4~12F^2
Multibit	1	1	2	3	1	2	2
Voltage	<1 V	<1 V	>10 V	>10 V	<1.5 V	<3 V	<3 V
Read time	~1 ns	~10 ns	~50 ns	~10 µs	<10 ns	<10 ns	<10 ns
Write time	~1 ns	~10 ns	10 µs–1 ms	100 µs–1 ms	<10 ns	~50 ns	<10 ns
Retention	N/A	~64 ms	>10 y	>10 y	>10 y	>10 y	>10 y
Endurance	>1E16	>1E16	>1E5	>1E4	>1E15	>1E9	>1E6~1E12
Write energy (J/bit)	~fJ	~10fJ	~100pJ	~10fJ	~0.1pJ	~10pJ	~0.1 pJ

Notes: F: feature size of the lithography. The energy estimation is on the cell-level (not on the array-level). PCRAM and RRAM can achieve less than 4F^2 through 3D integration. The numbers of this table are representative (not the best or the worst cases).

Figure 1.5: *Table of device characteristics and performance metrics for mainstream CMOS-based memories and main resistive-switching emerging memory technologies. Copyright 2016 IEEE. Reprinted, with permission, from [15].*

random access memory (RRAM), and spin-transfer torque magnetic random access memory (STT-MRAM) have received an increasing interest as strong competitors of the mainstream CMOS-based SRAM, DRAM and Flash memories for their advantageous properties in terms of compactness, power consumption, and operation speed summarized in the table shown in Fig. 1.5 [15]. In particular, simple structure, high switching speed, CMOS compatibility and low power operation make these resistive switching devices very attractive not only as non-volatile storage memory, but also as

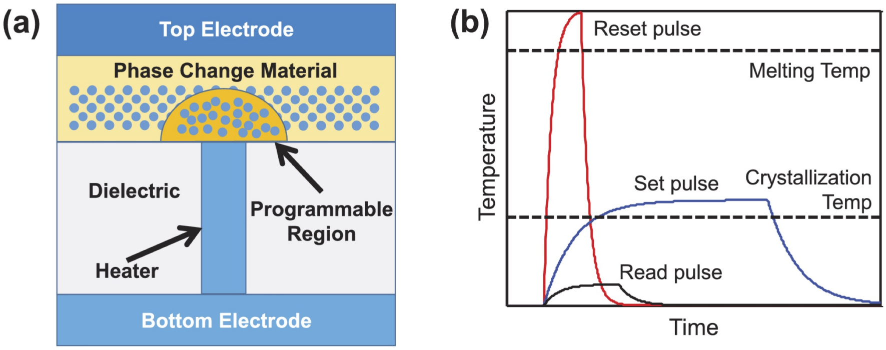


Figure 1.6: *(a) Cross-section scheme of a PCM device with mushroom structure evidencing a phase-change active material sandwiched between the top electrode and the bottom electrode acting as heater. (b) Cell programming or reading occurs by the application of current pulses inducing a temperature increase via Joule heating within the cell. As a long current pulse of medium amplitude leads temperature in PCM device to cross the crystallization temperature, a set transition from amorphous to crystalline phase is achieved. Conversely, as a fast current pulse with high amplitude leads temperature above melting temperature, a reset transition from crystalline to amorphous phase is activated. Copyright 2017 IEEE. Reprinted, with permission, from [26].*

computational memory for neuromorphic and in-memory computing applications [25].

1.2.1 Phase change memory (PCM)

Phase change memory (PCM) is a resistive switching memory device where a chalcogenide material, typically $Ge_2Sb_2Te_5$ (GST), acts as active material. As shown in Fig. 1.6(a), the structure of the conventional mushroom-type PCM cell consists of the GST chalcogenide layer sandwiched between the top electrode and a heater enabling to confine the heat and current, thus leading to a programmable region with hemispheric shape.

In PCM, depending on the Joule heating induced in the GST by applied electrical current pulses, the phase of active material can switch between the amorphous phase and the crystalline phase. Amorphous phase is achieved by the application of a reset current pulse with high amplitude of about 100 μA inducing a large temperature leading to a local melting of programmable region which is then abruptly quenched due to the short pulse width (typically $<$ 50 ns). Achieving the crystallization phase instead requires the application of a set current pulse leading GST to overcome crystallization temperature while keeping temperature below the melting temperature (620°C for GST). To this end, set pulse is designed with amplitude lower than reset pulse and duration in the range 100 ns-10 μs (Fig. 1.6(b)) [24, 26–30].

Set and reset transitions are thus asymmetric and lead to two states with low and high resistance, respectively, where low resistance state (LRS) is typically few kΩ while high resistance state (HRS) is few MΩ. As a result, PCM enables to encode two logic states featuring a relatively large resistive window of about 3 orders of magnitude which makes it potentially capable of multilevel cell (MLC) operation [31].

The most attractive features making PCM a promising candidate for next memory generation are good scalability, multilevel operation, high reliability and high endurance [24, 26, 27, 30]. However, in last two decades, main issues have challenged the development of PCM technology.

Because PCM stores information in the phase of active material, the downscaling of contact area between the bottom electrode and PCM layer results in the downscaling of melted active volume leading to a lower programming current, thus keeping constant the current density [27]. Despite this great advantage in terms of scalability due to the physics of PCM cell, the cell miniaturization process has been hindered by the lithography limits and the need for a transistor as select device to access individual cells in large arrays with no sneak paths.

In addition, another barrier toward the achievement of high device density is the demanding drive current requirement which forces to adopt large transistors [24, 26, 27, 30, 32].

In this scenario, to meet the challenge of high device density, the research efforts focused on the stacking along the third dimension leading to 3D architectures such as 3D-XPoint [20, 21] and investigation of novel bidimensional material interface to obtain better selectors.

At the same time, the attractive perspective to attain multiple stable resistive states (potentially up to 3 bits per cell [26]) in PCM devices is challenged by a physical phenomenon known as resistance drift. Because of this effect, the programming of PCM cell in the high resistance state is followed by a gradual resistance increase with logarithmic behavior over time changing from write cycle to write cycle and from cell to cell, thus making the write operation unpredictable [33, 34]. To mitigate drift resistance impact, various solutions including drift-invariant read techniques [35, 36] have been proposed.

Overall, the introduction of new efficient solutions combined with further advancements at the level of material engineering, cell design and array architecture can lead to a strong reduction of detrimental impact of

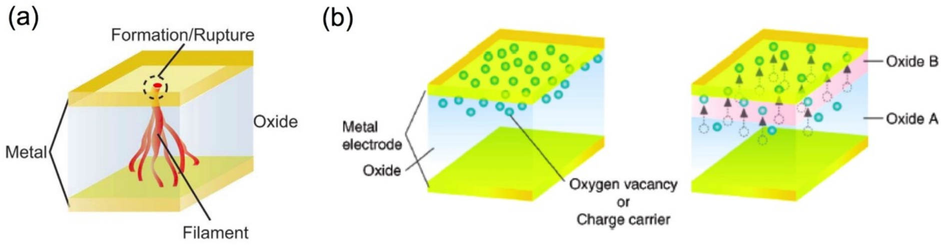


Figure 1.7: *Schematic illustration of RRAM devices based on filamentary resistive switching such as OxRAM and CBRAM (a) and uniform-switching (b). Copyright 2008 Elsevier Ltd. Adapted, with permission, from [37].*

discussed issues on PCM performance in near future, thus making PCM technology an extremely attractive candidate for storage-class memory and embedded non-volatile memory applications.

1.2.2 Resistive switching memory (RRAM)

Resistive switching memory (RRAM) generally consists of a metal-insulator-metal (MIM) stack, where resistance can change as a result of a local modification of the material composition, *e.g.*, along a conductive filamentary (CF), or within an interface layer. This marks the difference between RRAM and PCM, where the resistance change is dictated by a different phase of the active material [29], or magnetic random-access memory (MRAM), where the resistance change results from a re-orientation of the magnetic polarization within a ferromagnetic layer [38].

RRAM offers a simple two-terminal structure, compatibility with CMOS process, back-end of the line (BEOL) process, high speed and low power consumption. Given the large number of switching materials and their possible combination in MIM stacks [39], multilayers [40], and multi-terminal structures [41], RRAM offers an unprecedented flexibility to serve different demands of memory, storage and computing.

In particular, over the last decade, RRAM was used to achieve both large-scale prototypes capable of density higher than 1 GB via one-transistor/one-resistor (1T1R) [42] or cross-point architecture [19], and relatively small-scale (< 10 MB) protoypes suitable for embedded memory applications such as IoT applications [43].

However, significant issues such as the device reliability, namely the control of device variability and noise degrading the stability of data after write operation, and the choice of the best selector device to achieve high-density 3D crossbar architectures, are still open questions requiring additional investigations and improvements at the level of material engineering [25].

Resistive switching in RRAM devices

Resistive switching mechanisms in RRAM devices can be discriminated by the type of localization of the chemical modification responsible for the change of conductance. The two classes of switching phenomena are shown in Fig. 1.7: chemical/conductance modification occurs along a filamentary path, also known as conductive filament (CF), in filamentary switching (a), whereas the change of conductance and composition occurs on an interface region in the case of uniform, or interface, switching (b) [37].

Filamentary switching RRAM

As shown in Fig. 1.8, filamentary switching is generally triggered by a forming operation, namely a soft breakdown operation that creates a locally degraded region with large concentration of defects. In oxide-based RAM, also known as OxRAM, the dielectric switching layer consists of a transition metal oxide such as HfO_x, TiO_x and TaO_x, which is sandwiched between a top and a bottom metal electrode [37,44,45].

After forming, the CF shows a high concentration of metallic impurities and/or oxygen vacancies which are responsible for the low resistance state (LRS) or set state. The CF is electrically disconnected via a reset operation, which generally causes a defect depletion within a relatively limited region along the CF, thus leading to a high resistance state (HRS). The set process can recreate the CF thus supporting filamentary switching [25].

OxRAM can exhibit two switching modes depending on polarity of voltage pulses applied during set and reset operations. If both transitions occur under the positive polarity of applied voltage, resistive switching is referred to as unipolar. In unipolar OxRAM, which was originally reported in NiO [46,47], CF formation and rupture are explained by thermally activated redox reactions [48]. In particular, the reset process leads to CF oxidation resulting in the formation of a depleted gap located at the point of CF at maximum temperature [49, 50] while the set transition involves a chemical reduction of metal oxide induced by Joule heating. In bipolar switching, instead, set and reset processes occur under opposite voltage polarities. In bipolar OxRAM, ion migration driven by electric field and accelerated by temperature is responsible for the CF connection and disruption [51]. During reset, negatively biased top electrode attracts ionized defects such

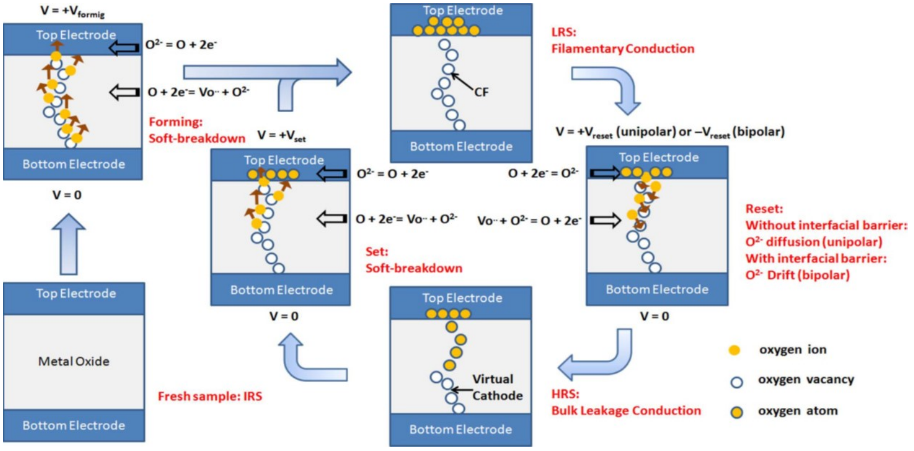


Figure 1.8: *Schematic illustration of the forming, set and reset processes for bipolar and unipolar filamentary RRAM devices. Copyright 2012 IEEE. Reprinted, with permission, from [45].*

as oxygen vacancies disconnecting the CF where the filament temperature is maximum. Set transition, instead, leads to a defect migration into the depleted gap region, causing the creation of a continuous CF whose size is limited by the maximum (compliance) current during the set transition, generally controlled by a transistor or resistance in series with the memory device. In particular, the same defects are migrated in one direction or the other during set/reset transitions in bipolar switching, whereas unipolar switching is assumed to require recreation of defects and their radial diffusion [52]. As a result, bipolar RRAM devices generally exhibit a higher endurance than unipolar RRAM, making bipolar switching overall more attractive for cycling intensive applications. There have been reports where the same device could show the coexistence of both unipolar and bipolar switching behaviors, such as the case of TiN/HfO_2 RRAM [53].

A second type of filamentary switching device is the conductive-bridge RAM, also known as CBRAM [54–57]. In CBRAM, metal impurities, typically cations supplied by Ag or Cu based metallic cap at the top electrode, are injected in a chalcogenide (GeSe, GeS) or oxide (SiO_2, Al_2O_3) electrolyte layer to create conductive paths. As evidenced in Fig. 1.9, set transition consists of the migration of Ag cations from the active top electrode toward the bottom electrode under a positive voltage resulting in the Ag-based CF formation and growth that is controlled by the compliance current (A-D). On the other hand, by applying a negative voltage to the top

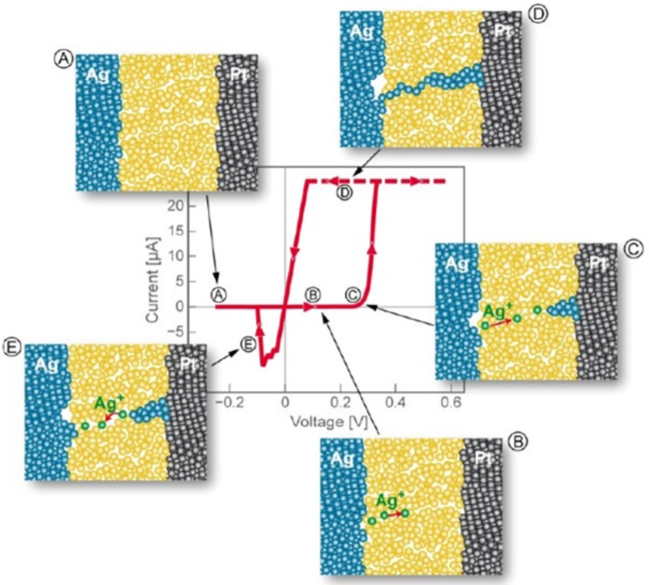


Figure 1.9: *Schematic illustration of the set transition (A-D) and reset transition (E) in CBRAM devices. Copyright 2011 IOP Publishing Ltd. Reprinted, with permission, from [54].*

electrode for reset process, cations migrate in the opposite direction causing a dissolution of the metallic CF (E) [54]. Unipolar switching has been sometimes reported in CBRAM [58]. Despite several similarities in terms of switching and reliability between OxRAM and CBRAM devices, some differences exist. CBRAM shows a ratio between HRS and LRS resistances of about 10^4 that is 2-3 orders of magnitude higher than OxRAM resistance window. The large resistance window is probably due to the higher mobility of Ag/Cu cations compared to the defects in OxRAM resulting in a larger gap and consequently in an increased HRS resistance after reset transition. As a result of the increased HRS, CBRAM devices can also operate at lower programming currents of about 10 pA [59], and feature multilevel cell operation [60].

Uniform switching RRAM

Uniform switching where chemical composition at the origin of the resistance change occurs within the whole device area, was evidenced in other classes of materials, such as perovskite-type oxides, *e.g.*, $Pr_{1-x}Ca_xMnO_3$ (PCMO) [61] and TaO_x/TiO_2 bilayers [62, 63]. Uniform switching was ex-

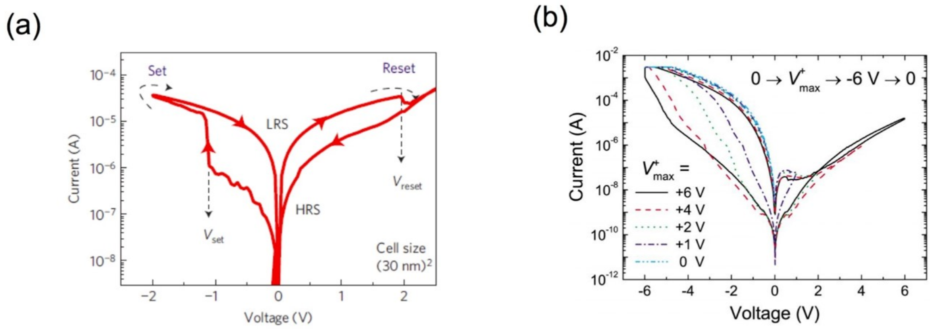


Figure 1.10: *Comparison between the typical I-V curve of a filamentary RRAM such as Ta$_2$O$_{5-x}$/TaO$_{2-x}$ RRAM device (a) and the typical I-V characteristics for a uniform switching RRAM such as Al/PCMO-based RRAM device (b). (a) Copyright 2011 Nature Publishing Group. Reprinted, with permission, from [40]. (b) Copyright 2009 AIP Publishing LLC. Reprinted, with permission, from [61].*

plained as a local chemical reaction taking place at the interface between 2 separate materials. For instance, field-induced oxygen exchange can occur between a reactive top electrode and the oxide layer, *e.g.*, between Sm top electrode and PCMO [64]. Alternatively, oxygen exchange occurs between TiO$_2$ and TaO$_x$, where the latter serves as the barrier oxide controlling HRS/LRS resistance values [62, 63].

Taking as reference the schematic illustration in Fig. 1.7(b), when a positive voltage is applied to the top electrode, oxygen ions and/or electrons drift from the bulk oxide layer toward the top electrode, thus inducing the oxidation of the electrode/oxide interface. Interface switching requires thus the use of a relatively reactive oxide, such as Al or Sm, while inert metals such as Pt do not yield significant resistance change. The resulting oxidized layer causes a resistance increase by enhancing the barrier height in a tunneling or Schottky barrier for electrons/holes injection. The application of a negative voltage results in a switching to LRS because of oxygen migration back to the bulk oxide layer. Since resistivity change occurs across the whole interface area, the HRS/LRS resistance values and the programming currents are generally proportional to the device area [65].

Filamentary and interface switching usually differ also by the shape of their I-V characteristics. Fig. 1.10 shows the I-V characteristics for a filamentary Ta$_2$O$_{5-x}$ /TaO$_{2-x}$ RRAM device [40] (a) and a uniform switching RRAM with Al/PCMO structure (b) [61]. Filamentary switching is marked by an abrupt set transition, which can be explained by a sudden voltage snap back due to the sudden self-accelerated formation and growth of a conductive filament [66]. On the other hand, uniform switching appears as smooth set/reset transition, and usually shows largely asymmetric characteristics due to rectification induced by Schottky barriers or asymmetric tunneling barriers.

1.2.3 Spin-transfer torque magnetic random access memory (STT-MRAM)

In last 20 years, physics and technology advancements have led to novel magnetic memory concepts such as spin-transfer torque magnetic random access memory (STT-MRAM). STT-MRAM device is based on the spin-transfer torque effect, which was theoretically predicted by physicists John Slonczeswki [67] and Luc Berger [68] in 1996 and then observed for the first time in the late 1990s [69–71].

As shown in Fig. 1.11(a), a spin-torque-based memory consists of a magnetic tunnel junction (MTJ) stack with an ultrathin ($\approx$ 1 nm) tunnel oxide layer, such as crystalline MgO, sandwiched between two ferromagnetic layers (generally CoFeB electrodes) referred to as pinned layer and free layer due to their fixed and variable magnetization orientation, respectively.

This structure is characterized by two stable states depending on the relative magnetic orientation of two ferromagnets, namely the parallel (P) state, depicted in Fig. 1.11(b), where the free magnetization direction M_{free} is aligned to fixed magnetization direction M_{fixed}, and the anti-parallel (AP) state, depicted in Fig. 1.11(c), in which M_{free} orientation is not aligned to M_{fixed}.

Also, AP and P states, which are achieved by a reset and set transition, respectively, as evidenced by typical I-V characteristics in Fig. 1.11(d), can exhibit a percentage change in resistance up to about 200% [74] as a result of the tunnel magnetoresistance (TMR) effect [75].

As illustrated in Fig. 1.11(e), the injection of a current through the MTJ can manipulate the magnetic polarization of the free layer leading STT-MRAM device to switch from P to AP state or from AP to P state depending on applied current sign [38, 76, 77].

If the electrons enter from the free layer of a STT-MRAM in P state (left), namely the applied current flows from fixed to free layer, the electrons with a spin opposite to the M_{free} are reflected back to the free layer inducing a spin-transfer torque (STT) capable of exerting a rotation of M_{free}, eventually leading STT-MRAM state from P to AP (reset transition), thus in HRS. Conversely, if the electrons enter from the fixed layer of a STT-

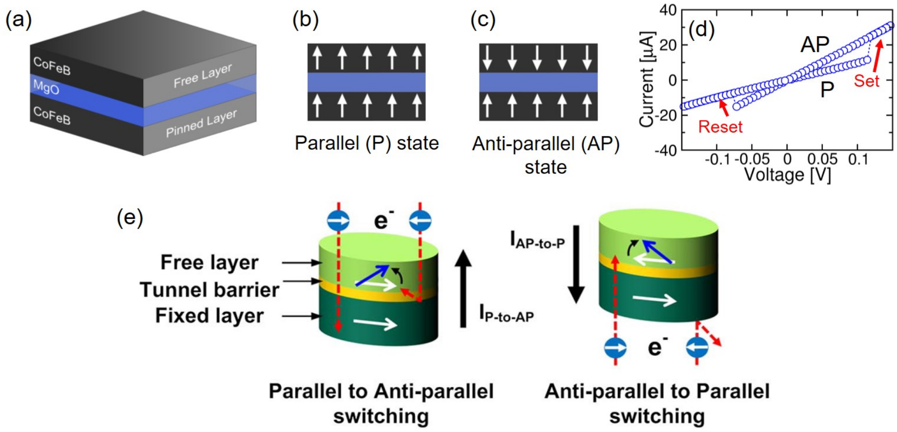


Figure 1.11: *(a) Sketch of MTJ structure based on MgO tunnel oxide sandwiched between two CoFeB-based ferromagnetic layers with a free and pinned magnetization orientation, respectively. MTJ enables two stable states due to (c) parallel and (d) anti-parallel alignment of magnetization in ferromagnetic electrodes. (d) Typical I-V characteristics for a STT-MRAM device evidencing AP-to-P (set) and P-to-AP (reset) transitions achieved by (e) a current-induced spin-transfer torque causing a rotation of free layer magnetization. (a)-(d) Copyright 2016 IEEE. Reprinted, with permission, from [72]. (e) Copyright 2014 IEEE. Reprinted, with permission, from [73].*

MRAM prepared in AP state (right), namely the applied current is forced from free to fixed layer, only the electrons with spin parallel to M$_{fixed}$ can reach free layer inducing a STT capable of exerting a rotation of M$_{free}$, eventually leading STT-MRAM state from AP to P (set transition), thus in LRS [73].

Compared to PCM and RRAM devices where to store information is needed to move atoms, the STT-MRAM storage principle based on the rotation of magnetization direction in the free layer makes STT-MRAM extremely attractive for high cycling endurance, generally referred to as virtually infinite [78]. However, even a relatively small voltage (< 1 V) can cause a sensible electrical stress to the MgO barrier capable of inducing a dielectric breakdown of thin tunnel layer, thus limiting device cycling endurance performance [72, 76, 79].

Regarding read/write operation, STT-MRAM exhibits an extremely fast access time, lower than 3 ns [80] and therefore dramatically faster than Flash memory. To improve this performance further, write currents must be downscaled and the stack should be suitably engineered to decrease the errors during writing phase due to the thermally activated nature of resistive

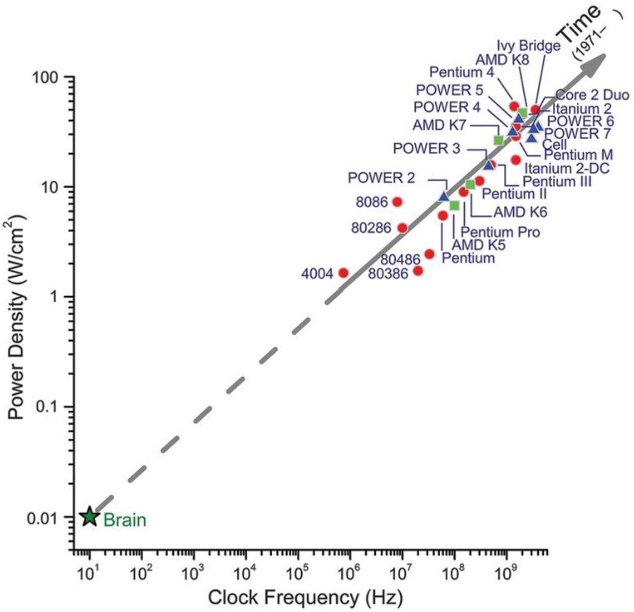


Figure 1.12: *Comparison between CMOS-based microprocessors with von-Neumann architecture built since 1971 and biological brain in terms of dissipated power density and operation frequency. Although the technology development driven by Moore's law has led to increasingly powerful processors, today's computers cannot compete with the brain efficiency in performing many complex cognitive tasks. Copyright 2014, American Association for the Advancement of Science. Reprinted, with permission, from [4].*

switching in STT-MRAM affecting data retention [76].

The great potential in terms of operation speed and endurance combined with good scaling capabilities, demonstrated down to 11 nm [81], makes STT-MRAM a promising alternative to DRAM memories. In addition, further advancements such as lower write currents and a higher TMR might even lead it to reach access times lower than embedded DRAM and thus to be used in the cache in the next future [76].

1.3 Novel approaches for beyond-CMOS computing

1.3.1 Neuromorphic computing

To tackle latency and energy burdens challenging digital computers based on conventional von Neumann architecture, various alternative paradigms have recently been explored. Among them, strong interest has been attracted by an intriguing approach referred to as neuromorphic computing taking inspiration from biological brain. Biological brain is a very complex system capable of massively parallel and error-tolerant computation thanks to its architecture based on very large networks of neurons connected by synapses (about 10^{11} neurons and 10^{15} synapses overall). Also, as shown in Fig. 1.12, it features a power consumption orders of magnitude lower than the most important clocked digital computers built so far, which stems from the event-based strategy used to process information contained in the spikes sparsely emitted by stochastic neurons [4, 82]. Therefore, the goal pursued by neuromorphic computing is to achieve complex cognitive functionalities with an energy efficiency comparable with that of biological brain mimicking its structure and fundamental mechanisms.

Neuromorphic computing by conventional CMOS technology

The path toward the building of neuromorphic systems started in the late 1980s as the scientist Carver Mead introduced pioneering advancements in bio-inspired microelectronics based on the analogy between the physics of MOSFET biased in the subthreshold region and the physical properties of biological neurons [83, 84].

Over the following decades, the application of this novel approach led to the design and building of silicon-based neuron [85, 86] and synapse circuits [87, 88], opening the way for significant hardware implementations of CMOS VLSI neuromorphic systems [89–99].

Among the large-scale hardware neuromorphic platforms have emerged in recent years [4, 100–102], the Manchester University project called SpiNNaker [100] is one of the most important. SpiNNaker is a highly distributed digital computer equipped with a custom communication framework that enables to interconnect many multi-core chips [100, 103]. Its architecture relies on the assembly of several boards including up to 48 packages, which incorporate a chip realized in 130 nm CMOS technology based on 18 ARM968 cores and a memory chip of 128 Mbyte Synchronous DRAM (SDRAM) (Fig. 1.13) [100, 103]. The SpiNNaker system enables to simulate the activity of large spiking neural networks on the biological time scale adopting many types of synapse and neuron behavioral models. However, a significant disadvantage is that the use of synapse and neuron models with increasing complexity causes a severe limitation on the network size that can be simulated in real time [6]. In addition, complete SpiNNaker operation involves a power consumption of about 50 kW which is not compatible with biological brain energy efficiency.

Another recent CMOS-based neuromorphic platform of great interest

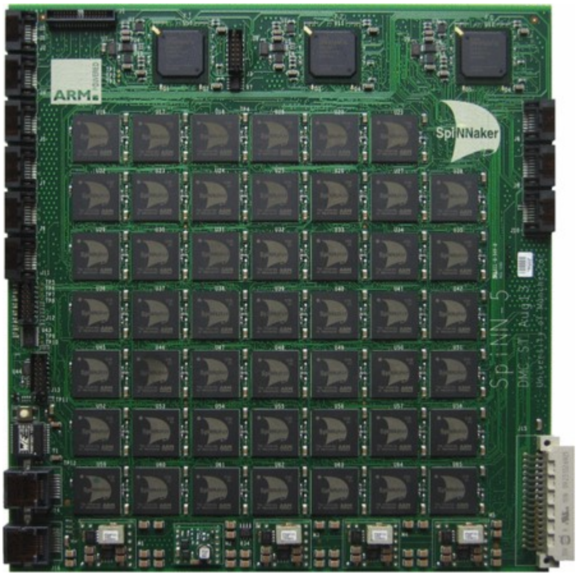


Figure 1.13: *A small-scale SpiNNaker circuit board with 48 packages. Copyright 2014 IEEE. Reprinted, with permission, from [100].*

is the IBM TrueNorth chip [4, 104]. TrueNorth is a 28 nm CMOS fully digital chip with area of 4.3 cm^2 based on 4096 neurosynaptic cores of spiking neural networks, each including 256 digital leaky integrate-and-fire neurons and 256 x 256 binary programmable synapses via a SRAM-based crossbar array (Fig. 1.14). Although the dynamics of neurons is controlled by a global clock signal at 1 kHz, the cores communicate in a fully asynchronous fashion evidencing a parallel operation driven by spike-based events delivered at synapses by firing neurons. Compared to SpiNNaker, IBM TrueNorth chip offers an extremely low power consumption (< 150 mW) and a memory distributed on the whole network, which makes this system highly parallel, modular and noise resilient. However, this high parallelism results in a relative loss of density efficiency because of all the unused synapses for a certain application. Also, a strong limitation for TrueNorth chip is that the synapses do not incorporate weight update mechanisms, thus preventing on-line learning. At the level of applications, TrueNorth chip was operated on 400-pixel-by-240-pixel video input at 30 frames per second achieving detection and recognition of multiple objects in real time with an extremely low-power dissipation of only 63 mW.

The goal of reproducing faithfully the dynamics of biological synapses in hardware led to the development of a new CMOS-based neuromorphic

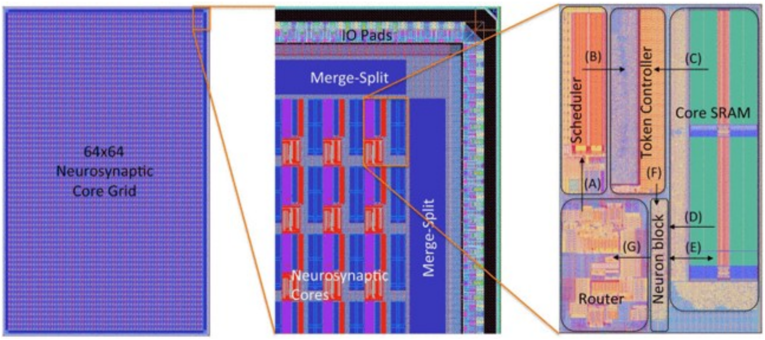


Figure 1.14: *TrueNorth chip layout consisting of a 2D grid of 64x64 neurosynaptic cores where each core includes a scheduler buffering input spikes to implement axonal delays, a token controller to manage operation in the core, a SRAM memory to store data for neurons, a time-multiplexed neuron block updating neuron internal potentials, and a router to transmit the spike events. Copyright 2015 IEEE. Reprinted, with permission, from [104].*

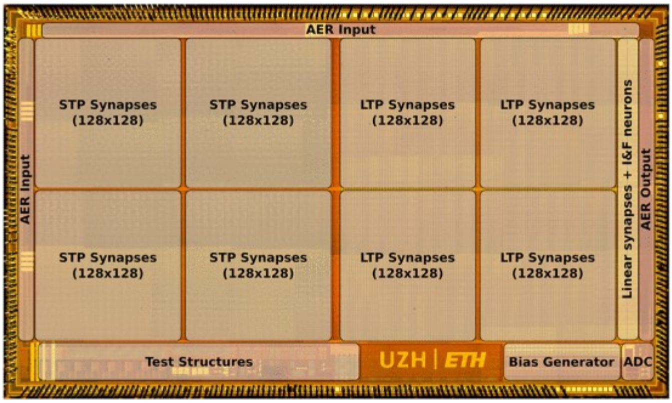


Figure 1.15: *Micrograph of ROLLS neuromorphic processor chip. To achieve memory storage and massively-distributed computation, most of chip area is designed to host non-linear synapse circuits capable of implementing short-term plasticity and long-term-plasticity. Copyright 2015 IEEE. Reprinted, with permission, from [6].*

processor system by Institute of Neuroinformatics (INI) of University of Zurich and ETH Zurich called ROLLS [6, 99]. This chip, fabricated in 180 nm CMOS technology, comprises 256 neurons and slightly more than 130.000 synapses within an area of 51.4 mm^2. The configurable spiking neural network implemented by ROLLS uses analog subthreshold circuits to mimic the real dynamics of neurons and synapses, and asynchronous digital logic circuits to control the event-based communication and configure the properties of the network as the synaptic connectivity in a flexible fashion. Fig. 1.15 shows the micrograph of ROLLS chip evidencing that chip area is mostly occupied by short-term memory and long-term memory nonlinear synapses serving as sites of memory storage and computation. In particular, the long-term memory synapses include analog bistable circuits capable of implementing learning according to a stochastic spike-based plasticity model where the arrival of pre-synaptic spikes can activate the update process leading the synaptic weight to the fully potentiated or fully depressed state based on the initial value [105]. The great computational power of ROLLS chip was demonstrated by the hardware implementation of attractor networks capable of associative memory thanks to their recurrent synaptic connectivity, and two-layer spiking neural networks for image classification tasks [99].

Neuromorphic computing by memristive devices

Despite great complexity and high performance, state-of-art fully CMOS neuromorphic hardware is not still competitive with the human brain in terms of integration density because of the area-expensive neuron and synapse circuits [106]. Also, power dissipation achieved in these platforms is much higher than human brain consumption, which is estimated around 20 W [82, 107].

To overcome these limitations, the research on nanoscale materials has evidenced that emerging non-volatile device technologies such as RRAM and PCM feature characteristics making them suitable for the implementation of synapses in neuromorphic circuits [39, 107–110]. First, RRAM and PCM devices can enable non-volatile storage of multiple resistance states in a very compact area. Moreover, their tunable resistance can be exploited to replicate in hardware bio-realistic synaptic plasticity rules such as the spike-timing dependent plasticity (STDP).

STDP rule, which was experimentally observed for the first time in cultured hippocampal neurons in the late 1990s [111], describes the modulation of synapse weight or efficacy as a function of the relative time delay between the spikes emitted by the pre-synaptic neuron and the post-synaptic neuron, respectively. If pre-synaptic neuron fires just before the fire emitted by the post-synaptic neuron ($\Delta t > 0$), the synapse weight increases resulting in the synaptic long-term potentiation (LTP). Conversely, if the pre-synaptic neuron fires just after the post-synaptic neuron ($\Delta t < 0$), the synapse weight decreases resulting in the synaptic long-term depression (LTD). Also, these experimental data show that the extent of synaptic weight change is maxi-

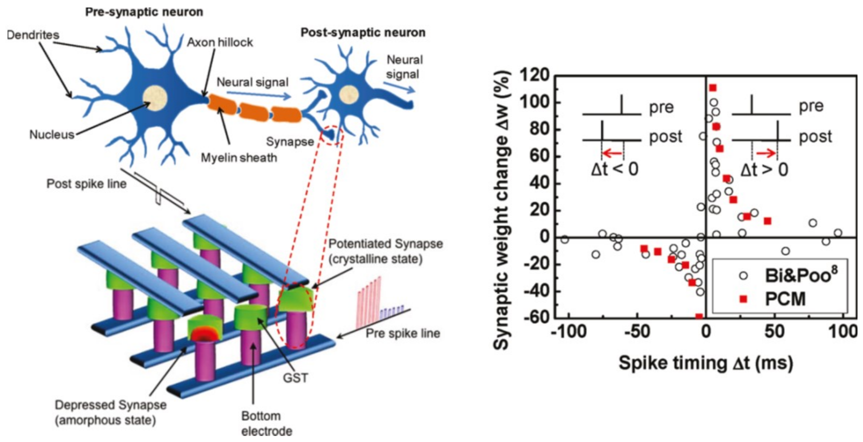


Figure 1.16: *(left) Hardware implementation of the neural network building block comprising pre-synaptic neuron (PRE), synapse and post-synaptic neuron (POST) by a memristor-based crossbar structure where PCM devices operate as synapses while bottom/top electrode lines serve as PRE/POST spike lines. (right) PRE and POST spike overlapping across PCM synaptic devices for variable time delays results in a STDP hardware implementation evidencing a nice agreement with original biological data. Copyright 2011 American Chemical Society. Adapted, with permission, from [113].*

mum for very short time delays and then decreases rapidly with increasing time delay [111, 112].

In last 10 years, STDP rule was demonstrated in individual RRAM [114, 115] and PCM [113] devices (Fig. 1.16), and in hybrid CMOS/memristive structures such as the one-transistor/one-resistor (1T1R) structure [116–118] and the 2-transistor/1-resistor (2T1R) structure [119,120] using schemes based on overlapping spikes at synapse terminals. Note that a more complex hybrid CMOS/memristive synapse circuit that does not adopt an overlap scheme to implement STDP and limits the integrated current in the post-synaptic neuron was also recently proposed [121].

Moving from device to network level, many memristive spiking neural networks capable of capturing cognitive abilities such as unsupervised learning and recognition of visual or auditory patterns via STDP rule have been implemented in simulation [116, 121–128]. Nevertheless, only implementations of spiking neural networks with a limited number of memristive synapses capable of STDP have been achieved in hardware up to date [118, 129–132].

Although STDP has received considerable attention as main mechanism underlying synaptic plasticity, biological experiments in the early 2000s evidenced that spike rate plays a key role in controlling plasticity of real synapses [133]. This thus led to the exploration of another synaptic plasticity mechanism called spike-rate dependent plasticity (SRDP) which was connected to spike triplets [134–136]. As a result, new synapse models incorporating weight update based on spike rate have been implemented in both simulation and hardware [137–144].

However, STDP and SRDP do not complete the wide range of biological mechanisms controlling synaptic plasticity in human brain. The goal of replicating more faithfully the synaptic dynamics has thus stimulated the investigation of memristive devices based on materials such as Ag_2S [145] and SiO_xN_y:Ag [146] which, thanks to their volatile resistive switching mechanism, have enabled to achieve the hardware implementation of an additional effect displayed by biological synapses called short-term plasticity.

In addition, various memristive devices have been investigated to replicate not only synaptic behavior but also the neuron behavior [132, 147–149].

Specifically, as evidenced in Fig. 1.17, a PCM device can be used to implement the neuronal integration (Fig. 1.17(a)) matching the increase of membrane potential triggered by incoming spikes with the increase of crystalline phase of chalcogenide active material (Fig. 1.17(b)) [149]. Compared to the use of conventional bulky CMOS-based integrate-and-fire neuron circuits, this approach could lead to an additional improvement of the device density in hardware neuromorphic systems with memristive devices.

Although neuromorphic computing is strongly linked to the biological world, important demonstrations of complex cognitive capabilities have also been achieved using schemes different from brain-inspired unsupervised learning such as the supervised learning based on backpropagation rule [150–152].

In this frame, motivated by the desire to achieve the very high performance of artificial intelligence in performing machine-learning tasks such as image recognition [152, 153], speech recognition [154], translation of sentences [155], natural language processing [156, 157] and other complex applications [158–160] via software-based deep neural networks (DNNs), valuable hardware demonstrations with memristive devices have been achieved, including electroencephalography pattern recognition by a cross-point RRAM synapse array [161], face recognition by an 1T1R RRAM array [162] and image classification by multi-layer fully connected artificial neural network (ANN) with 1T1R RRAM [163] and 1T1R PCM [164, 165] synaptic devices.

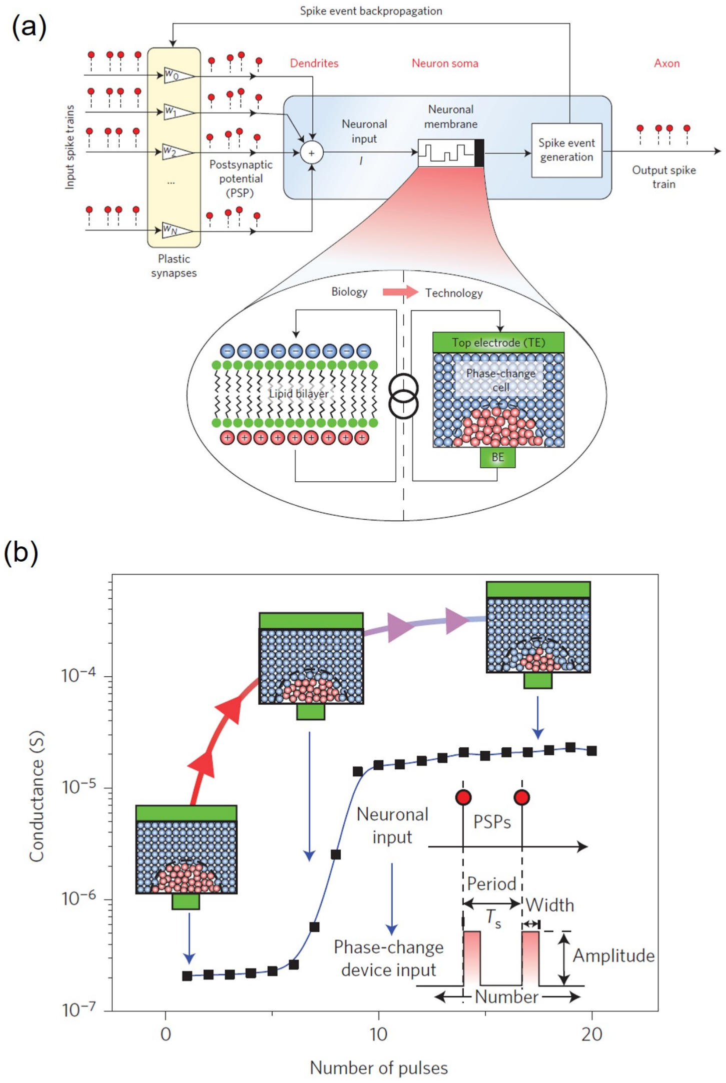


Figure 1.17: *(a) Schematic illustration of a PCM-based neuron which (b) stores its membrane potential due to integration of input spikes in the crystalline phase configuration of PCM cell. Copyright 2016, Springer Nature. Adapted, with permission, from [149].*

Fig. 1.18(a) shows the scheme of the 3-layer fully connected ANN used in [164, 165] to demonstrate image classification task on MNIST handwritten digit dataset [151, 166]. In this feedforward network architecture, each layer includes software-based neurons implementing a nonlinear activation function and drives the next layer by weights w_{ij} implemented in hardware

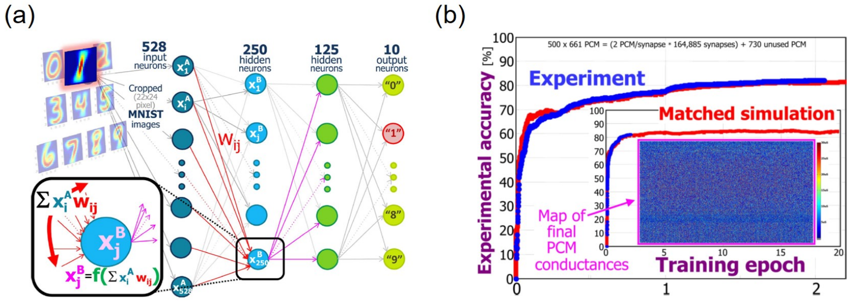


Figure 1.18: *(a) Schematic illustration of a feedforward fully-connected deep neural network (DNN) used to demonstrate supervised learning and classification of handwritten digit images from MNIST database. (b) Measured and calculated accuracy performance for both training and classification phase achieved by 3-layer DNN with PCM-based synaptic weights. Copyright 2014 IEEE. Adapted, with permission, from [164].*

by an array of 1T1R 2-PCM synapses. During training, such weights are adjusted in response to the submission of 5000 images, which represent the training dataset used in this experiment, according to an experimental overlapping pulse scheme replicating backpropagation algorithm. Backpropagation rule is a well-known computational algorithm that allows to optimize the learning efficiency of the neural network on a large image dataset. The submission of an input pattern to the network results in an output signal, whose distance from the ideal value is called error signal. After calculating the error signal at the output stage, it is backward propagated from the output stage to the input layer leading to the update of all the synaptic weights of an amount directly proportional to the product of the input and the error signals [150–152]. After training, neural network classification capability was tested by the presentation of new handwritten digit images to the input layer. Fig. 1.18(b) shows the measured learning and test accuracies with corresponding simulations evidencing a maximum accuracy around 83% which is due to the detrimental impact of PCM asymmetry and nonlinearity during weight update process.

To overcome this strong limitation due to PCM non-idealities, a new synaptic "2PCM + 3T1C" unit cell, which combines a pair of PCM devices, each with own transistor serving as selector, with a 3-transistor/1-capacitor analogue conductance device implementing volatile weight storage, was recently proposed in [167] enabling to experimentally achieve classification accuracy values equivalent to those obtained by software-based DNNs not only for MNIST dataset, but also for CIFAR-10 and CIFAR-100 datasets

[168]. Most importanly, the efficiency and throughput performance demonstrated by mixed hardware-software DNNs based on these analogue memory devices are 280 and 100 times better, respectively, than the corresponding performance of recent GPUs, thus suggesting analogue memory and computing devices as key elements for low-power acceleration of training in hardware DNNs [167].

In summary, the co-integration of memristive devices such as RRAM and PCM and CMOS-based circuits is a promising solution to design brain-inspired neuromorphic systems with enhanced energy efficiency and integration density. Although the objective to achieve the massive parallelism and the outstanding computational abilities of biological brain is still far and many issues at both device and system level such as the device reliability and the integration strategy need to be still addressed, the great potential of neuromorphic computing approach and the future findings in neuroscience can pave the way for the building of truly brain-inspired computing systems.

1.3.2 In-memory computing

Carrying out calculations where the data are stored is the only strategy to totally remove the memory wall. Recent works have evidenced that this approach, known as in-memory computing [10, 22, 23, 169], can be achieved using the physics of resistive switching devices such as RRAM [169–172] and PCM [148, 173–175], and the fundamental laws of electrical circuits such as Ohm's law and Kirchhoff's laws [176]. Also, depending on the binary or gradual nature of resistive switching phenomenon, different schemes such as in-memory digital computing and in-memory analogue computing can be implemented by memristive devices [22].

In-memory digital computing by RRAM devices

The increasing appeal of emerging non-volatile memory devices for implementation of computing tasks has pushed digital computing to explore new in-memory logic gate and circuit concepts to carry out digital Boolean operations saving energy and area compared to corresponding CMOS implementations [169–173].

Among different types of resistive-switching devices, RRAM is the most suitable technology for digital computing due to its binary operation during set. Also, RRAM enables high scalability, a direct access to the cell by interconnections and the device reconfiguration via voltage pulses.

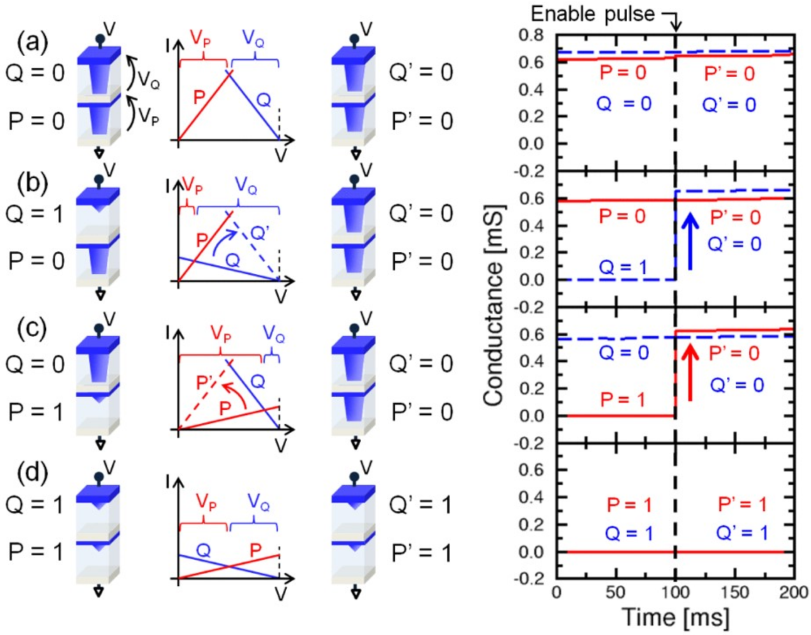


Figure 1.19: *Schematic illustration of implementation of an AND gate using a serial configuration of RRAM devices evidencing initial state for P and Q RRAM cells, I-V characteristics for both RRAM devices, final state for P and Q RRAM cells and time evolution of measured conductance before and after application of the voltage pulse (from left to right). AND logic function was demonstrated testing RRAM-based structure operation preparing devices in all the initial configurations, namely (a) Q = 0 and P = 0, (b) Q = 1 and P = 0, (c) Q = 0 and P = 1, and (d) Q = 1 and P = 1. Copyright 2015 IEEE. Reprinted, with permission, from [171].*

For example, as proposed in [171], an AND logic gate can be implemented by two serially connected RRAM devices. Moving from left to right, Fig. 1.19 schematically shows the initial configuration of two devices called Q (top RRAM) and P (bottom RRAM), respectively, the I-V operation characteristics, the final state and the corresponding experimental demonstration covering all combinations from the case where both devices are prepared in LRS (a) to the case where both devices are prepared in HRS (d). The logic operation is driven biasing the devices by the application of a driving voltage V at the top electrode of the structure with the intermediate electrode left floating and the bottom electrode at ground. Specifically, because V can have values between 2V$_C$ and 2V$_{set}$, it is always positive and, as a consequence, only abrupt set transitions can occur in each resistive cell.

As evidenced in Fig. 1.19(a), if both Q and P are initialized in LRS (state 0), no switching event occurs in the devices since they are already in LRS. Thus, the final state is (Q', P') = (Q, P) = (0, 0). Unlike previous case, if Q and P are prepared in HRS (state 1) and LRS, respectively, *i.e.*, (Q, P) = (1, 0), the applied voltage V mostly drops across Q inducing a set transition leading to a final state (Q', P') = (Q, P) = (0, 0) with both devices in LRS (Fig. 1.19(b)).

Fig. 1.19(c) instead shows the case with Q and P prepared in LRS and HRS, respectively, i.e., (Q, P) = (0, 1). Because P cell is initially in HRS, the V drops almost totally across P causing a set transition for P. As a result, (Q', P') = (Q, P) = (0, 0).

Finally, in the case shown in Fig. 1.19(d) where both Q and P are initialized in HRS, *i.e.*, (Q, P) = (1, 1) , V divides equally between two resistive switches, thus preventing any switching event in Q and P cells, resulting in a final state (Q', P') = (Q, P) = (1, 1), since the voltage drop across each device is lower than V$_{set}$.

In summary, the output states (Q', P') show that this circuit based on serial connection of two RRAM switches implements an AND logic function of input states (Q, P), thus supporting RRAM as viable technology for in-memory digital computing.

In-memory analogue computing by PCM devices

While in-memory digital computing takes advantage of binary resistive switching, a gradual resistance change is instrumental in capturing other concepts such as analogue computing [22].

A fundamental example of in-memory analogue computing is the implementation of the arithmetic summation in a PCM device thanks to the gradual crystallization of amorphous chalcogenide material [148, 149]. As shown in Fig. 1.20, to carry out a sum operation in base-10, the cell is first prepared in a fully amorphous HRS configuration (R = 500 kΩ) by reset transition and 10 consecutive pulses are needed to achieve the partial crystallization corresponding to full LRS. Specifically, to perform the sum 1+3, the application of 1 pulse (state-1) is followed by the application of 3 sequential pulses with the same amplitude and width. As a result, the PCM cell reaches the resistance state achievable by the application of 4 consecutive pulses (state-4), thus simultaneously obtaining computation and storage of the sum result [148]. Therefore, this task supports PCM device as in-memory accumulator capable of tackling the high area consumption of CMOS implementations.

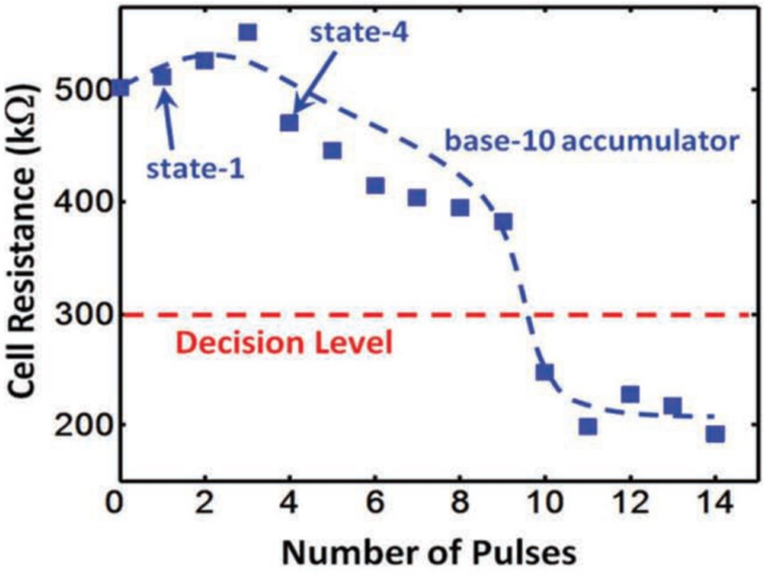


Figure 1.20: *Measured resistance response of a base-10 PCM-based accumulator. Starting from the cell in high-resistance amorphous configuration, the application of 1 pulse followed by the application of additional 3 pulses leads the PCM device from state-1 to state-4 via the gradual chalcogenide crystallization process, thus enabling the implementation of the arithmetic sum. Copyright 2012, John Wiley and Sons. Reprinted, with permission, from [148].*

In-memory analogue computing by crossbar arrays

In addition to the use of physics of memristive devices, in-memory computing can be achieved exploiting the fundamental laws of electrical circuits such as Ohm's law and Kirchhoff's laws [22, 23]. This approach is particularly suitable for the analogue implementation of the matrix-vector-multiplication (MVM) operation by crossbar arrays.

Fig. 1.21 shows the architecture of a crossbar array where each intersection or crosspoint between orthogonal column and row electrodes hosts a memristive device. Based on this architecture, the application of a voltage V_j at each j-th column induces a current $V_j \cdot G_{ij}$, where G_{ij} is the conductance of memristor device located at crosspoint between the i-th row and j-th column. The sum of all currents activated at each i-th row results in a total current $I_i = \sum_j G_{ij} \cdot V_j$, thus evidencing the analogue implementation of MVM operation simply using the Ohm's law and Kirchhoff's law. Most importantly, crossbar architecture enables MVM in only one opera-

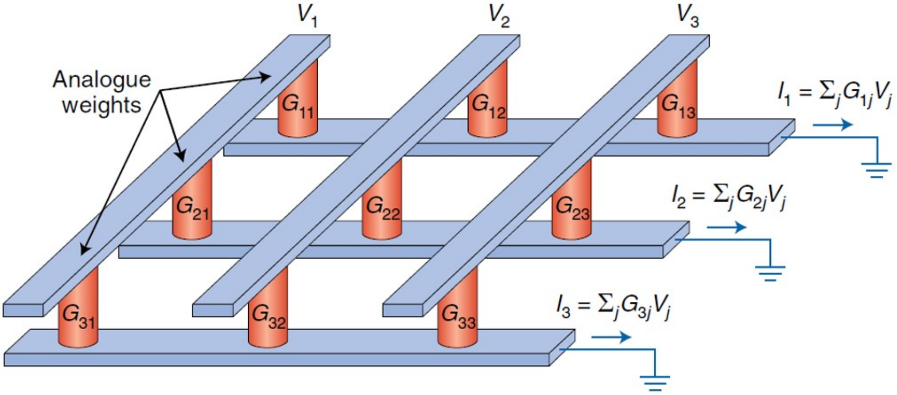


Figure 1.21: *Sketch of a 3x3 crossbar array enabling the physical implementation of a matrix-vector multiplication (MVM) operation by Ohm's law and Kirchhoff's law for in-memory analogue computing. Copyright 2018, Springer Nature. Reprinted, with permission, from [22].*

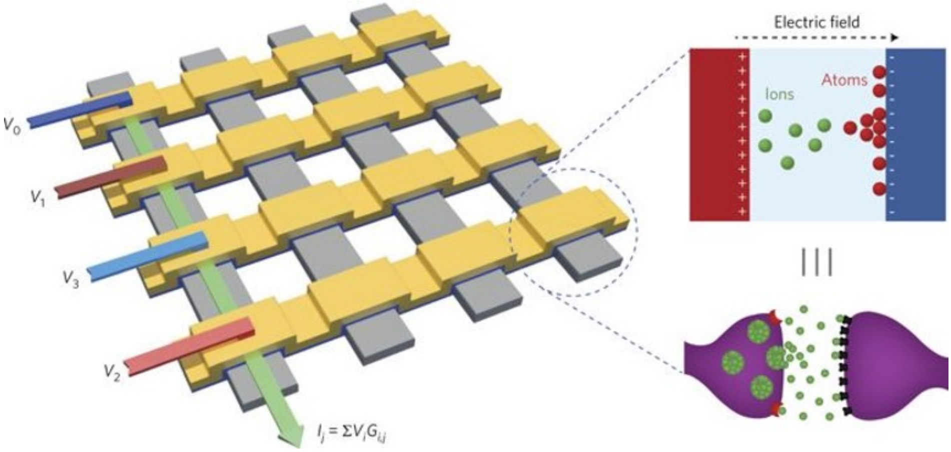


Figure 1.22: *Artificial neural network implemented by a crossbar array based on memristive devices capable of emulating biological synapses. Copyright 2018, Springer Nature. Reprinted, with permission, from [23].*

tion step, thus decreasing time and energy costs compared to digital MAC operation carried out in current processors. In addition, thanks to MVM accelerated computation, crossbar architecture has attracted strong interest for hardware implementation of high density ANNs with memristive devices serving as synaptic weights [23], as shown in Fig. 1.22, and image processing [177].

Issues and challenges for in-memory computing

Despite in-memory computing features great potential to overcome von-Neumann bottleneck, many issues have yet to be addressed [22].

First, the inherent variability of resistive switching mechanism such as the cycle to cycle and device to device variability of V_{set}, and the memory instability are two limiting factors toward RRAM-based implementation of reliable in-memory digital logic gates. While memory instability mostly burdens RRAM devices due to the extreme sensitivity of LRS and HRS from displacement of single atoms close to the conductive filament, PCM is mostly affected by drift and strong non-linearity of HRS resistance. As a result, crosspoint array computation becomes limited only to error-resilient applications such as pattern recognition [22].

The future development of in-memory computing is closely related to the technology scalability. To achieve higher density is first needed to down-scale the computing element dimension, which however involves an undesired increase of cell-to-cell variability. In addition, as a result of cell miniaturization, interconnection lines have also to down-scale, inducing an increasing series resistance and consequently high parasitic voltage drops [22].

Such issues resulting from in-plane scaling could be successfully solved by the hardware implementation of 3D-crossbar arrays since they allow to increase device density by multi-layer stacking, thus avoiding the concerns due to the cell miniaturization. Also, to further increase device density, the distance between two adjacent cells could be reduced by decreasing the thickness of memristor switching layer, which makes RRAM more suitable than PCM as computing element for crosspoint array because of its very thin switching layer [22].

In summary, in-memory computing with memristive devices exhibits a very strong potential in terms of energy and computational efficiency thanks to its peculiar feature to enable calculations in situ. However, significant challenges including the improvement of device performances such as endurance, variability and power consumption, and the building of efficient 3D crosspoint arrays are issues to be addressed in order to achieve an extensive use of this approach. Therefore, major efforts at research and industrial level are still needed before this paradigm enables to overcome the memory wall and consequently becomes a feasible technology for beyond-CMOS computing.

1.3.3 Quantum computing

Since the Nobel Laureate physicist Richard Feynman began to speculate on the possibility to build a new type of computer to efficiently simulate quantum physics in 1982 [178], the prospect of achieving a quantum computer, that is a computing machine capable of storing, processing and transmitting information using quantum mechanical phenomena, has attracted growing interest from both academic community and industry.

Unlike the traditional computers where the information is encoded into bits, which can only be in either of two states, namely 0 or 1, quantum computers store information in quantum bits (qubits) which can be either in the basis states $|0\rangle$ and $|1\rangle$, or in any pure state given by their quantum superposition, namely $|\psi\rangle = \alpha|0\rangle+\beta|1\rangle$ where α and β are two complex numbers representing the probability amplitudes such that $|\alpha|^2 + |\beta|^2 = 1$. Therefore, the building of a quantum computer with an internal memory based on N qubits may in principle allow to store information in the quantum superposition of 2^N basis states described by 2^N complex coefficients. As a result, a quantum computer may store an amount of information exponentially larger than the classical counterpart, which would make it dramatically more powerful from the computational viewpoint at least for certain applications [179–181]. Among such tasks, the best-known examples are the prime factorization of large integers, which can be exponentially accelerated using the Shor's polynomial-time quantum algorithm [182], and the database search, which can be carried out with a quadratic speedup applying the Grover's quantum algorithm [183].

To understand how classical and quantum computers work, Fig. 1.23 shows a comparison between the two computation strategies. In classical computers, starting from all bits prepared in the state 0 with unity probability (a), computation leads to changes in the memory state via Boolean logic operations (b), until reaching a new state a with unity probability (c).

On the other hand, in quantum computers, starting from all qubits initialized in the basis state $|0\rangle$ with unity probability, which involves a delta-like probability density distribution in the $|0\rangle$ state (d), the quantum state defined by a wavefunction $|\psi\rangle$ evolves with time according to the Schrödinger equation (e) until the probability density distribution given by $|\psi|^2$ centers around a new state $|a\rangle$ (f). As a result, whereas classical computation always provides a deterministic output, quantum computation only gives a probabilistic output [180].

Capturing the enormous potential in terms of storage capacity predicted for quantum processors is not however easy since preserving and maintain-

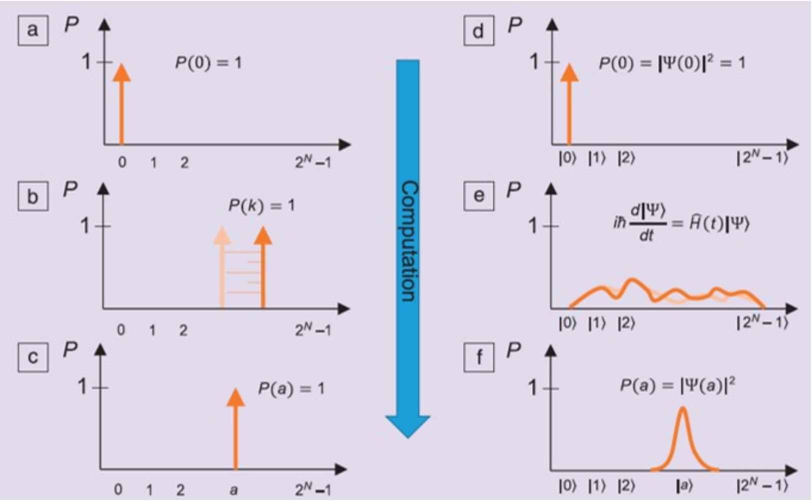
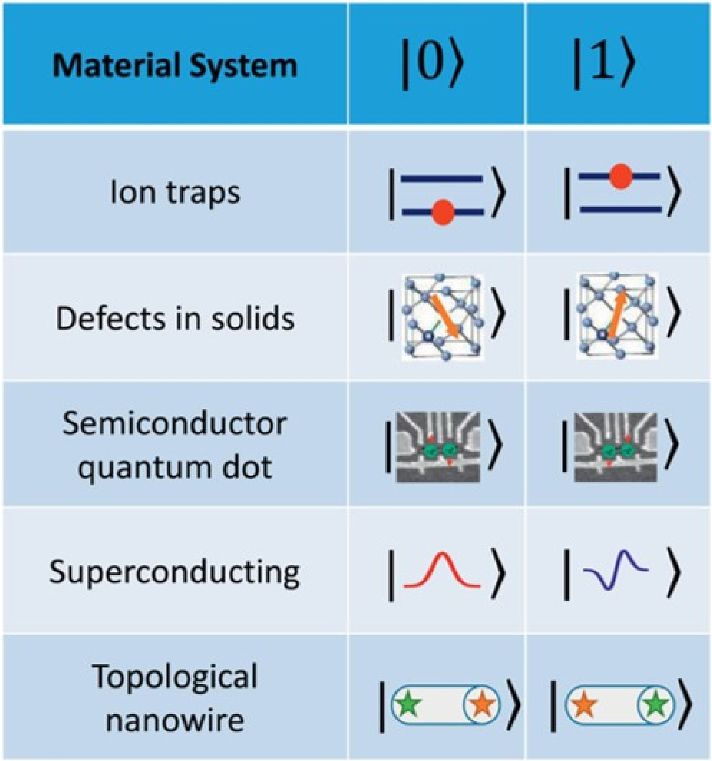


Figure 1.23: *Comparison between (a-c) classical and (d-f) quantum computating approaches. In classical computers, starting from all bits set to the state 0 (a), the memory state follows a deterministic evolution driven by Boolean logic operations (b) achieving another state a with unit probability (c). In contrast, in quantum computers, the memory state initially prepared in $|0\rangle$ state (d) follows a probabilistic evolution according to Schrödinger equation (e) achieving a final state $|a\rangle$ with a certain probability distribution. Copyright 2013 Materials Research Society. Reprinted, with permission, from [180].*

ing information in the qubits is strongly challenged by two fundamental physical processes.

As a qubit typically consists of a two-level quantum system, quantum information contained in a qubit is first degraded by the decay from the excited state to the ground state due to the energy relaxation mechanism being governed by a timescale T_1. In addition, it is also affected by a second process, referred to as quantum decoherence, consisting of the unwanted interaction of qubit with its surrounding physical environment resulting in the degradation of the qubit phase over a timescale T_2, which is lower than T_1 in the most cases. Consequently, decoherence has a more detrimental impact than energy relaxation on quantum computation, which can be mitigated by the application of schemes referred to as quantum error correction (QEC) techniques [181].

In particular, note that the effect of noise on computation marks an important distinction between quantum computing and brain-inspired neuromorphic computing. Whereas the inevitable interaction of qubits with their physical environment is the main cause of loss of coherence and thus of



Material System	$\|0\rangle$	$\|1\rangle$
Ion traps		
Defects in solids		
Semiconductor quantum dot		
Superconducting		
Topological nanowire		

Figure 1.24: *List of 5 different material systems to make qubits in hardware. Copyright 2013 Materials Research Society. Reprinted, with permission, from [180].*

information for quantum computation, the presence of noise was proven to play a beneficial role in brain-inspired neuromorphic computing enabling to capture increasingly complex cognitive primitives [184]. Therefore, although the quantum computing promises to capture a computational power even higher than that of the biological brain using quantum parallelism and entanglement, the demanding technological requirements to achieve extremely low levels of noise have considerably hindered the development of practical quantum computing systems with respect to that of neuromorphic systems in these years. To overcome this obstacle, the novel concept of quantum neuromorphic computing has recently been proposed and is currently under investigation [185, 186].

Regarding the building of the qubits, several material systems have been explored. In this framework, Fig. 1.24 evidences five different engineering solutions receiving great interest [180].

A first approach to make qubits consists of the electromagnetic confinement of single ions in harmonic potentials (ion traps) at extremely low temperature by application of voltages to the electrodes close to ions. The state of a qubit based on trapped ion is given by combination of two hyperfine states used as $|0\rangle$ and $|1\rangle$, respectively [180], and is measured thanks to the photon generated by the ion. To achieve a quantum computer based on trapped ion technology, the project is to exploit the entanglement of multiple ions to create scalable networks of quantum nodes communicating by emitted photons. In these years, the intensive research on this technology has led to significant advances [187] that culminated with the recent achievement of two-qubit logic gates, namely fundamental blocks for quantum computing, capable of combining the state-of-art level of precision or fidelity (99.8 %) with a new record of speed (1.6 μs) [188].

A second manner to achieve qubits relies on the use of defects in solids. In particular, great interest has been gained by the nitrogen-vacancy (NV) centers in diamond lattice, namely defects formed by a missing carbon-atom close to a nitrogen atom replacing one carbon atom [189]. This type of qubit system exploits the spin degrees of freedom of the NV center associated with its bound electrons and near nuclear spins [189]. Its great appeal origins from atomic-like properties such as robust spin quantum states and precise optical transitions combined with the solid-state structure, which enables a fast electrical and magnetic control [189]. Most importantly, this technology has the unique feature to achieve long coherence times even at room temperature [180,189,190]. Specifically, coherence times longer than 1 s at room temperature for nuclear spins were reported in [191]. Therefore, the solid-state structure, which can also be exploited to design large-scale systems, and room temperature operation make this scheme promising for the building of commercially-available quantum computing platforms.

Another attractive solid-state material system to make qubits is provided by quantum dots, namely nanoscale semiconductor regions realizing the confinement of single electrons in potential wells. Specifically, they capture qubit functionality by exploiting the spin degree of freedom of the electrons confined within semiconductor structures such as GaAs/AlGaAs system of Group III-V and Si/SiGe system of Group IV. This approach promises great advantages in terms of scalability and compatibility with conventional integrated circuits even though significant improvements in material engineering such as the growth and deposition of oxides at ultra-low temperatures are crucial to make it a competitive technology for development of commercial quantum computing machines [192].

In addition to previous material systems, academic and industry research has also focused on the building of qubits, called superconducting qubits, by electronic circuits including capacitors, inductors, interconnections and mainly Josephson tunnel junctions, namely devices based on a very thin ($\approx$ 1 nm) insulating layer interposed between two supeconducting electrodes, made via lithography and cooled at ultra-low temperatures of the order of 20 mK. At these temperatures, such circuits exhibit a behavior similar to that of the quantum oscillators evidencing discrete quantum states. Among these, the first two levels at lowest energies typically are those that effectively form a superconducting qubit. Also, depending on the elements used within the superconducting circuits, three different types of superconducting qubits, referred to as charge qubit, phase qubit, flux qubit, respectively, can be implemented [193]. This approach shows interesting features such as the great prospectives for scalability by use of existing lithographic techniques, increasing coherence times and control at microwave frequencies. However, fundamental challenges such as device operation at ultra-low temperature for large-scale systems, spourious cross-coupling among many qubits and the imperfections in fabrication process of circuits will have to be suitably addressed in near future so that superconducting qubits meet demanding requirements for implementation of quantum computing machines working not only in the laboratories [194].

Finally, semiconductor nanowires have also been explored to build qubits. In this frame, as reported in [195], when a one-dimensional semiconductor nanowire exhibiting high spin-orbit coupling is combined with a proper magnetic field and the proximity effect of a superconductor, a curious particle called Majorana fermion can be observed. Starting from these experiments, topological qubits based on the location of Majorana fermions were demonstrated [196]. The great advantage of this approach consists of tackling the loss of information due to the interactions between qubit and local field by exploiting the delocalization of quantum state [180].

With the rapid progress of technology, very important companies in the field of information technology have announced the building of their first quantum computer prototypes in the last years. After realizing a first quantum machine based on five superconducting qubits called IBM Quantum Experience in May 2016 [197], which was the first quantum computer usable by the public via cloud, in November 2017 IBM announced the development of both a new commercial quantum computer based on 20 superconducting qubits and a first prototype of 50 qubit quantum computer [198]. In addition to IBM, in early 2018, Intel unveiled his 49-qubit superconducting quantum-processor chip called Tangle Lake [199] and the Google's Quantum Artificial Intelligence Lab announced a new 72-qubit superconducting quantum computing chip called Bristlecone [200].

Overall, quantum computing is a fascinating computing paradigm being expected to introduce a dramatic acceleration in solving problems too computationally expensive in terms of time and cost for today's computers in various fields including the simulation of complex quantum systems investigated in physical chemistry and materials science, the discovery of new life-saving drugs, optimization of large systems such as transportation routes, machine learning and finance [201, 202]. In addition, it is believed that it could lead to a deep revolution in the field of data security in that the encryption schemes currently used in many activities such as private messaging and banking could be potentially broken by future quantum algorithms [203].

Despite the great advances recently made in the exploration of new approaches at the level of material systems, the implementation of large-scale practical applications with commercially available quantum computers is however far from being achieved shortly. To solve problems computationally intractable for classical computers by exploiting unique phenomena such as quantum entanglement and superposition, future quantum computers will have to indeed use over thousands of highly interconnected qubits operating with very low error rates against the some tens of qubits used in current implementations [204]. The extremely high level of precision with which the qubits have to be created, manipulated, and measured in order to achieve a reliable computation thus imposes remarkable technical challenges [185]. In particular, much attention has been focused on the need to strongly limit the errors arising from spurious interactions among the qubits and to optimize the qubit connectivity. To achieve these goals, systems based on trapped ions seem the leading technology thanks to their superior precision and high connectivity yields [201]. Also, another critical challenge is the fact that very low error levels can be typically achieved in most material systems only keeping qubits at ultra-low temperatures of the order of millikelvins [204]. To tackle this additional issue, an interesting solution being recently under investigation of research community could be the use of qubits based on NV centers in diamond system which offer the great advantage to work at room temperature and to be natural light emitters, which facilitates the measurement process [201]. In addition, diamond system exhibits a solid-state crystal structure which could be exploited in order to overcome another crucial challenge as scalability of the technology by adapting nanofabrication techniques currently used in semiconductor industry to the development of integrated quantum devices [189]. Unfortunately, the current manifacturing methods of NV centers require significant improvements to realize enough defects in a reliable manner [194].

In the end, quantum computing is an emerging technology approach evidencing a disruptive potential which is attracting increasing attention and investments by leader companies in information technology and governments in the world. Nevertheless, its development is still in the early stages and it will be many years before universal quantum computers are available on the market and thus a new information revolution begins.

1.4 Conclusions

This chapter provides an overview of main solutions at the level of device, circuit and system receiving much attention to solve the fundamental issues currently challenging performance improvement of today's computers. After discussion of the solutions for the scaling issues and heat wall, novel approaches to remove memory wall have also been presented. On the one hand, the great potential of emerging non-volatile memory devices has been described reviewing physical mechanisms and key characteristics of three fundamental device technologies such as RRAM, PCM and STT-MRAM. On the other hand, new computing schemes aiming to overcome limitations of von Neumann architecture such as neuromorphic computing and in-memory computing have been extensively described, showing that the use of memristive devices can play a key role to achieve future computing beyond standard CMOS technology. Finally, another attractive paradigm for future computing, namely quantum computing, has also been presented discussing both its impressive potential in tackling computational problems extremely hard for conventional processors and the concerns slowing down the market entry of a truly universal quantum computer.

CHAPTER 2

Physics-based modeling of HfO_2 RRAM devices

2.1 Introduction

The semiconductor industry is currently challenged by the emergence of Internet of Things, Big data, and deep-learning techniques to enable object recognition and inference in portable computers. These revolutions demand new technologies for memory and computation going beyond the standard CMOS-based platform.

In this scenario, resistive switching memory (RRAM) is extremely promising in the frame of storage technology, memory devices, and in-memory computing circuits, such as memristive logic or neuromorphic machines. To serve as enabling technology for these new fields, however, there is still a lack of industrial tools to predict the device behavior under certain operation schemes and to allow for optimization of the device properties based on materials and stack engineering.

To address this strong limitation, various types of computational models have been developed across the whole hierarchy of materials-level atomistic simulations, device simulation, and compact models for exploring RRAM applications in memory and computing [205].

Among these modeling approaches, finite element method (FEM) numerical models have raised a strong interest for their ability to grasp the switching mechanisms at the device scale (few tens of nm^3) [206–208]. Specifically, these simulation models have the added value of providing a direct output in the form of calculated current-voltage characteristics, or calculated response to applied pulses.

Most importantly, these numerical simulations allow to visualize the local dynamics of defect concentration leading to set/reset processes, thus providing the basis for the development of compact models consisting of a simplified set of analytical equations for microscopic parameters, such as the conductive filament (CF) diameter, the gap length and the local temperature [51,66,209–212].

Compact models are essential tools for circuit simulations, to anticipate the demonstration of storage/computing concepts [211,212], thus supporting RRAM in various application frameworks to strengthen the short-term impact on the market and industry evolution.

This chapter, which is based on the works [66, 127, 205, 206], covers an extensive description of various physics-based models of HfO$_2$ RRAM devices. A previous numerical model capable of a detailed understanding of the switching mechanisms and a previous analytical model are first reviewed. Finally, a stochastic simulator of set/reset statistical variability in HfO$_2$ RRAM devices, which provides a variability-aware framework for the design and simulation of neuromorphic circuits, is also described.

2.2 Numerical model for HfO$_2$ RRAM devices

Numerical FEM models can provide an accurate microscopic understanding of the switching dynamics in RRAM devices, while accurately describing the current-voltage characteristics (I-V characteristics) and the time evolution of the device [206–208].

The FEM simulation of RRAM device consists of the self-consistent solution of three fundamental partial differential equations, namely (i) the carrier continuity equation for electronic conduction, (ii) the steady-state Fourier equation of heat transport, and (iii) the drift/diffusion continuity equation of ionized defects, which describes the ion migration processes at the origin of the set and reset transitions [206].

Fig. 2.1 shows the simulated geometry used in [206], consisting of the initial configuration of the RRAM device in the set state. Here, a continuous CF with an ideal cylindrical shape connects the top electrode (TE) and the

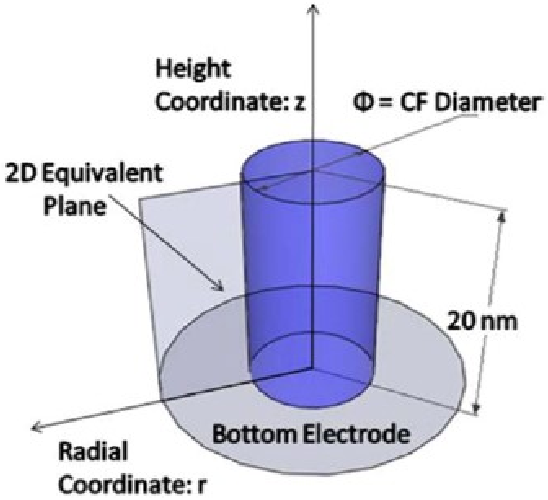


Figure 2.1: *Schematic illustration of simulated geometry used in the numerical model to describe the set state of HfO$_2$ RRAM device. Copyright 2012 IEEE. Reprinted, with permission, from [206].*

bottom electrode (BE). In the simulation, the CF consists of a region with enhanced concentration of defects such as oxygen vacancies and metallic impurities, thus causing a low resistance path between the two electrodes. In particular, the switching layer, which will be assumed to be HfO$_2$ if otherwise noted, is assumed to have a 20 nm thickness.

In this numerical model, ion migration is described as a result of the combined effects of diffusion and drift forces, according to the ionic hopping phenomenon. Diffusion consists of the random hopping of defects along potential wells separated by an average energy barrier E$_A$ (Fig. 2.2(a)). On the other hand, the presence of an applied electric field can induce drift, because of the energy barrier lowering by a factor αqV, where V is the applied voltage, in the field direction (Fig. 2.2(b)). Drift predominates in set and reset transitions, where a strong field is applied to induce a fast-directional ion migration and change the resistance.

Combining the diffusion flux j$_{diff}$ and the drift flux j$_{drift}$, the total ion flux j$_D$ is given by:

$$j_D = j_{diff} + j_{drift} = -D\nabla n_D + \mu F n_D, \tag{2.1}$$

where D is the ion diffusivity, n$_D$ is the ionized defect concentration, μ is the ion mobility and F is the applied electric field.

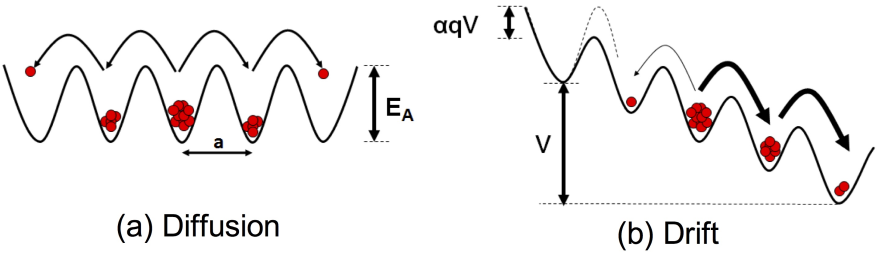


Figure 2.2: *Schematic illustration of physical mechanisms controlling hopping-based migration of ionized defects in bipolar RRAM. (a) Ionic diffusion is driven by temperature and concentration gradient, while (b) ionic drift is driven by the electric field. Copyright 2012 IEEE. Reprinted, with permission, from [206].*

Note that ion diffusivity is temperature activated according to the Arrhenius law, namely:

$$D = D_0 e^{-\frac{E_A}{kT}}, \tag{2.2}$$

where D$_0$ is a pre-exponential factor, k is the Boltzmann constant, E$_A$ is the energy barrier for hopping transport in Fig. 2.2(a), and T is the temperature.

In addition, ion mobility μ depends on ion diffusivity D according to the equation:

$$\mu = \frac{qD}{kT}, \tag{2.3}$$

which is known as Einstein relation.

The drift-diffusion ionic continuity equation $\nabla j_D = 0$ must then be solved with the Poisson continuity equation for electron current, which yields F to enter Eq. 2.1, and the Fourier equation to calculate T entering Eq. 2.2. Note that this model attributes resistive switching to a pure migration of defects, without any significant generation or recombination of defects.

These are assumed to be generated at forming, and remain confined in the CF region with negligible loss during the set/reset cycling. The migration of ions within an active region, generally consisting of the CF area, results in a change of chemical composition which affects the local resistance. To describe the impact of composition on resistivity, the defects, *e.g.*, oxygen vacancies and hafnium ions, can be considered to act as dopants in the metal oxide [206]. In fact, increasing the defect density in a metal oxide is known to affect the local density of states (DOS), by introducing states in the gap which can act as doping [213, 214]. According to this picture, the

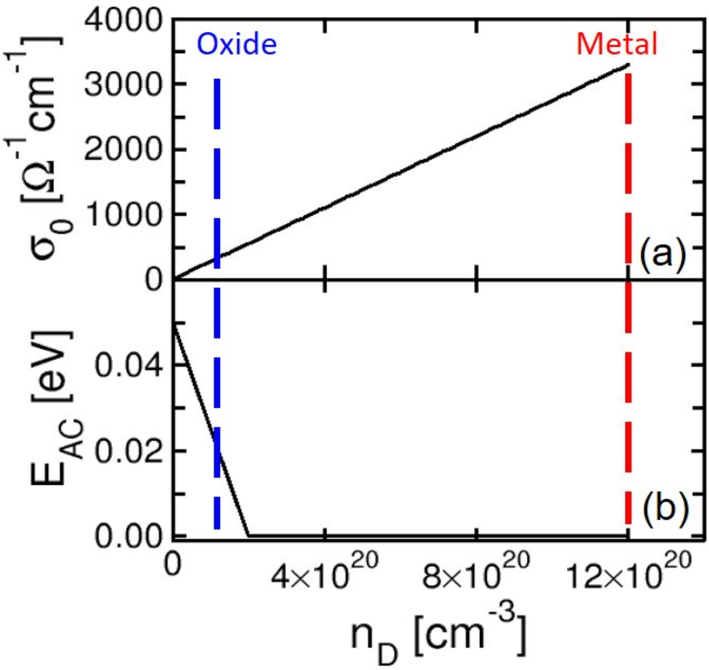


Figure 2.3: *Calculated evolution of electrical conductivity parameters in Eq. 2.4, namely (a) the pre-exponential factor σ_0 and (b) the activation energy E_{AC} at increasing of defect density n_D. Copyright 2012 IEEE. Reprinted, with permission, from [206].*

local defect concentration n_D controls the electrical conductivity σ, which is assumed dependent on temperature via an Arrhenius law given by:

$$\sigma = \sigma_0 e^{-\frac{E_{AC}}{kT}}, \tag{2.4}$$

where σ_0 is a pre-exponential factor and E_{AC} is the activation energy for electrical conduction. In Eq. 2.4, electrical transport is assumed to obey to a thermally activated hopping mechanism, such as Poole-Frenkel, which has indeed been evidenced at relatively low conductance in RRAM devices [215].

Fig. 2.3 shows (a) σ_0 and (b) E_{AC} as a function of n_D [206]. A linear increase of σ_0 is assumed in the calculation, to describe the transition from HRS, at low defect concentrations, to LRS at high defect concentration approaching a maximum value $n_D = 1.2 \cdot 10^{21}$ cm^{-3} at which the local conductivity becomes virtually metallic. The linear increase of σ_0 is consistent with both the Poole-Frenkel picture of conduction, where each carrier is thermally emitted from a localized state, and the doping theory

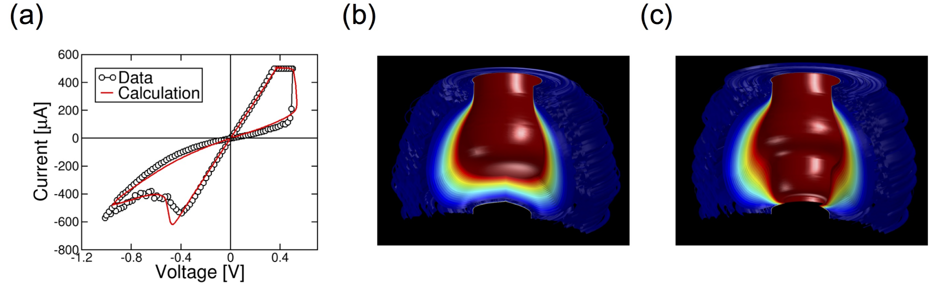


Figure 2.4: *(a) Measured and calculated I-V characteristics for HfO$_2$ RRAM device, and 3D color plots of (b) reset and (c) set states obtained by the FEM model described in [206]. In the color plot, red and blue regions indicate high and low concentration of defects, respectively. Reprinted from [205].*

in semiconductors, where carriers originate from the ionization of doping atoms. The activation energy E$_{AC}$ is assumed zero for high n$_D$, because of the doped-semiconductor or metallic-like conduction of CF in the set state, while E$_{AC}$ is assumed to linearly increase for decreasing n$_D$ close to zero as a result of a Poole-Frenkel-type electrical conduction in the case of disconnected filament.

Fig. 2.4(a) shows the measured and calculated I-V characteristics of HfO$_x$-based bipolar RRAM evidencing an abrupt set transition and a more gradual reset process. The latter is due to the migration of ionized defects activated by field and temperature toward the negatively biased top electrode resulting in a depleted gap along CF [51, 206]. The depletion process is seen to start close to the middle of CF, where T generally reaches its maximum value along the CF [206]. This physical explanation of reset process is supported by the evolution of the defect density calculated by the numerical FEM model [206], which is shown in Fig. 2.4(b) at the end of the reset transition, *i.e.*, for the HRS. In fact, the map evidences a clear depletion region, or depleted gap, extending close to the bottom electrode. In this depleted gap, the concentration of defects is so low that the conductivity pre-factor σ_0 is relatively small, while the energy barrier is large according to Fig. 2.3, therefore resulting in a relatively large resistance in the depleted region which is at the origin of the resistance rise during the reset process.

On the other hand, when a positive voltage is applied to the top electrode, ionized defects migrate in the direction of the electric field toward the bottom electrode, causing a fast increase of defect density in the depleted gap. The map of n$_D$ at the end of the set transition, namely for the

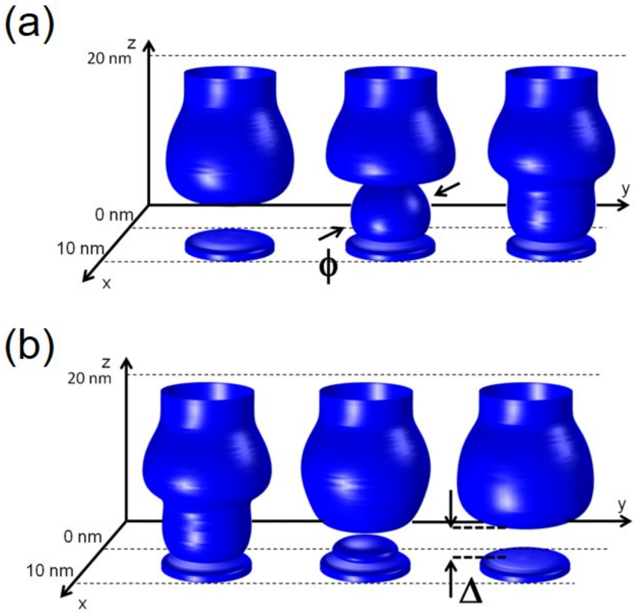


Figure 2.5: *3D contour plots of the defect concentration illustrating the evolution of (a) set transition by the formation and growth of the CF and (b) reset transition via a gradual opening of a depleted gap. Copyright 2014 IEEE. Reprinted, with permission, from [216].*

LRS, in Fig. 2.4(c) shows no depleted gap and a continuous CF with low resistance. More details about the evolution of the CF during set transition are obtained by 3D contour plots of defect density shown in Fig. 2.5(a).

From the initial HRS, the set process results in the connection of top and bottom stubs via formation of a sub CF whose diameter ϕ increases until reaching a maximum value limited by the compliance current. Fig. 2.5(b) illustrates the evolution of CF shape during reset transition, showing the formation and the gradual opening of the depleted gap with length Δ reaching a maximum value in the HRS [206, 216].

Fig. 2.6 shows the (a) measured and (b) calculated current during the reset transition as a function of the absolute value of voltage. The I-V curves are shown for various initial set states (S_1, S_2, S_3 and S_4) differing by their diameter ϕ, namely initial resistance increases from S_1 to S_4 as ϕ decreases due to a decreasing compliance current I_C used during the previous set transition [217]. Note that the reset voltage V_{reset} is almost constant for all set states, thus the reset current linearly increases with LRS conductance

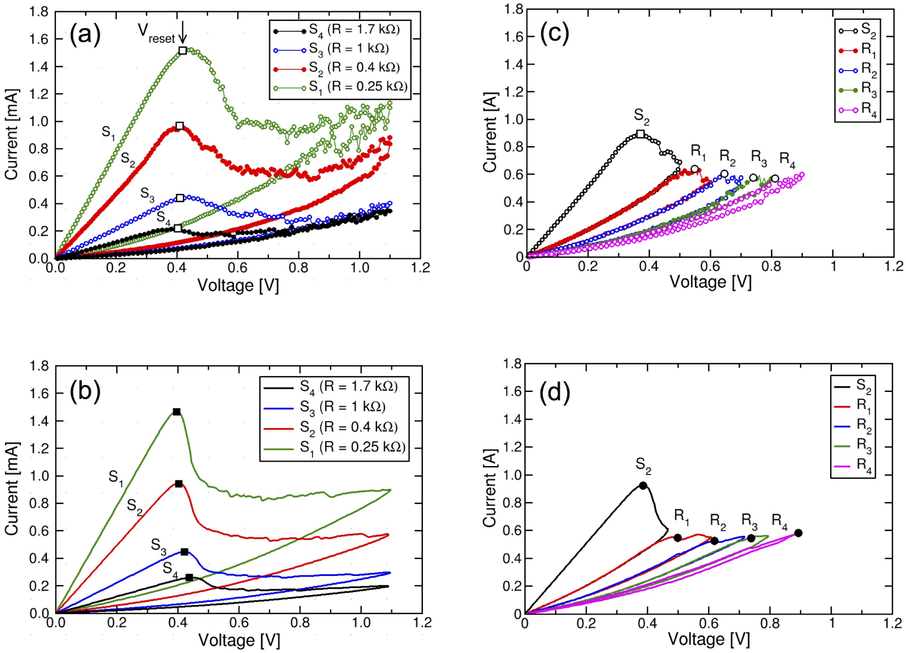


Figure 2.6: *(a)-(b) Measured and calculated I-V characteristics showing reset transitions at variable initial LRS resistance (S_1 - S_4). Both measured and simulated curves evidence that V_{reset} does not depend on initial state. (c)-(d) Measured and calculated I-V curves for variable HRS (R_1 - R_4) obtained by voltage sweeps at increasing V_{stop} starting from the set state S_2 of resistance R = 0.4 kΩ. V_{reset} increases with the initial resistance of the HRS. Copyright 2012 IEEE. Reprinted, with permission, from [217].*

1/R, or equivalently with the cross-sectional area of the CF. Also, note that I$_{reset}$ $\approx$ I$_C$ in Fig. 2.6 (a) and (b) since V$_{reset}$ is almost equal to V$_C$, *i.e.*, the critical voltage controlling ionic migration during set transition.

Fig. 2.6 also shows the (c) measured and (d) calculated I-V curves of reset transition for various initial states, including a set state S$_2$ of resistance R = 0.4 kΩ and four reset states (R$_1$, R$_2$, R$_3$, and R$_4$) of increasing resistance. These reset states were obtained by applying consecutive reset sweeps with increasing stop voltage V$_{stop}$, namely the maximum voltage in the reset transition. As V$_{stop}$ increases, the depleted gap length Δ increases in the final reset state, thus R also gradually increases from R$_1$ to R$_4$. The first reset state R$_1$ was obtained by resetting S$_2$ with V$_{stop}$ = 0.5 V.

Afterward, starting from R$_1$, a second voltage sweep with V$_{stop}$ = 0.6 V is applied causing the device resistance to increase to a higher value corresponding to the reset state R$_2$. Finally, R$_3$ and R$_4$ are obtained by the

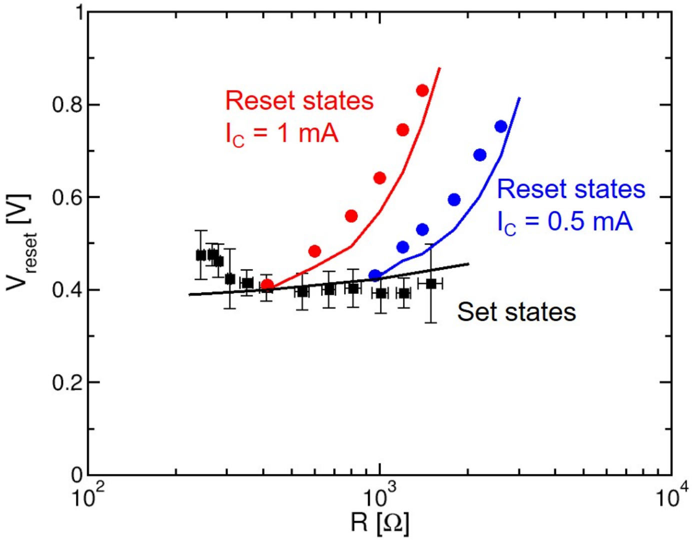


Figure 2.7: *Measured and simulated V_{reset} as a function of R for variable set states, differing by I_C in the set transition, and variable reset states, differing by V_{stop} in the reset transition. Reset states resulting from set states obtained at two different values of I_C (0.5 mA and 1 mA) are compared in the figure. Note that V_{reset} is almost constant for set states, while it increases with R for reset states. Copyright 2012 IEEE. Reprinted, with permission, from [217].*

application of further consecutive sweeps at increasing V$_{stop}$ resulting in a further increase of R. Note that V$_{reset}$, defined as the first voltage evidencing an increase of R, increases with the initial resistance of the reset state in both the experimental data and the calculations, which is in contrast with the behavior of V$_{reset}$ observed for set states in Fig. 2.6 (a) and (b).

The different behavior of V$_{reset}$ is further summarized in Fig. 2.7, collecting the measured and calculated V$_{reset}$ for variable set and reset states. Set states are achieved at variable I$_C$ while reset states are obtained at variable V$_{stop}$ starting from 2 initial set states with I$_C$ = 1 mA and I$_C$ = 0.5 mA, respectively. In the case of the set states, V$_{reset}$ remains essentially constant at 0.4 V since the maximum electric field and maximum temperature in the CF are not affected by any change in CF diameter and cross-section [215]. On the other hand, reset states with increasing R show an increasing V$_{reset}$, as a result of the increasing length of the depleted gap. In fact, the electric field is strongly localized at the depleted gap, and the longer is the depleted region, the smaller is the remaining field across the conductive region of the CF, where F drives ionic migration at the origin of the reset transition.

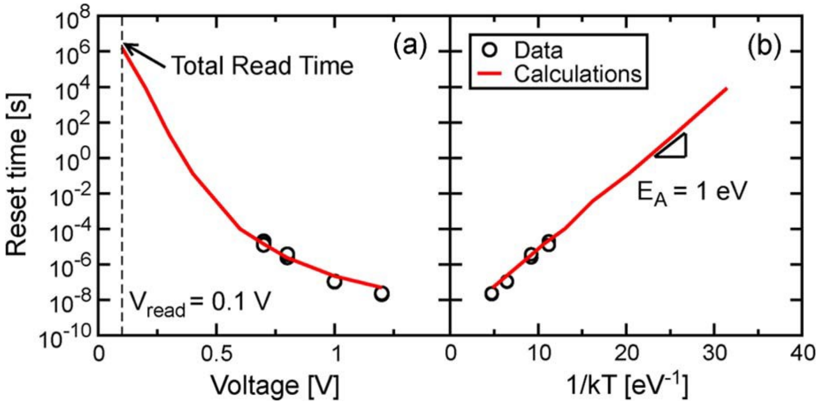


Figure 2.8: *Measured and calculated evolution of reset time as a function of (a) pulse amplitude and (b) 1/kT, where T indicates the maximum temperature in the CF calculated by the numerical model. Copyright 2012 IEEE. Reprinted, with permission, from [206].*

As a result, to activate ion migration in reset states, V$_{reset}$ must increase according to the gap extension.

In addition to static DC characteristics as in Figs. 2.4 and 2.6, the numerical drift-diffusion model can provide accurate prediction of AC-type measurement results, such as V$_{reset}$ under variable sweep rate, or reset time at constant voltage. Fig. 2.8 shows the measured and calculated reset time defined as the time to observe an increase of resistance by 60 % with respect to the initial value during the reset transition at constant voltage [206,217]. The reset time in Fig. 2.8(a) shows a highly non-linear dependence on the absolute value of the applied voltage. This can be explained by the Arrhenius dependence of diffusion kinetics in Eq. 2.2, where the local T is induced by Joule heating, thus increases approximately with the square of the applied voltage [51]. To support this explanation, Fig. 2.8(b) shows the reset time as a function of 1/kT, where T was evaluated from the model as the maximum temperature along the CF at the reset transition. Data and calculations show a clear exponential dependence, thus evidencing the Arrhenius dependence and supporting the crucial role of temperature in accelerating ion migration and reset transition.

The FEM model thus shows a full capability to predict device behavior under both basic lab-type experiment, such as quasi-static I-V curves, and more application-driven explorations of device speed, thus satisfying the need for industrial technology computer-aided design (TCAD) device simulations.

2.3 Analytical model for HfO$_2$ RRAM devices

For the design and simulation of circuits comprising RRAM devices, the numerical model of Section 2.2 is not suitable because of a relatively high computational cost and long solution time [205]. This limitation can be overcome by physics-based compact models, where the device characteristics can be calculated by the solution of simplified analytical equations [66].

In general, the starting point for developing a compact model is to learn the switching mechanism from a detailed device simulation, such as the FEM simulation of filamentary switching shown in Fig. 2.5. Here, the CF shows distinctly different evolutions during set and reset processes: set transition consists of a sudden appearance of defects within the depleted gap, followed by a CF growth in terms of defect density and CF diameter within the depleted gap. On the other hand, reset transition is due to an increased length of the depleted gap. The “explosive” nature of set process agrees well with the abrupt change of current in the I-V curves, compared to the more gradual transition in the reset process.

In the analytical model for RRAM switching described in [66], the different dynamics of set and reset processes can be understood by the positive or negative feedback of electric field, temperature, and the defect distribution along the CF [66, 206]. In fact, defects during set transition migrate in response to the large electric field across the depleted gap. As defect migration starts to take place, the depleted gap length decreases, thus the local electric field increases,which further accelerates defect migration. Such positive feedback effect would result in a destructive failure of the device; however, current limitation (compliance) systems introduce an external negative feedback which allows to reduce the voltage during set transition, thus preventing destructive breakdown and enabling a detailed control of the final CF size and resistance [51, 66].

On the other hand, defect migration during reset transition is triggered by a relatively low electric field across the continuous CF. As the depleted gap starts to form, the electric field decreases in the CF regions where defects are located, thus slowing down the migration kinetics. As a result of such negative feedback effect, the voltage must be increased to further sustain the reset transition, resulting in the gradual increase of resistance.

Fig. 2.9 shows the CF evolution in a filamentary-type RRAM during (a) set and (b) reset transition [66]. The CF evolution mimics the observed set/reset migration dynamics in Fig. 2.5, namely, set transition evolves via the growth of CF diameter ϕ within the depleted gap region (a), whereas reset transition occurs by the gradual increase of the depleted gap length

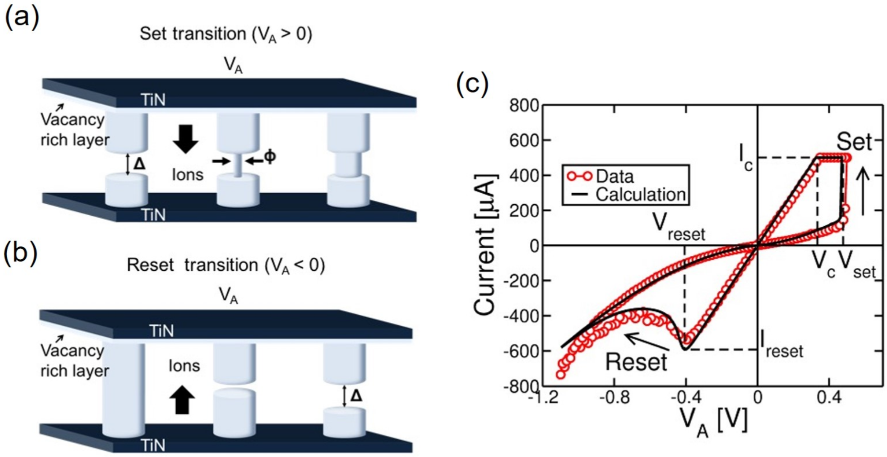


Figure 2.9: *Schematic illustration of filament evolution during switching in RRAM for (a) set transition, (b) reset transition, and (c) I-V curve calculated with an analytical model, compared to experimental data for a TiN/HfO_2/TiN device. Copyright 2014 IEEE. Adapted, with permission, from [66].*

Δ (b). Formally, the rate equations for ϕ and Δ resemble the drift/diffusion equations governing the continuous FEM modeling of RRAM [206], namely:

$$\frac{d\phi}{dt} = Ae^{-\frac{E_A}{kT_{inj}}}, \tag{2.5}$$

for set transition, where A is a pre-exponential constant, E_A is a voltage-dependent energy barrier for migration, and T_{inj} is the local temperature at the injecting CF tip, namely the one with positive potential. A similar rate equation was assumed for reset transition, namely:

$$\frac{d\Delta}{dt} = Ae^{-\frac{E_A}{kT_{inj}}}, \tag{2.6}$$

where T_{inj} is again calculated at the positively biased, injecting CF tip [66]. These equations can be viewed as a simplified description of the CF evolution mechanism, where the CF evolves via Arrhenius-type migration dynamics controlled by an energy barrier E_A, and driven by the local electric field and the local temperature T_{inj}.

Fig. 2.9(c) shows the measured and calculated I-V curve obtained by this model: simulation results show the same abrupt change of resistance during set transition, and a gradual change of resistance during reset transition,

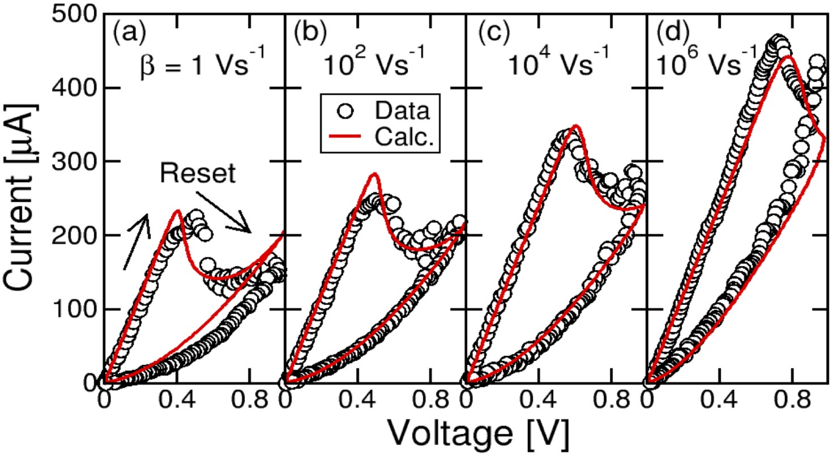


Figure 2.10: *Measured and calculated I-V characteristics showing reset transition at increasing sweep rate, namely (a) β = 1 V/s, (b) β = 10^2 V/s, (c) β = 10^4 V/s and (d) β = 10^6 V/s. Copyright 2014 IEEE. Reprinted, with permission, from [66].*

thus demonstrating that it correctly captures the positive/negative feedback loops controlling the microscopic CF evolution. Among the model equations, it is necessary to include (i) a shape-resistance relationship allowing to derive R for each value of ϕ and Δ, and (ii) a simplified electro-thermal model allowing to estimate the local temperature T$_{inj}$ based on the dissipated power V·I, and based on a detailed description of the thermal resistance controlling heat exchange across the time-varying CF and the surrounding oxide layer [66].

In the simulation results of Fig. 2.9, a migration energy barrier E$_A$ = 1.2 eV was assumed, thus similar to the values derived from time-dependent analysis of switching by numerical simulations [206], and similar to independent ab-initio studies of diffusion barriers in amorphous HfO$_2$ [218].

To better support the feasibility of Eqs. 2.5 and 2.6 combined with this value of E$_A$, Fig. 2.10 shows the measured and calculated I-V curves describing the reset transition at variable rate of the applied voltage sweep [66]. As the sweep rate $\beta = \frac{dV}{dt}$ was increased from 1 Vs^{-1} to 10^6 Vs^{-1}, the reset voltage and corresponding reset current increased by about a factor 2, although the initial LRS resistance was kept constant. This is due to the time-dependent reset dynamics, where a higher local T$_{inj}$, hence a higher V$_{reset}$, is needed to trigger ionic migration within a shorter time according to the Arrhenius law in Eqs. 2.5 and 2.6.

The analytical simulations in Fig. 2.10 agree very well with the ex-

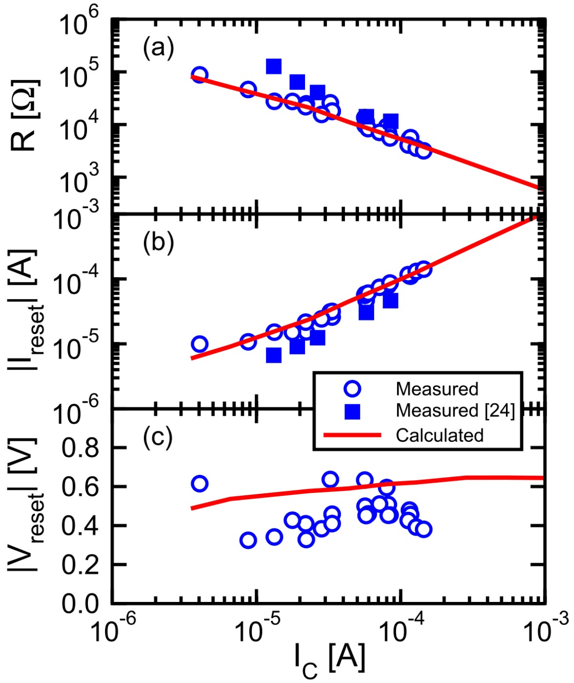


Figure 2.11: *Measured and calculated (a) average LRS resistance R, (b) reset current I_{reset} and (c) reset voltage V_{reset}, as a function of the compliance current I_C. Data were collected for integrated one-transistor/one-resistor (1T1R) structures allowing control of the LRS in the range 10-100 kΩ for I_C in the range 10-100 μA. Calculations agree very well with experimental data, supporting multilevel cell control of LRS resistance and low power operation of RRAM. Copyright 2014 IEEE. Reprinted, with permission, from [216].*

perimental data, supporting the accuracy of the rate equations and of the energy barrier E$_A$ assumed in the calculations of resistance switching in TiN/HfO$_2$/TiN. Note that a different material and/or stack would lead to different values of A and E$_A$ in the equations; thus, this compact model requires careful adjustment to describe a specific RRAM technology.

The model also accounts for the dependence on current compliance I$_C$ via the LRS resistance.

Fig. 2.11 shows (a) the measured and calculated resistance R, (b) reset current I$_{reset}$ and (c) reset voltage V$_{reset}$, as a function of I$_C$. These experimental results were collected for integrated one-transistor/one-resistor (1T1R) structures, where the small parasitic capacitance allowed for a tight control of the maximum current during set transition close to I$_C$ and without significant overshoots [219]. As I_C decreases, LRS increases as a result of the reduced maximum CF size reached within the experimental time, which was about 1 s in the DC experiments of Fig. 2.11. In fact, a relatively small I_C causes a negative-feedback-induced voltage snap back to occur at relatively low current, thus forcing the final resistance to a relatively high value R = V_C/I_C, where V_C is a characteristic voltage capable of inducing ionic migration on experimental time scale [51, 66]. Analysis of data in the figure indicates V_C = 0.5 V for these experimental devices, in agreement with other RRAM device technologies including both unipolar and bipolar switching RRAMs [66]. The reset current increases with I_C as a result of the decreasing R and of the constant reset voltage V_{reset}. This is almost equal to V_C, thus suggesting a symmetric behavior of ionic migration with respect to voltage polarity. Two device types differing in $\mathrm{HfO_2}$ thickness and deposition recipe are compared in the figure [216, 220], however indicating only minor deviations. In particular, the value of V_C was shown to depend only slightly on the device material/stack and geometry parameters, such as the thickness of the oxide layer, or the length of the CF [51]. This can be explained by the analytical formula for the maximum temperature along the CF, given by:

$$T = T_0 + \frac{R_{th}}{R}V^2 = T_0 + \frac{V^2}{8\rho k_{th}}, \tag{2.7}$$

where T_0 is the room temperature, R_{th}/R is the ratio between thermal and electrical resistances of the CF, V is the voltage drop across the CF, ρ is the electrical resistivity and k_{th} is the thermal conductivity of the CF materials. The equation indicates that the local temperature does not depend on CF thickness, but is solely controlled by applied voltage since R_{th}/R is approximately constant. The balancing effect of thermal/electrical resistances can be explained as follows: as the thickness increases, the power dissipation P = V^2/R within the CF decreases, while the corresponding temperature along the CF increases. As a result, the same voltage V_C is needed to achieve the critical temperature needed to induce migration within the time scale of the experiment [51].

2.4 Stochastic model for $\mathrm{HfO_2}$ RRAM devices

A key challenge of RRAM devices is the switching variability, where a single cell operated for several set/reset cycles displays different switching characteristics from cycle to cycle [216, 220]. Similarly, RRAM devices

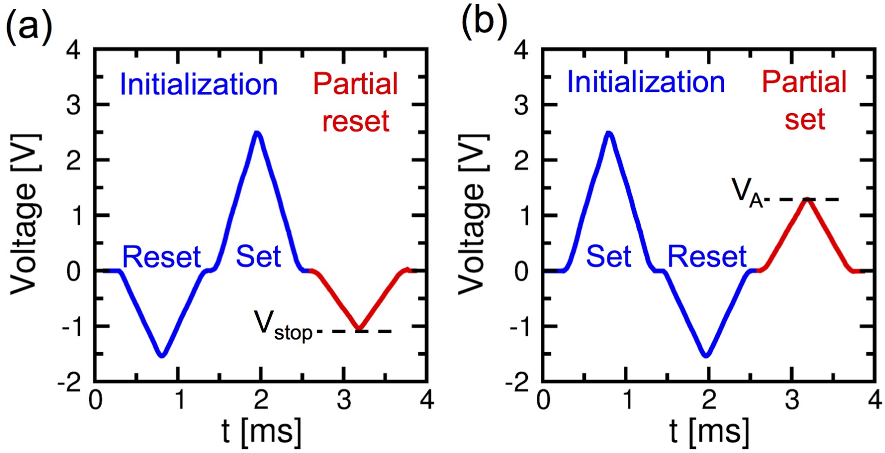


Figure 2.12: *Sequences of voltage pulses used to characterize (a) random reset and (b) random set transitions in a 1T1R structure. Copyright 2016 IEEE. Reprinted, with permission, from [127].*

can display a cell-to-cell variability, where different cells display different switching processes.

To study switching variability, the I-V characteristics are usually collected on Ti/HfO_x/TiN RRAM samples with 1T1R structure [127]. To address the stochastic variation of conductance, set and reset transitions were experimentally studied at variable applied voltage with the application of triangular pulses as shown in Fig. 2.12.

First, the reset statistics at variable V_{stop} was studied by the pulse sequence in Fig. 2.12(a), including an initial reset/set sequence to prepare the device in the LRS, followed by a negative voltage pulse with increasing amplitude V_{stop} to induce a reset transition. On the other hand, Fig. 2.12(b) shows the sequence adopted to study the set statistics, including an initial set/reset sequence to prepare the device in the HRS, followed by a final triangular voltage pulse with increasing positive amplitude V_A. Both set and reset experiments were repeated for 10^3 times to obtain resistance distributions based on sufficient statistics [127].

Fig. 2.13(a) shows the cumulative distributions of resistance obtained via the reset measurements with variable V_{stop} from -0.7 V to -1.6 V. As V_{stop} increases, the average resistance increases from LRS (insufficient voltage to induce reset) to HRS [221]. The figure also shows calculated distributions according to an empirical Monte Carlo model aimed at predicting the HRS distributions as a function of V_{stop}. The distributions were

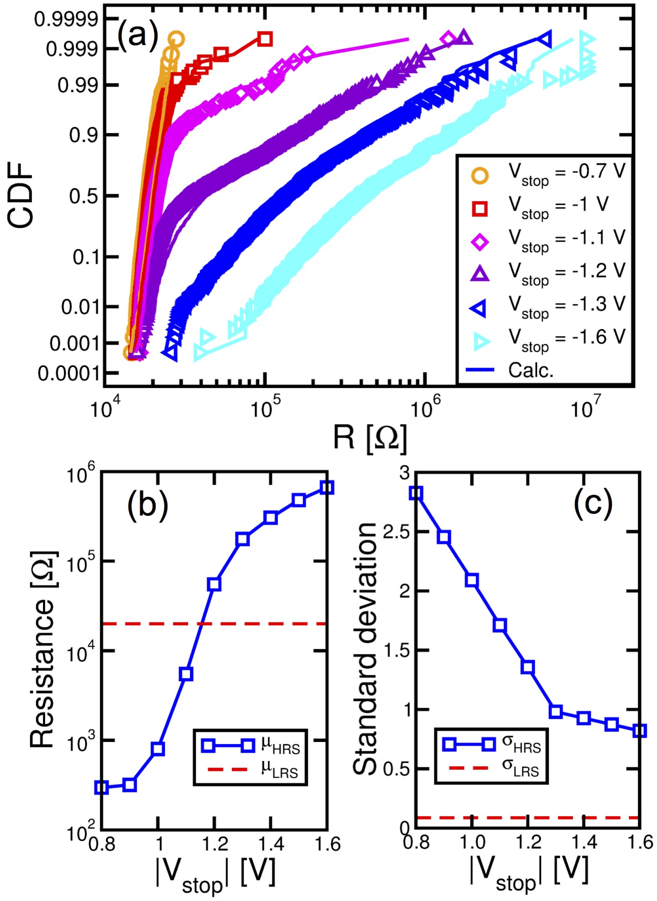


Figure 2.13: *(a) Measured and calculated cumulative distributions of resistance describing reset states at increasing |V$_{stop}$|, (b) average values μ_{LRS} and μ_{HRS}, and (c) standard deviations σ_{LRS} and σ_{HRS} as a function of V$_{stop}$. The average values and standard deviations can be used to calculate the reset statistical distribution by a Monte Carlo model of synaptic weight update. Copyright 2016 IEEE. Reprinted, with permission, from [127].*

calculated by the superposition of a lognormal sub-distribution of LRS with average value μ_{LRS} and standard deviation σ_{LRS}, and a lognormal sub-distribution of HRS with average value μ_{HRS} and standard deviation σ_{HRS}. Fig. 2.13(b) shows the average values μ_{LRS} and μ_{HRS}, while Fig. 2.13(c) shows the standard deviations σ_{LRS} and σ_{HRS} as a function of V$_{stop}$. The use of this set of parameters allows for the calculation of the HRS distributions after synaptic depression at a generic voltage V$_{stop}$ [127].

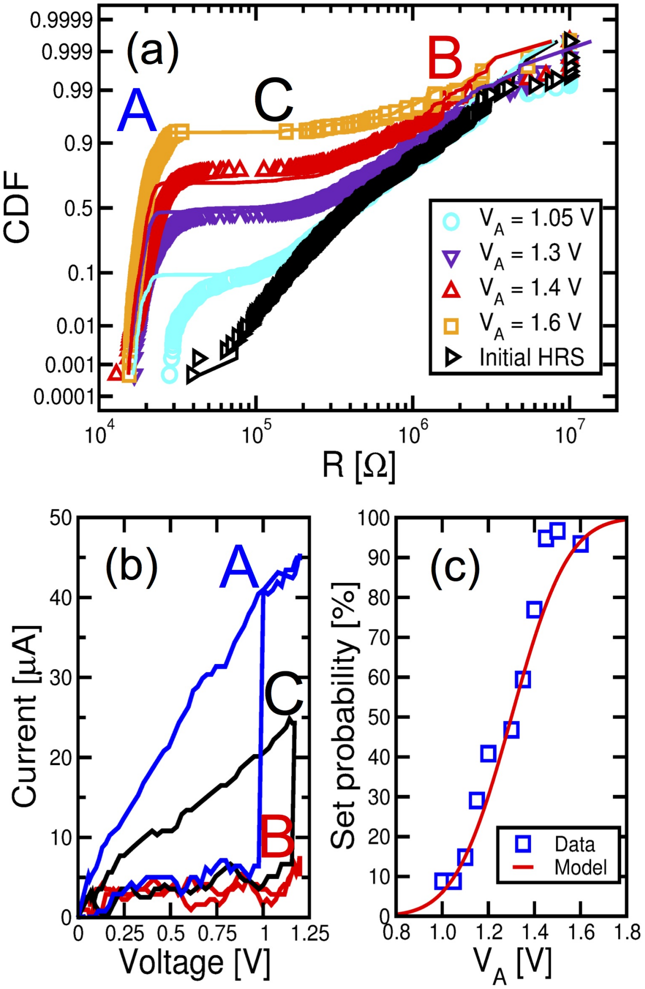


Figure 2.14: *(a) Measured and calculated cumulative distributions of resistance obtained by the random set experiments at increasing V_A. (b) I-V characteristics corresponding to 3 possible results of the set process, namely set transition (state A), no set transition (state B), and incomplete set transition (state C). (c) Measured and calculated set probability P_{set} as a function of V_A. Copyright 2016 IEEE. Reprinted, with permission, from [127].*

Fig. 2.14(a) shows the measured distributions of resistance with increasing applied voltage V_A for set transition. Starting from the HRS, the application of a pulse of voltage V_A results in set transition only for a fraction of cycles, as shown in Fig. 2.14(b). Here, the set process for a given voltage (V_A = 1.2 V) appears to be random, as a result of the statistical variability of the set voltage V$_{set}$. For instance, the set transition takes place for cycle A as V$_A$ is larger than V$_{set}$ for that particular cycle. On the other hand, no set transition is seen for cycle B, due to $V_A < V_{set}$.

Finally, an incomplete set transition occurs for cycle C, possibly due to V$_A$ being very close to V$_{set}$, and the migration taking place for a relatively short time during the pulse. States A (LRS), B (HRS), and C (incomplete LRS) are shown in the distributions of Fig. 2.14(a) for reference. Fig. 2.14(c) shows the probability for set transition as a function of V$_A$, with the criterion that the resistance R is below 80 kΩ. The set probability increases as V$_A$ increases compared to the average V$_{set}$. The figure also shows the calculated results of the compact formula for P$_{set}$ given by:

$$P_{set} = \frac{1 + erf(\frac{V_A - \mu_V}{\sqrt{2}\sigma_V})}{2}, \tag{2.8}$$

where μ_V = 1.3 V and σ_V = 0.193 V are the average value and standard deviation of V$_{set}$, respectively. Based on P$_{set}$, it is possible to predict the distributions of R in Fig. 2.14(a) by a Monte Carlo model combinating random HRS and LRS resistance with probabilities 1-P$_{set}$ and P$_{set}$, respectively. The results are in good agreement with the measured distributions, supporting the stochastic Monte Carlo model for RRAM resistance distributions. For example, the stochastic set process in Fig. 2.14(a) can be useful for true random number generator (TRNG), aimed at generating random bits by the inherent stochastic phenomena in the device physics [222, 223].

2.5 Conclusions

In this chapter, a previous FEM numerical model of HfO$_2$ RRAM devices capable of providing a deep understanding of set and reset processes and their microscopic interpretation has been first reviewed. Afterward, a previous compact analytical model based on detailed device characteristics achieved by numerical model has been discussed evidencing its capability of accurately capturing the different dynamics of set and reset processes by positive and negative feedback effects, respectively. Finally, a stochastic Monte Carlo simulator capable of accurately predicting the statistical variability of set and reset transitions in HfO$_2$ RRAM devices, which makes it particularly suitable for calculation of synaptic weight updates in neuromorphic circuits with RRAM-based synapses, has been presented.

CHAPTER 3

Resistive switching synapses for neuromorphic computing

3.1 Introduction

Neuromorphic computing is attracting an increasing interest for cognitive functions, such as pattern recognition [152] and natural language processing [157]. In a neuromorphic circuit, integrate-and-fire (I&F) neurons are connected by synapses, and usually process information by event-driven spiking activity [6]. Spikes serve for both carrying the information and inducing plasticity in the synapses, which forms the basis for learning. Brain-inspired learning rules are generally based on the timing of the spike arriving from the pre-synaptic neuron, or PRE, and the spike delivered by the post-synaptic neuron, or POST. For instance, in spike-timing dependent plasticity (STDP), the change of synaptic weight is dictated by the delay between PRE and POST spikes. STDP has been demonstrated to occur in certain synapses in the brain [111, 112], and are currently among the most popular approaches for unsupervised training of neural networks [123, 224, 225].

Other learning rules have been considered to be responsible for learning in biological neural networks. According to the Bienenstock-Cooper-Munro (BCM) theory [226], synaptic plasticity is governed by the PRE and POST spike frequency, rather than the timing of a pair of PRE and POST spikes. A high frequency of PRE and POST spikes leads to potentiation, while a low frequency leads to depression. This spike-rate dependent plasticity (SRDP) has been recognized as a bio-realistic learning rule [227], and linked to triplet based learning rules [135], where potentiation relies on the temporal occurrence of 3 spikes [134]. Integrated circuits capable of learning by STDP or SRDP rules generally require complicated and large synaptic blocks hosting multiple transistors and capacitors [106, 228]. To enable small-area synapse, hence high density neural circuits, emerging memories such as resistive switching memory (RRAM) and phase change memory (PCM) have recently attracted a strong interest [114, 115, 118, 119, 121, 126–128, 137, 141, 229, 230]. The development of RRAM-based SRDP synapses is still a major challenge for neuromorphic engineering [129, 138–140, 146, 231].

In this chapter, which is based on works [118, 127, 129, 143, 232], two RRAM-based synaptic structures capable of STDP and SRDP, respectively, are presented. First, a hybrid structure comprising a RRAM device serially connected to one transistor, referred to as one-transistor/one-resistor (1T1R) structure, is described to demonstrate STDP rule via experiments and simulations at the level of single device. In addition, a resistive synapse with 4-transistors/one-resistor (4T1R) structure is described to demonstrate a SRDP learning rule where potentiation and depression processes are achieved via 3-spike overlapping according to a modified triplet rule. To support simulation results, frequency-dependent synaptic operation was also tested on a synapse prototype providing extensive experimental characteristics.

3.2 1T1R synapse for STDP learning

3.2.1 1T1R synapse structure

Fig. 3.1(a) shows the sketch of a hybrid CMOS/RRAM synapse with 1T1R structure consisting of a RRAM device, which is based on a Si-doped HfO$_x$ layer interposed between a TiN bottom electrode (BE) and a Ti top electrode (TE), serially connected to a field-effect transistor (FET). In 1T1R synapse, FET serves as selector element enabling to access the device for the gate voltage V$_G$ above the threshold voltage and also the limitation of compliance current I$_C$ to control the CF diameter, hence the current consumption during set and reset processes. Fig. 3.1(b) shows the measured

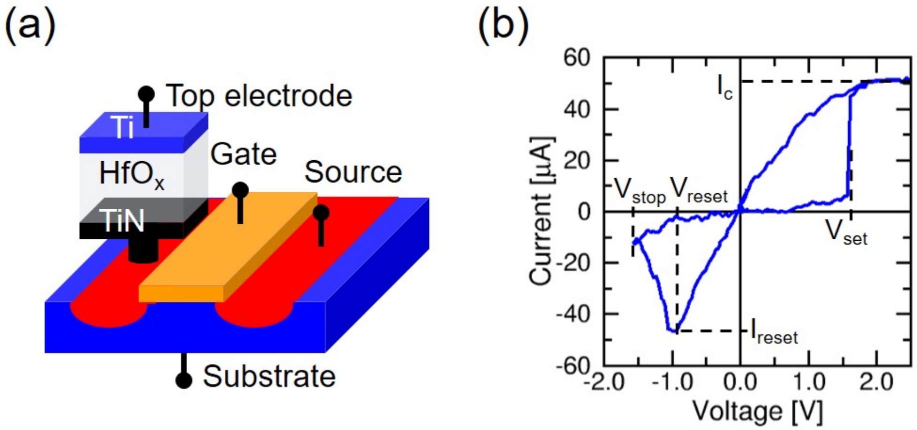


Figure 3.1: *(a) Sketch of an electronic synapse based on a Ti/HfO$_x$/TiN RRAM device with 1T1R configuration, and (b) measured I-V characteristics of a 1T1R synapse with compliance current I_C = 50 µA during the set transition. The gate voltage V_G allows to control the compliance current I_C. Copyright 2016 IEEE. Adapted, with permission, from [127].*

I-V characteristics of a Ti/HfO$_x$/TiN RRAM, with the compliance current of 50 μA controlling the conductance of the low resistance state (LRS), while the high resistance state (HRS) is controlled by V$_{stop}$. Also, note that set transition to achieve LRS is activated at voltage V$_{set}$ $\approx$ 1.5 V while reset transition to achieve HRS is activated at voltage V$_{reset}$ $\approx$ -1 V [127].

In Fig. 3.2, the 1T1R synapse is shown as a connecting element between a pre-synaptic neuron (PRE) and a POST-synaptic neuron (POST) [127, 233, 234]. The PRE is connected to the gate of the FET in the 1T1R structure, while the POST receives the synaptic current from the BE while controlling the voltage at the TE of the 1T1R structure.

The operation of the 1T1R synapse can be understood as follows: as the PRE emits a positive voltage spike, the FET acts as a pass-transistor enabling a synaptic current proportional to the RRAM synaptic conductance. The current spike enters the POST via the BE which can collect incoming currents from several synaptic channels, as in the ideal McCulloch-Pitts (MCP) neuron scheme [235]. The currents are integrated in the I&F POST circuit, eventually leading to a fire event as the integral signal V$_{int}$ hits the threshold V$_{th}$. At fire, the POST generates a spike toward the next layer of neurons, and additionally applies a feedback spike to the synapse TE. The feedback spike consists of the sequence of a positive pulse with amplitude

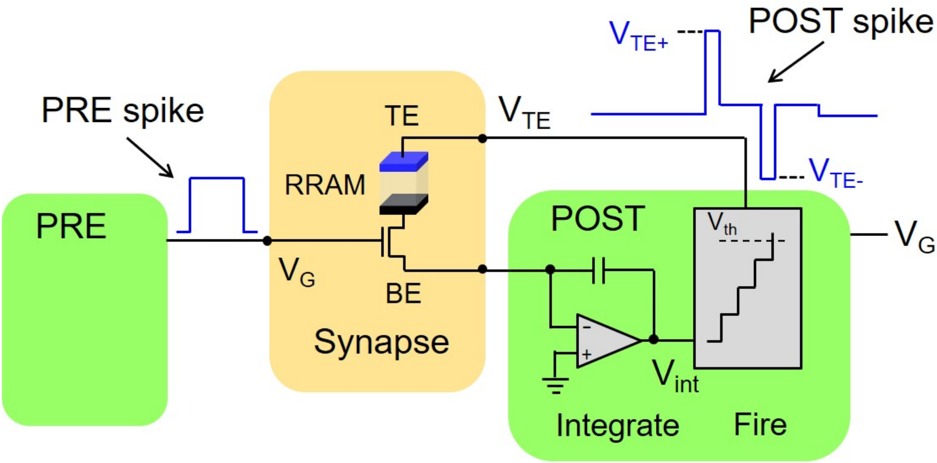


Figure 3.2: *Sketch of a resistive synapse with 1T1R configuration connected to a pre-synaptic (PRE) neuron and a post-synaptic (POST) neuron. As a spike is emitted by PRE, a current spike is activated across the synapse leading to an increase of V_{int} within the POST. As POST integration leads V_{int} to exceed the internal threshold V_{th}, a fire spike is backward applied to the TE of the 1T1R synapse, causing a weight update according to STDP. Copyright 2016 IEEE. Adapted, with permission, from [233].*

$V_{TE+} > V_{set}$ and a negative pulse with amplitude $V_{TE-} < V_{reset}$, which can induce a weight update depending on the relative timing with the PRE spike referred to as spike-timing dependent plasticity (STDP).

3.2.2 Potentiation and depression in 1T1R synapse

To achieve STDP functionality, spike waveforms are designed so that if PRE spike shortly precedes the POST spike (Fig. 3.3(a)), namely the relative time delay Δt = t_{POST} - t_{PRE} is positive (0 < Δt < 10 ms), PRE spike only overlaps with the positive pulse within POST spike, thus inducing a set transition in RRAM device, which results in the long-term potentiation (LTP) of 1T1R synapse (synaptic weight increase) as a result of RRAM conductance change from HRS to LRS.

On the other hand, if the PRE spike shortly follows the POST spike (Fig. 3.3(b)), namely the relative time delay Δt is negative (-10 ms < Δt < 0), PRE spike only overlaps with the negative pulse within POST spike causing a reset transition in RRAM device which results in the long-term depression (LTD) of 1T1R synapse (synaptic weight decrease) as a result of RRAM conductance change from LRS to HRS.

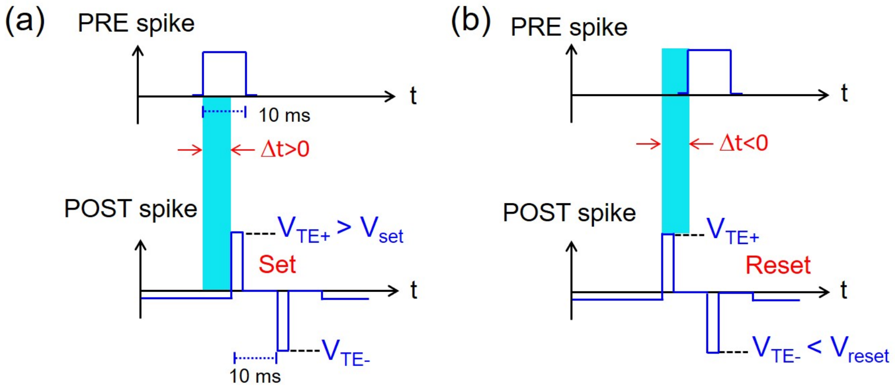


Figure 3.3: *STDP implementation by overlapping PRE and POST spikes. (a) If the PRE spike occurs before the POST spike ($\Delta t > 0$), the resistance decreases due to the positive TE spike causing set transition, or synaptic potentiation. (b) Otherwise, if the PRE spike occurs after the POST spike ($\Delta t < 0$), the resistance increases due to the negative TE spike causing reset transition, or synaptic depression. Copyright 2016 IEEE. Adapted, with permission, from [233].*

3.2.3 STDP experimental demonstration by 1T1R synapse

STDP implementation by 1T1R synapse was corroborated by experimental measurements at device level [118]. Fig. 3.4(a) shows a pair of PRE and POST voltage spikes with a positive delay Δt = 3 ms to explore LTP and a pair of PRE and POST voltage spikes with a negative delay Δt = -7 ms to study LTD. As a result, Fig. 3.4(b) shows the measured resistance evidencing a resistance decrease from HRS to LRS, namely a synaptic potentiation event, in response to the application of first PRE/POST spike pair and a resistance increase from LRS to HRS, namely a synaptic depression event, in response to the application of second PRE/POST spike pair.

Also, Fig. 3.4(c) shows the correlation plot of the resistance $R(t_{i+1})$ measured after the spike application as a function of $R(t_i)$ measured before the spike application, for variable Δt [118]. Under potentiation condition, namely for positive delay satisfying 0 < Δt < 10 ms, a RRAM prepared in HRS undergoes a set transition to the LRS, whereas if the RRAM device is initially in LRS, no resistance variation occurs because the RRAM is already at its minimum resistance state [118, 127, 236]. For negative delay satisfying -10 ms < Δt < 0, corresponding to the condition for synaptic depression, a resistance transition is activated when the RRAM device is initialized in its LRS. Finally, if Δt assumes values outside the plasticity

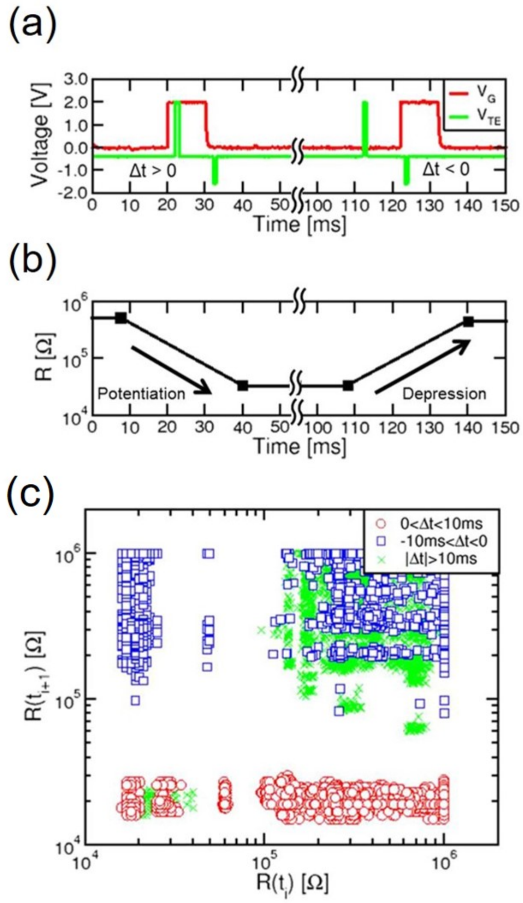


Figure 3.4: *(a) PRE and POST voltage waveforms applied to the gate and the TE, respectively, in case of positive delay (left) and negative delay (right). (b) Measured resistance evidencing a transition from HRS to LRS (synaptic potentiation) in the case of positive delay and a transition from LRS to HRS (synaptic depression) in the case of negative delay. (c) Correlation plot of the RRAM resistance at epoch t_{i+1} R(t_{i+1}) as a function of the RRAM resistance at epoch t_i R(t_i) for variable Δt, corresponding to the cases of potentiation, depression, and no change of weight because of excessive delay. Adapted from [118].*

window (|Δt| > 10 ms), the PRE and POST spikes do not overlap, therefore the RRAM resistance does not change. As a result of the full set/reset operations taking place in the plasticity mechanism, the 1T1R synapse only displays HRS and LRS resistive levels, thus evidencing the binary operation of the 1T1R synaptic device due to the relatively abrupt set and reset

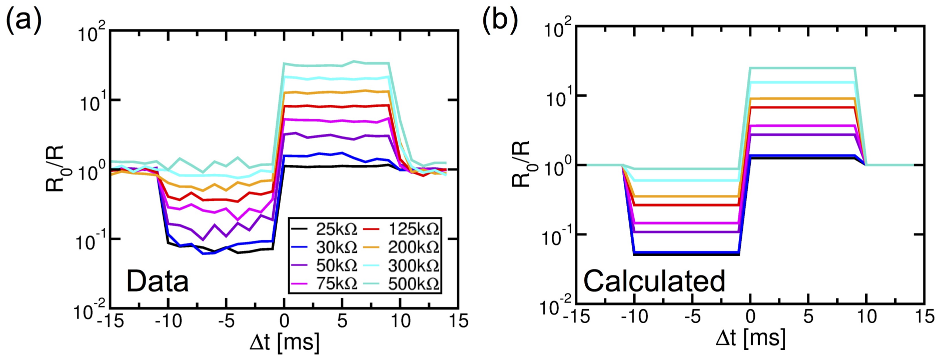


Figure 3.5: *(a) Measured and (b) calculated STDP characteristics indicating the relative change of resistance R_0/R as a function of Δt for variable initial resistance states R_0, from full LRS (R_0 = 25 kΩ) to full HRS (R_0 = 500 kΩ). Potentiation and depression are both a function of time delay and the initial synaptic state, resulting in the final state being either HRS or LRS. Copyright 2016 IEEE. Reprinted, with permission, from [127].*

transitions [118].

Note that more resistance levels can be achieved by time dependent modulation of the PRE and POST spikes as in the 2T1R synapse circuit proposed in [119]. In this synapse architecture, the waveform of the PRE spike allows for time-dependent potentiation, where a longer Δt corresponds to a smaller conductance due to the lower compliance current during set transition. On the other hand, the waveform of the POST spike allows for time-dependent depression, where a longer Δt corresponds to a smaller resistance due to the lower voltage applied during reset transition [119]. The enhanced functionality comes at the expense of a slightly higher complexity of the 2T1R synapse circuit, requiring 2 transistors instead of only one in the 1T1R synapse.

3.2.4 STDP characteristics

To further support the dependence of STDP on initial state in the 1T1R synapse, Fig. 3.5 shows the (a) measured and (b) calculated STDP characteristics, namely the ratio between the initial resistance R_0 and the final resistance after potentiation/depression, as a function of Δt for increasing R_0 [127]. Calculations were done based on the analytical model for RRAM devices [66] discussed in the section 2.3.

These results show binary STDP behavior, where the amount of potentiation and depression is a function of R_0. The variable change of resistance

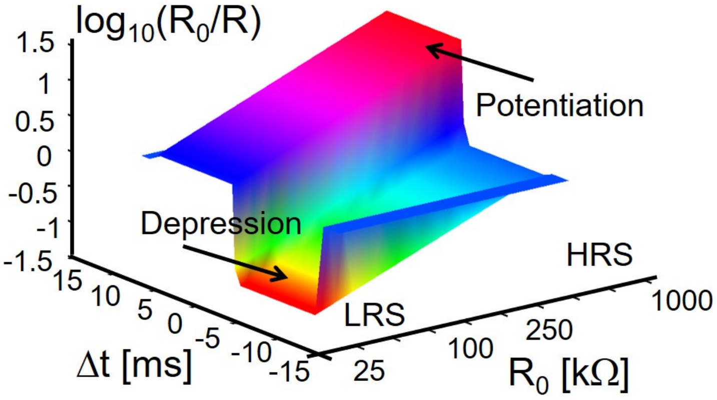


Figure 3.6: *3D color plot of calculated STDP characteristics shown in Fig. 3.5(b). Copyright 2016 IEEE. Reprinted, with permission, from [127].*

allows the final resistance to be equal to either HRS or LRS, in strong analogy with biological synapses where the weight update is limited between two boundary states.

Finally, Fig. 3.6 further illustrates the three-dimensional (3D) color map of the calculated STDP characteristics, evidencing the increase of potentiation/depression level of 1T1R synapse with increasing/decreasing R_0 for positive/negative Δt [127].

3.3 4T1R synapse for SRDP learning

3.3.1 4T1R synapse structure

According to some biological experiments [133], the rate of the spiking activity has a significant impact on plasticity of biological synapses.

To emulate the dependence of weight update process on spike rate observed in a biological synapse as one schematically depicted in Fig. 3.7(a), a synapse circuit capable of capturing SRDP was developed. The synapse shown in Fig. 3.7(b) consists of a hybrid CMOS/RRAM structure, combining 4 MOS transistors and a bipolar-switching RRAM device [118, 237], and serving as connection between a PRE and a POST [129]. In the synaptic circuit, the transistors are arranged in 2 branches, namely transistors M_1 and M_2 which are responsible for synaptic long-term potentiation (LTP), and transistors M_3 and M_4 for synaptic long-term depression (LTD). The

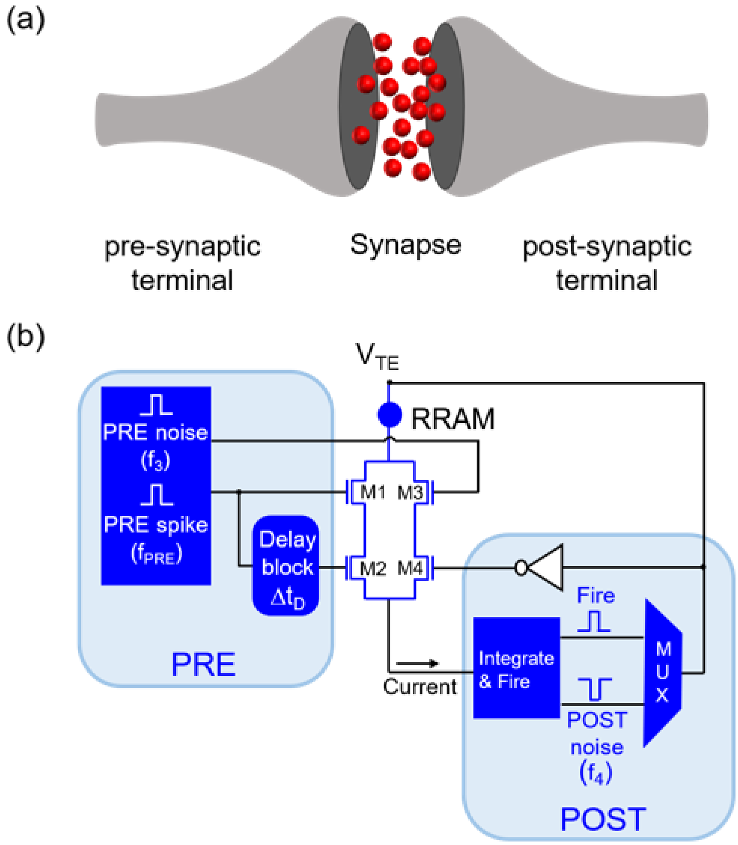


Figure 3.7: *(a) Sketch of a biological synapse connecting PRE- and POST-synaptic neurons and (b) schematic illustration of corresponding PRE-synapse-POST circuit. 4T1R synapse is capable of LTP via M_1/M_2 branch, which is controlled by PRE spikes at average frequency f_{PRE} induced by external stimuli, and LTD via M_3/M_4 branch, which is activated by PRE and POST noise spikes at average frequencies f_3 and f_4, respectively. Copyright 2018 IEEE. Reprinted, with permission, from [143].*

RRAM device is connected in series to the parallel of branches M_1/M_2 and M_3/M_4. The PRE spike is applied to the gate of M_1 and, after a delay by a time Δt_D, to the gate of M_2. The gate of M_3 is driven by a random noise PRE spiking. The POST consists of an I&F circuit, which delivers a fire spike to the TE of RRAM device as the internal potential resulting from integration exceeds a certain threshold [98, 127]. The POST also generates a negative noise spike that is alternatively submitted to the TE and, after inversion, to the gate of M_4. The POST multiplexer (MUX) activates the

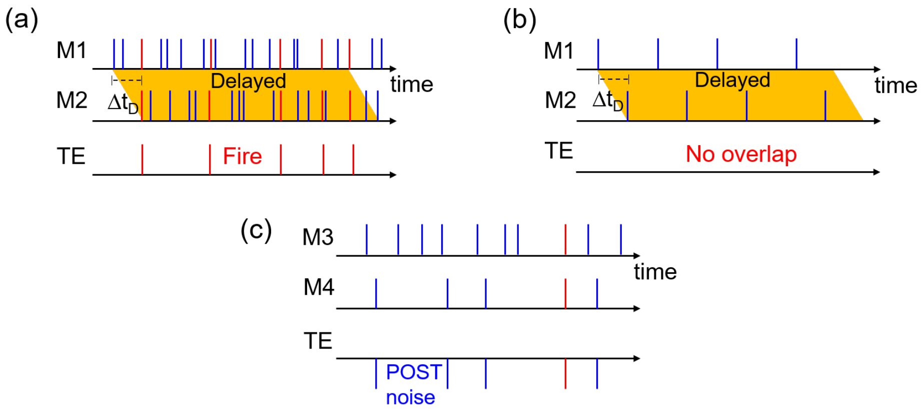


Figure 3.8: *Illustrative description of spike timing inducing (a) LTP for high frequency PRE spiking activity, (b) no LTP for low frequency PRE spiking activity, and (c) stochastic LTD via PRE and POST noise spikes. Copyright 2016 IEEE. Adapted, with permission, from [129].*

fire channel on at every POST fire, temporarily inhibiting the noise channel to the TE. Noise spikes can be obtained by tunable random number generator circuits, *e.g.*, by amplification of thermal noise, *e.g.*, 1/f noise [238] or random telegraph noise [239], or by random set processes in RRAM devices [222, 240].

The hybrid CMOS/RRAM structure of 4T1R synapse has some key advantages compared to previous approaches where SRDP was demonstrated by specific RRAM materials, such as Ag_2S [231], Ag/AgInSbTe/Ag [138], Pt/FeO_x/Pt [139], Al/TiO_{2-x}/AlO_x/Al [140], and Ag/SiON [146]. In particular, 4T1R synapse relies on memory-grade RRAM technology with fast switching, long endurance and long-term retention, which might be used in a multipurpose system-on-chip (SoC) for several functions, including embedded nonvolatile memory for code/data storage, generation of random keys for hardware security functions, such as a physical unclonable function (PUF) [223], and neuromorphic synapse/neuron circuits for on-chip cognitive computation.

3.3.2 Potentiation and depression in 4T1R synapse

Potentiation at high PRE-spike frequency

Synapse potentiation takes place at high frequency of PRE spiking, as shown by Fig. 3.8(a). In fact, if the PRE frequency is higher than $\Delta t_D{}^{-1}$

(f_{PRE} > $\Delta t_D{}^{-1}$), there is a strong probability for the gate of M_1 (activated by a spike at time t) and the gate of M_2 (activated by a previous spike delayed by Δt_D) to be stimulated at the same time. The repeated and simultaneous activation of M_1 and M_2, forming a NAND gate, results in current spikes which are integrated in the I&F circuit and finally cause fire. The fire spike is then delivered to the TE of RRAM such that the overlapping spikes at M_1, M_2 and TE induce a set process of the resistive device, hence a LTP event. Note that the positive fire spike is also applied to the gate of M_4 after inversion, which deactivates the M_3/M_4 branch.

In summary, a high PRE spiking frequency causes LTP through the M_1/M_2 branch. This result supports the need for a triplet of spikes (PRE-PRE-POST) to induce a frequency dependent potentiation of a synapse [134, 135].

Depression at low PRE-spike frequency

As shown in Fig. 3.8(b), PRE spiking at low frequency ($f_{PRE} \ll \Delta t_D{}^{-1}$) cannot activate the NAND-type M_1/M_2 branch, thus LTP cannot take place. On the other hand, random noise spikes from the PRE and the POST can simultaneously activate M_3 and M_4, respectively, as shown in Fig. 3.8(c). Since the negative POST noise is applied to the TE, the simultaneous noise spiking of the PRE and POST leads to a stochastic reset process of the RRAM device, hence synaptic LTD event. As a result, the SRDP undergoes synapses LTP or LTD depending on the competition between the spike-controlled activation of the M_1/M_2 and the M_3/M_4 branch, respectively [129].

Note that the 2-branch, 4T1R structure might be relatively expensive from the viewpoint of area consumption, *e.g.*, compared to 1T1R synapses [127] and 2T1R synapses [119] for STDP. However, this is the minimum structure to serve the function of online potentiation/depression from rate-coded spiking information.

3.3.3 SRDP characteristics

The potentiation/depression dynamics of the 4T1R synapse was studied by individually testing each branch by an integrated 2T1R structure, consisting of 2 transistors and a HfO_2 RRAM device in series [241].

The bipolar-switching RRAM used in these experiments had a Ti TE and a TiN BE. The active material was Si-doped HfO_2 deposited with an amorphous phase. The Ti top electrode also plays the role of creating an oxygen exchange layer (OEL), by inducing an oxygen vacancy rich layer by

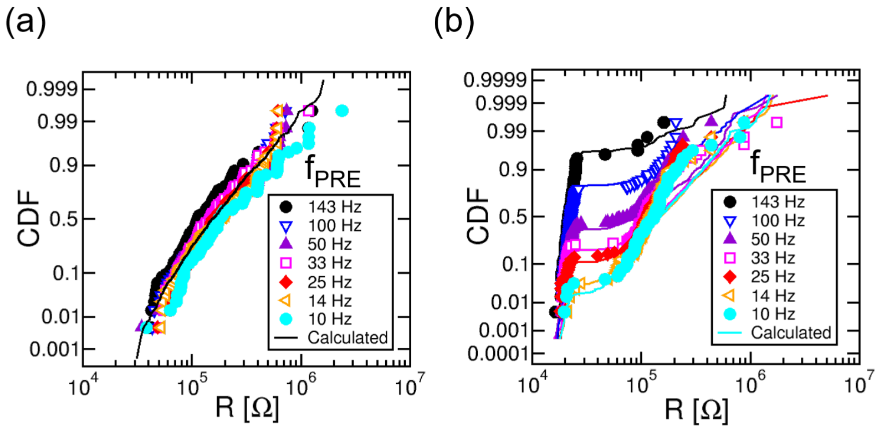


Figure 3.9: *Measured and calculated cumulative distributions of resistance R (a) before and (b) after learning and (c) measured and calculated average R for increasing f_{PRE}. (d) Number of overlapping PRE spikes activating M_1 and M_2 as a function of f_{PRE}. Copyright 2016 IEEE. Reprinted, with permission, from [129].*

oxygen gettering [241]. The TiN layer served as inert BE to prevent breakdown during the bipolar switching operation of the device. In addition, the size of transistors used in the structure was W/L = 3 μm/1.45 μm [242].

To demonstrate the synaptic potentiation induced by a high frequency PRE spiking, the LTP branch was characterized by applying a constant positive voltage of 2 V to the TE, while the gate of M_1 was stimulated by a train of random spikes with amplitude 3.2 V, pulse-width 1 ms and average frequency f_{PRE}. The same train was delayed by a time Δt_D = 10 ms, then applied to the gate of M_2. The M_2 pulse amplitude was also reduced to 1.6 V to limit the overall current to a compliance level I_C = 50 μA during set process for a controlled LTP. The RRAM device was prepared in a HRS of about 150 kΩ to check the LTP statistics during a 0.75-s-long training process with given value of f_{PRE}. The training experiment was repeated 1000 times on the same devices for each value of f_{PRE}.

Fig. 3.9 shows the measured and calculated distributions of R (a) before and (b) after each training process, for increasing f_{PRE}. The initial distribution in Fig. 3.9(a) corresponds to the initial HRS, as obtained by a reset pulse of -1.6 V applied to the TE with gate voltage 3.2 V applied to M_1 and M_2. The distributions in Fig. 3.9(b) after training show increasing fractions of LRS for increasing f_{PRE}, with average LRS resistance of 20 kΩ. In particular, note that the probability of set transition is high only

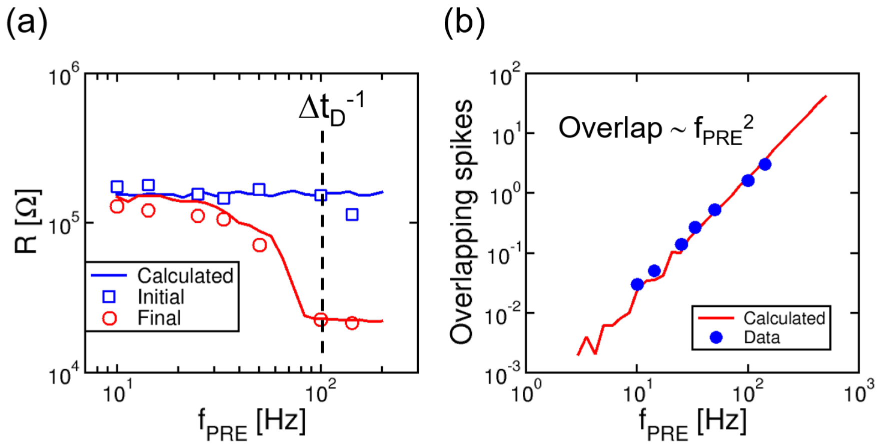


Figure 3.10: *(a) Measured and calculated average R for increasing f_{PRE}. (b) Number of overlapping PRE spikes activating M_1 and M_2 as a function of f_{PRE}. Copyright 2016 IEEE. Reprinted, with permission, from [129].*

for $f_{PRE} \geq$ 100 Hz, corresponding to an average time between 2 consecutive spikes of about Δt_D. Figs. 3.9(a) and (b) also show calculated distributions obtained by the stochastic model of RRAM device discussed in Sec. 2.4 [127], derived from the analytical model of bipolar RRAM discussed in Sec. 2.3 [66]. The distributions were accurately predicted by calculating the probability for spike overlap within the 0.75-s-long training sequence, and assuming a R-dependent variability for LRS and HRS [127]. Fig. 3.10(a) summarizes the results by showing the measured and calculated average R as a function of f_{PRE}. The transition to the LTP regime occurs abruptly for f_{PRE} = $\Delta t_D{}^{-1}$.

Note that the SRDP synapse works as a binary synapse, namely, the RRAM device in the 4T1R structure is always found in either LRS or HRS. This is due to the rather abrupt transitions of set and reset process in the adopted HfO_2 RRAM [127]. However, the adoption of RRAM devices with materials capable of gradual set/reset processes, such as PCMO [243] or TaO_x/TiO_x bilayers [244], might result in analog SRDP of the synapse, with advantages in terms of gray-scale learning [118].

These results can be understood by the increasing probability for spike overlapping at M_1 and M_2 for increasing f_{PRE}, as shown in Fig. 3.10(b). Both experiments and calculations show that the overlap probability increases with $f_{PRE}{}^2$, as expected for the joint probability of 2 independent spikes in the Poissonian train exciting the LTP branch at the same

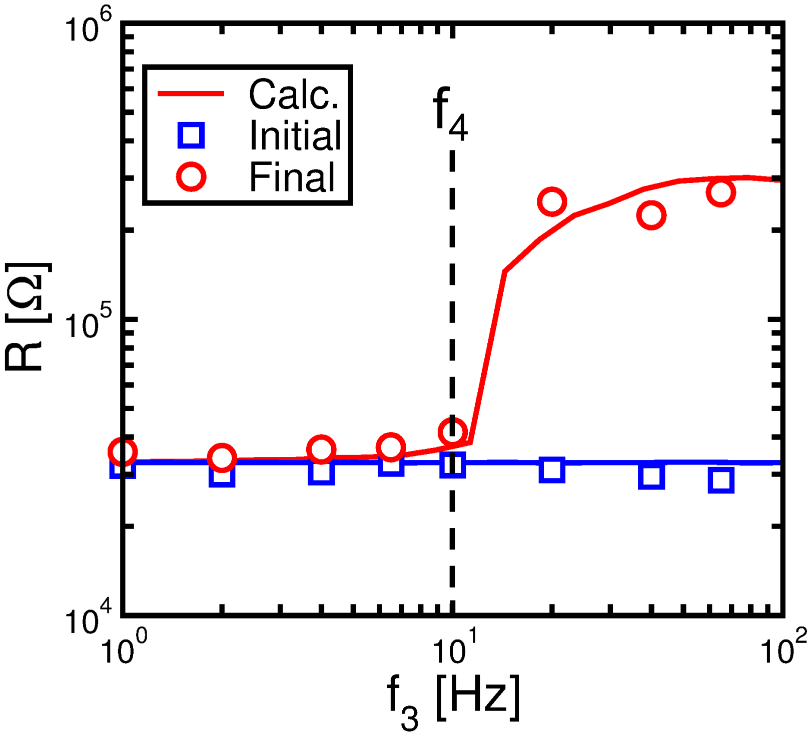


Figure 3.11: *Measured and calculated average R resulting from LTD branch characterization for increasing PRE noise frequency f_3 at fixed POST noise frequency f_4. Copyright 2016 IEEE. Reprinted, with permission, from [129].*

time [129].

To demonstrate LTD, the same 2T1R structure was tested by stimulating one transistor (M_3) by a spike train of amplitude 3.2 V at variable frequency f_3, while the other transistor (M_4) was stimulated by a spike train of amplitude 1.6 V and average frequency f_4 = 10 Hz. The same pulse sequence of the gate of M_4 was applied after inversion to the TE. This training sequence was maintained for 6000 epochs, equivalent to 6 s, and each experiment was repeated 5 times after preparing the device in the LRS. Fig. 3.11 shows the measured and calculated R as a function of f_3, indicating a transition to the LTD regime for $f_3 > f_4$, as the overlap probability becomes sufficiently large to allow for at least one reset transition [129].

Note that the particular choice of frequency operation for potentiation and depression is dictated by the analogy with biological systems, *e.g.*, experiments on synaptic plasticity *in vitro* [227]. Note however that, by tuning Δt_D, f_{PRE} and noise frequencies f_3 and f_4, it is possible to freely vary the

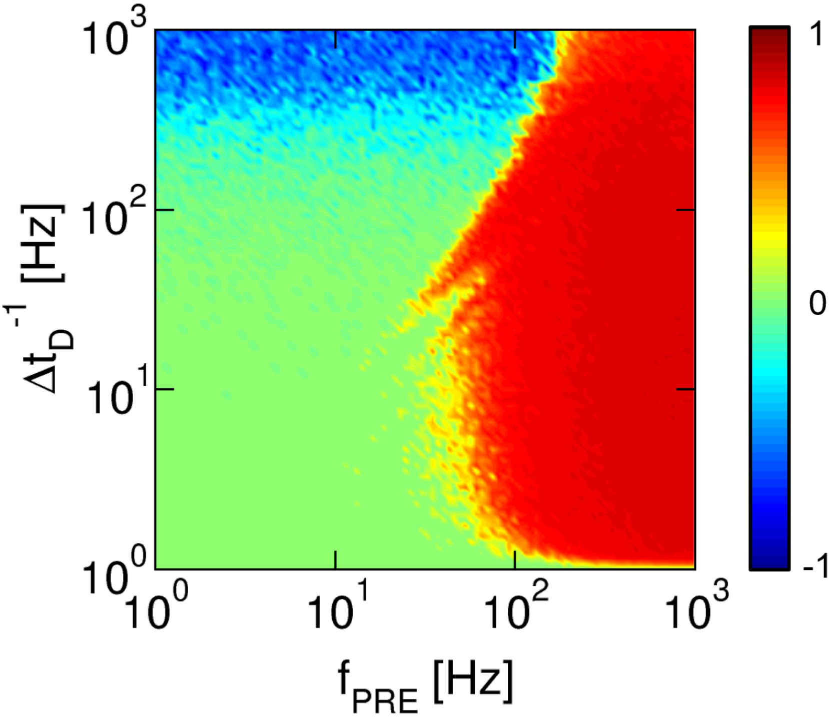


Figure 3.12: *Calculated color map of synapse conductance change R_0/R for variable f_{PRE} and $\Delta t_D{}^{-1}$ evidencing LTP (red), LTD (blue), and no weight change (green). Copyright 2018 IEEE. Reprinted, with permission, from [143].*

operation frequency, *e.g.*, for accelerated training of neural networks. The ultimate frequency for SRDP synapse is in the range of 1 GHz, because of limitations in the RRAM switching time of a fraction of ns [245, 246].

3.3.4 SRDP simulation in 4T1R synapse

To support the experimental study of 4T1R synapse, extensive simulations at level of single synapse device were carried out using the stochastic model of RRAM device [127]. All the calculated results were collected in a color map, reported in Fig. 3.12, showing synaptic weight change R_0/R as a function of f_{PRE} and the reciprocal of time delay $\Delta t_D{}^{-1}$ by settling an initial intermediate resistance R_0 = 100 kΩ and training time of 1 s. Ideally, LTP transition should take place for any $f_{PRE} \geq \Delta t_D{}^{-1}$ however, being the training time limited to 1 s, no conductance change is observed as f_{PRE} and $\Delta t_D{}^{-1}$ assume low values because no spike overlap events occur.

In addition, the map evidences that LTD transition can be also observed for f_{PRE} < $\Delta\mathrm{t}_D{}^{-1}$ provided that PRE and POST noise rates, both set to $\Delta\mathrm{t}_D{}^{-1}$/10, are sufficiently high.

3.4 Conclusions

In this chapter, the well known STDP rule has been demonstrated by a 1T1R RRAM synapse structure via experiments and simulations at device level. In addition to 1T1R synapse, another synapse circuit implementing SRDP rule that is considered as a fundamental learning rule in the human brain has also been presented. This hybrid synapse combines one RRAM device with 4 MOS transistors arranged in 2 NAND-type branches, serving the LTP and LTD functions in SRDP. Noise is used to induce LTD of synapses connected to neurons spiking at low frequency. Finally, to extensively investigate LTP and LTD in 4T1R synapse, experiments on integrated 2T1R structures and simulations at the level of device have been carried out supporting the feasibility of SRDP algorithm in hybrid CMOS/RRAM synapses.

CHAPTER 4

Feedforward spiking neural networks with 1T1R RRAM synapses for unsupervised pattern learning

4.1 Introduction

In last 60 years, the development of computing machines capable of human-like cognitive behavior, also known as artificial intelligence (AI), has been the subject of intensive research [247]. Specifically, human abilities including image recognition and classification [152], speech recognition [154], translation of sentences [155] and playing games such as AlphaGo [158, 159] have recently been demonstrated with outstanding accuracy via software-based deep neural networks (DNNs) trained on central processing units (CPUs) or graphic processing units (GPUs) accelerators of conventional computing platforms based on von Neumann architecture.

However, although AI has achieved very high levels of performance in this class of machine/deep learning tasks [152], CMOS-based digital computers perform brain-inspired tasks inefficiently. This is first due to the large area and slow data movement resulting from physical separation of processing and memory units in von Neumann architecture. In addition, increasing complexity of emerging cognitive tasks demands a power dissipation much higher than the power consumption of biological brain, which is approximately 20 W [107], as a result of the radically different computing schemes used in conventional digital computers and brain, respectively.

In fact, while in digital computers information is transmitted through bits according to a clock signal at a very high frequency in the range of GHz, in the human brain information is transmitted through spikes sparsely emitted by biological neurons at a very low frequency of 10 Hz [82, 107]. As a result, this event-driven computing scheme makes the brain operation extremely energy-efficient since the energy consumption occurs only where and when the information is processed [6].

To achieve brain's energy efficiency and massive parallelism in hardware, a new class of material-based devices called memristors [17] has been intensively investigated in recent years for implementing artificial synapses in high-density hardware spiking neural networks [39, 107, 110]. Unlike synapse circuit implementations in CMOS technology which are very expensive in terms of area [4, 106], emerging memory devices such as PCM and RRAM feature a great potential to meet the hard challenge of replicating the massive synaptic density ($\approx 10^4$ synapses per neuron on average) and low power consumption of the brain thanks to their nanoscale size, tunable resistance and low current operation [114].

In addition to synaptic density, synaptic arrangement also plays a fundamental role in the brain to achieve learning. In this frame, taking inspiration from significant computational abilities shown by a simple 2-layer feedforward neural network model referred to as perceptron in selective learning and recognition of incoming sensory patterns [248, 249], the use of spiking neurons and memristive plastic synapses in feedforward neural networks has recently attracted an increasing interest leading to demonstrations of unsupervised learning and recognition of visual/auditory patterns at the level of both simulation [116, 121–128, 250] and hardware [118, 129–131].

In this chapter, which is based on the works [118, 127, 129, 232, 251], unsupervised learning and recognition of visual patterns is demonstrated by 2-layer feedforward neural networks equipped with 1T1R RRAM synapses capable of STDP in both simulation and hardware. In addition to single pattern learning, other fundamental cognitive functionalities such as on-line learning of two sequential patterns and on-line learning of multiple patterns are also demonstrated. Finally, the impact of noise on learning performance in a RRAM-based perceptron neural network is extensively investigated via experiments supported by simulations.

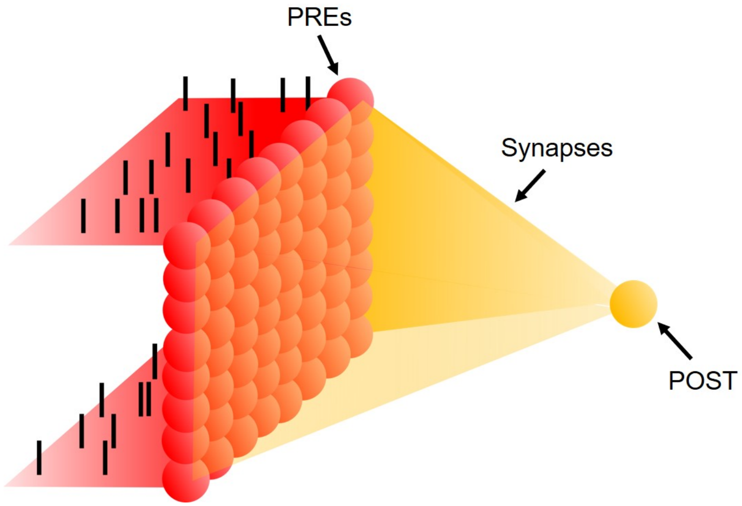


Figure 4.1: *Sketch of a two-layer perceptron neural network consisting of 64 pre-synaptic neurons (PREs) within the input layer and only 1 post-synaptic neuron (POST) within the second layer.*

4.2 Unsupervised learning of a single visual pattern

Learning ability in biological brain is considered to arise from changes in synapse strength driven by the spiking activity of neurons in neuronal networks [252].

Thus, to achieve a conclusive proof of concept for brain-inspired learning is not enough to implement learning rules such as STDP at the level of single synaptic element, as discussed in the Sec. 3.2, but rather demonstrations at network level are essential.

Here, a classical 2-layer feedforward neural network referred to as perceptron [248, 249] is designed and simulated to demonstrate unsupervised learning of visual patterns.

Fig.4.1 shows a schematic illustration of the simulated perceptron neural network consisting of a first layer, called pre-synaptic layer, with 64 pre-synaptic neurons (PREs) fully connected to the only one post-synaptic neuron (POST) of second layer by synapses implemented via 1T1R structures. To achieve unsupervised learning of a visual pattern, the network is operated according to a stochastic approach, namely, at any epoch consisting of a 10 ms time step, either the pattern "O" shown in Fig. 4.2(a), or random noise, as for instance one shown in Fig. 4.2(b), is alternatively submitted with equal probability of 50% to the PREs [127].

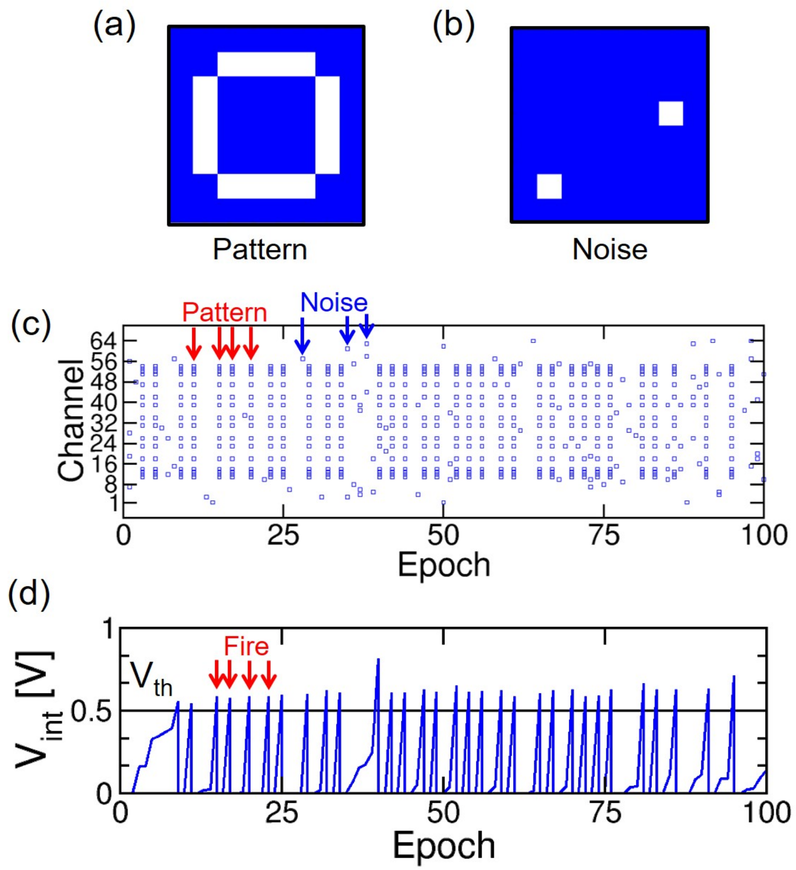


Figure 4.2: *Color plots of (a) pattern "O" and (b) an example of random noise image being submitted to the perceptron network for a single pattern learning simulation. (c) At any epoch, pattern and noise are alternatively presented with equal probability to PREs, (d) leading POST to fire as V_{int} hits the threshold V_{th}.*

In response to the stochastic presentation of pattern and noise shown in Fig. 4.2(c), PREs activate synaptic currents across the corresponding 1T1R synapses which are integrated by the I&F POST, thus causing an increase of POST internal potential V_{int} eventually crossing the fire threshold V_{th}. At that point, as indicated in Fig. 4.2(d), POST emits a fire which is backward delivered at TEs of 1T1R synapses by inducing the weight update within only selected synapses according to scheme shown in Fig. 3.3.

If a pattern submission induces a POST fire, namely $\Delta t > 0$ as described in Fig. 3.3(a), potentiation occurs at all synapses within the pattern. On the other hand, if the POST fire is followed by the presentation

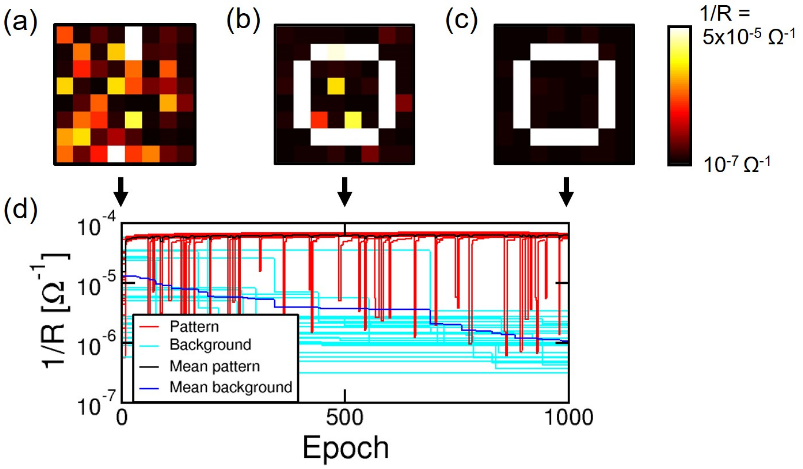


Figure 4.3: *Color plots of calculated synaptic weights during pattern learning at (a) epoch 0, (b) epoch 500, and (c) epoch 1000. (d) Time evolution of calculated synaptic weights evidencing pattern learning from an initial random weight configuration via a sudden potentiation of pattern synapses (red lines) and a gradual depression of background synapses (cyan lines) activated by random noise. Black line and blue line indicate the mean evolution of pattern and background conductance, respectively, supporting the different timescale between potentiation and depression process.*

of random noise, namely $\Delta t < 0$ as described in Fig. 3.3(b), depression occurs at synapses stimulated by random noise. Note that a refractory time was adopted for preventing the submission of the pattern at two consecutive epochs so that a pattern presentation can only be followed by a noise presentation leading to depression within background synapses by the sequence pattern-fire-noise [251].

As a result, pattern synapses are selectively potentiated, whereas all the other synapses, called background synapses, are stochastically depressed, thus supporting the network ability to learn the submitted visual pattern irrespective of the initial state of each synapse [118, 127].

Fig. 4.3 shows calculated learning of pattern "O" by perceptron neural network. The simulation was carried out using the stochastic Monte Carlo model of RRAM resistance distributions [127] discussed in the Sec. 2.4 to describe the set and reset transitions in binary 1T1R synapses. Specifically, potentiation (set transition) and depression (reset transition) were induced by POST spike pulses designed with amplitude V_{TE+} = 1.6 V and V_{TE-} = -1.6 V, respectively.

Starting from initial weights randomly distributed between HRS and

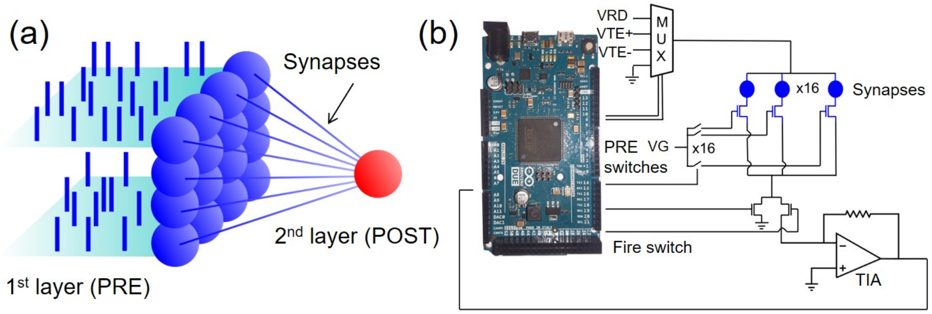


Figure 4.4: *(a) Schematic illustration of a 4x4 perceptron neural network with 16 PREs and 1 POST. (b) Circuit scheme of a 4x4 perceptron network including 16 PRE switches, an Arduino 2 microcontroller, a multiplexer and a transimpedance amplifier. Copyright 2018 IEEE. Adapted, with permission, from [251].*

LRS as shown in Fig. 4.3(a) by color plot of synaptic weights at epoch 0, pattern submission alternated with noise activates selective potentiation within pattern synapses and stochastic depression within background synapses (Fig. 4.3(b)) leading to pattern learning within 1000 epochs as shown by color plot in Fig. 4.3(c). Fig. 4.3(d) shows the detailed evolution of calculated synaptic conductance for increasing epoch, evidencing fast convergence of pattern synapses to high conductance values (within about 10 epochs) and a more gradual convergence of background synapses to low conductance values, thus supporting learning of a visual pattern via a perceptron network equipped with 1T1R synapses capable of STDP.

Note that pattern learning achieved in Fig. 4.3 for a small-scale perceptron network can be also demonstrated for larger perceptron networks properly tuning the threshold of POST according to pattern density, namely the number of the activated pixels divided the total number of pixels of the pattern image, and the density of submitted noise, namely the ratio between the average number of activated pixels within a random noise image and the total number of pixels of the image, to achieve the best trade-off between stability and speed of learning process [253].

To validate pattern learning functionality in experiments, a perceptron neural network consisting of 16 PREs, 16 synapses and 1 POST, schematically illustrated in Fig. 4.4(a), was realized in hardware [118, 251]. As shown by schematic circuit of the network in Fig. 4.4(b), PRE spikes are implemented via digital switches (PRE switches) controlled by the an Arduino Due microcontroller (μC) enabling the application of a voltage V_G to the gate of the 16 RRAM-based integrated synapses with 1T1R struc-

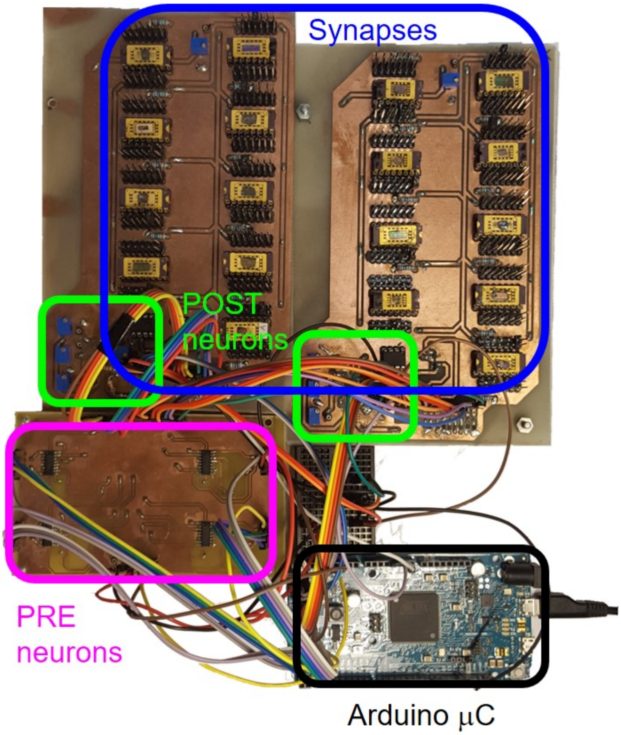


Figure 4.5: *Hardware implementation of a 4x4 perceptron network on a PCB. Copyright 2018 IEEE. Adapted, with permission, from [251].*

ture. The synaptic currents activated by PRE spikes are collected and sent to a transimpedance amplifier (TIA) to be converted into an analog voltage which is then submitted to the μC for digital integration. As integrated current exceeds the threshold, the μC generates the feedback spike driving a multiplexer (MUX) to provide the appropriate voltage to the TE of 1T1R synapses according to the scheme described in Fig. 3.3. In addition, Fig. 4.5 shows the printed-circuit board (PCB) realized to implement the 4x4 hardware perceptron spiking neural network [118, 251].

Fig. 4.6 shows an experimental demonstration of unsupervised learning of a 4x4 visual pattern representing a diagonal feature by the hardware spiking neural network. Similar to pattern learning simulations, a stochastic operation was adopted to train the network [118]. Fig. 4.6(a)-(d) shows the color plots of 16 synaptic conductances measured during the experiment. Starting from initial weights randomly prepared between LRS and HRS, the pattern/noise stochastic presentation to the PREs with the same probability

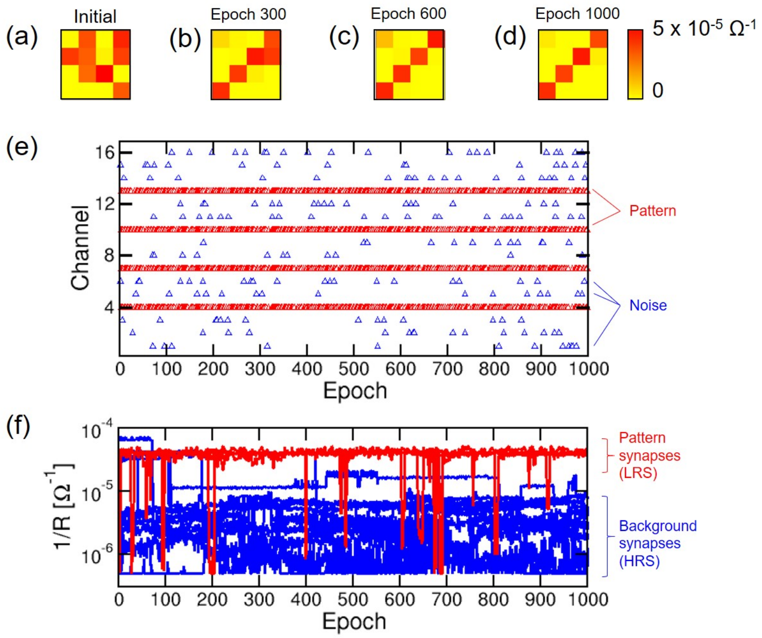


Figure 4.6: *Color plots of measured synaptic conductance during an experiment of unsupervised pattern learning via the hardware neural network in Fig. 4.5 at (a) epoch 0, (b) epoch 300, (c) epoch 600 and (d) epoch 1000. The measured synaptic weights at the end of experiment demonstrate the learning of submitted diagonal pattern. (e) Raster plot of spikes in each of 16 PRE channels resulting from stochastic submission of pattern and noise. (f) Time evolution of synaptic weights for pattern synapses (red lines) reaching high conductance values and background synapses (blue lines) reaching low conductance values. Reprinted from [118].*

of 50% leads the network to learn the submitted pattern within 1000 epochs via the potentiation of synaptic weights within the diagonal feature and the noise-induced depression of the other synapses within background.

Note that the noise stochastically alternated to the diagonal pattern at any epoch of the experiment (Fig. 4.6(e)) features a relatively low noise density of 3% to avoid that learning dynamics becomes unstable because of unwanted fires induced by noise input spikes [251].

Finally, Fig. 4.6(f) shows the evolution of measured synaptic weights with increasing epoch evidencing an abrupt transition to high conductance for pattern synapses and a more gradual transition to low conductances for background synapses as expected by simulation results. Both simulation and experiment thus support learning of visual patterns by perceptron neural networks with 1T1R synapses implementing STDP learning rule.

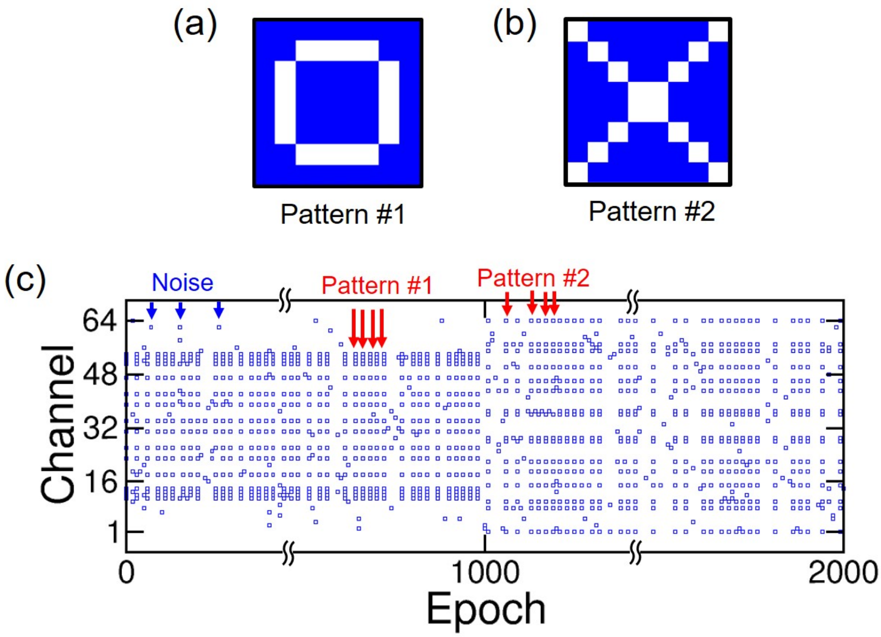


Figure 4.7: *Color plots of pattern #1 (a) and pattern #2 (b) submitted to an 8x8 perceptron network to demonstrate on-line pattern learning. (c) Raster plot of PRE spikes during learning evidencing the sequential application of two patterns. Pattern #1 is alternatively presented to noise during the first training phase from epoch 0 to epoch 1000, while pattern #2 is alternatively presented with noise during the second training phase from epoch 1001 to epoch 2000.*

4.3 Unsupervised on-line learning of sequential patterns

Another fundamental cognitive ability of human brain is to learn visual patterns in real time via synaptic weight adaptation. To test the on-line pattern learning, the perceptron network in Fig.4.1 was trained in simulation by a sequence of two 8x8 visual patterns shown in Fig. 4.7(a) and Fig. 4.7(b), representing the letters “O” and “X”, respectively. Fig. 4.7(c) shows the PRE spikes in response to pattern/noise presentation as a function of time evidencing the submission of the pattern #1 (“O”) stochastically alternated with random noise with equal probability for 1000 epochs followed by the presentation of the pattern #2 (“X”) stochastically alternated with random noise with equal probability for the following 1000 epochs.

After preparing the initial synaptic weights in a random resistance state between LRS and HRS (Fig. 4.8(a)), the presentation of pattern #1 and

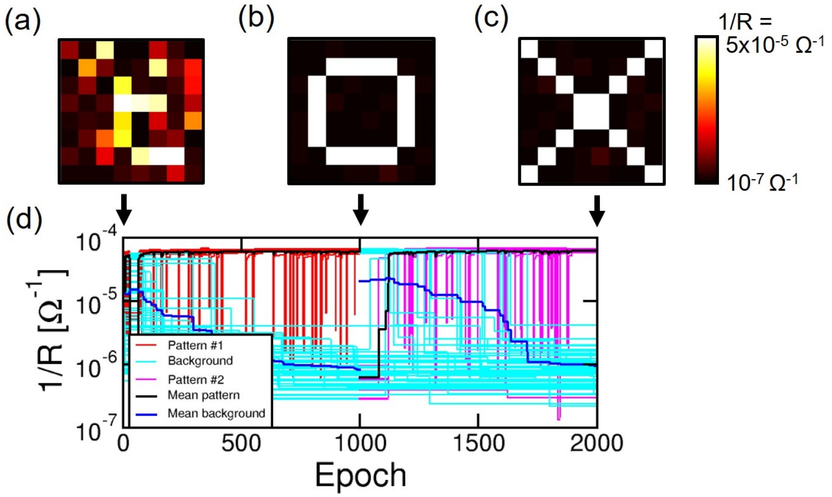


Figure 4.8: *Color plots of synaptic weights at (a) epoch 0, (b) epoch 1000, and (c) epoch 2000. (d) Evolution of calculated synaptic weights as a function of time evidencing the sequential learning of two patterns by a sudden increase of conductance for synaptic weights within pattern #1 (red lines) and pattern #2 (magenta lines) combined with a slower depression of background synapses for each pattern (cyan lines). Black and blue lines indicate the mean evolution of conductance within pattern and background during both training phases, respectively, thus supporting on-line learning of "O" and "X".*

noise for 1000 epochs leads to the selective potentiation of synapses within pattern #1 and the stochastic depression of background synapses according to STDP rule, which results in the learning of "O" as shown in Fig. 4.8(b).

At epoch 1000, the submitted input pattern is changed from the pattern #1 to the pattern #2. As a result, synapses within the pattern #2 undergo potentiation while all the background synapses undergo depression (Fig. 4.8(c)), thus demonstrating the network ability to remove or "forget" the previously stored pattern and learn a new pattern applied in sequence by the on-line adaptation of synaptic weights based on STDP rule. Finally, Fig. 4.8(d) shows the time evolution of calculated synaptic conductance evidencing the abrupt dynamics of potentiation process leading pattern synapses to high conductances and the gradual dynamics of depression process within the background synapses reaching low conductances in both learning phases.

After achieving on-line pattern learning in simulation, this cognitive functionality was also experimentally demonstrated by the hardware neu-

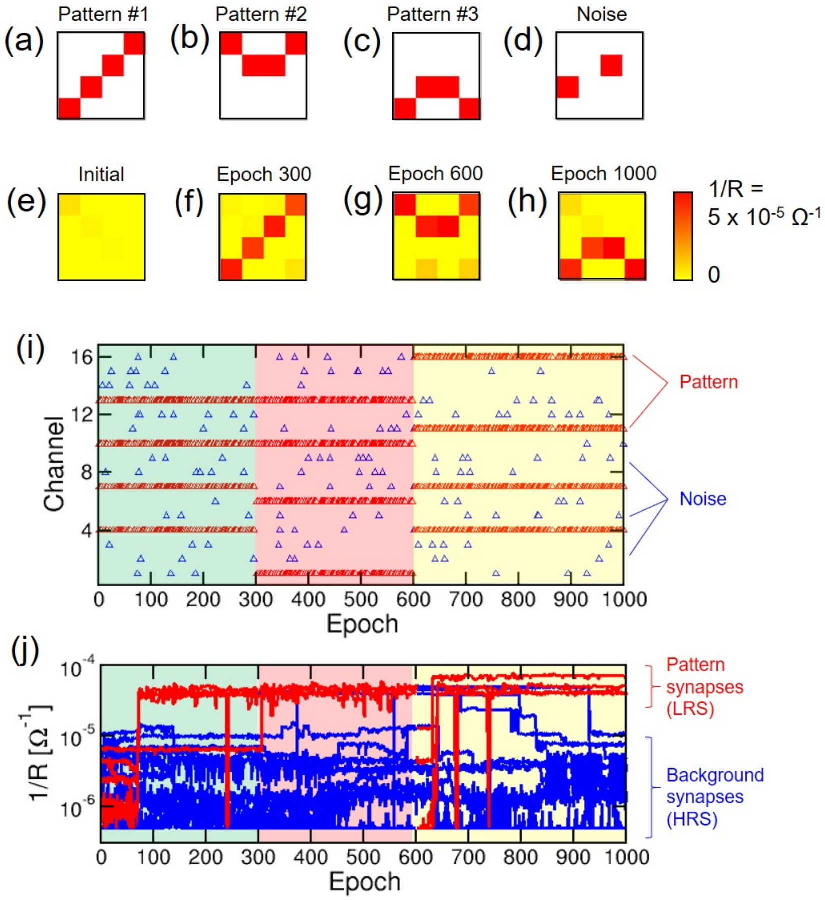


Figure 4.9: *Experimental demonstration of learning of 3 sequential patterns, namely (a) pattern #1, (b) pattern #2, and (c) pattern #3, stochastically submitted with (d) random noise. Color plots of (e) initial weights prepared in a random state between LRS and HRS and synaptic weights after (f) 300 epochs, (g) 600 epochs and (h) 1000 epochs. (i) Raster plot of PRE spikes and (h) time evolution of synaptic weights during whole training process of hardware network. Reprinted from [118].*

ral network shown in Fig. 4.5 [118]. Fig. 4.9(a)-(c) shows the 3 patterns which were sequentially presented to the PREs of perceptron network during the experiment, while Fig. 4.9(d) shows an example of noise which was stochastically alternated with each of patterns. Starting from synaptic weights in HRS (Fig. 4.9(e)), the network was externally stimulated by pattern #1 for initial 300 epochs (3 s), resulting in the potentiation of pattern synapses and depression of background synapses, as evidenced in

Fig. 4.9(f). In the following 300 epochs (epochs 301-600), the submitted pattern is changed from pattern #1 to pattern #2, thus leading the network to adapt to new pattern and remove the previous one via selective potentiation process and noise-activated depression process. Note that the percentage of randomly activated PRE channels within a noise image presented during this experiment is 3% on average. Finally, during the last 300 epochs of experiment (epochs 601-1000), pattern #2 was replaced by pattern #3 as input pattern leading to a new update of synaptic weights resulting in the stable learning of pattern #3 as evidenced in Fig. 4.9(h).

In addition, to capture more details of the on-line pattern learning experiment, Fig. 4.9(i) shows the raster plot of spikes within all the PRE channels evidencing the sequential presentation of the 3 patterns stochastically alternated with random noise while Fig. 4.9(j) shows the time evolution of measured conductance of synaptic weights evidencing a fast convergence to LRS for pattern synapses and a more gradual convergence to HRS for background synapses during each training phase. These experimental results thus corroborate unsupervised on-line pattern learning in harwdare perceptron networks equipped with RRAM-based 1T1R synapses capable of STDP.

4.4 Unsupervised on-line learning of multiple patterns

To match brain ability to learn multiple visual patterns simultaneously, the perceptron network used to demonstrate single pattern learning and on-line pattern learning in simulation was extended by the introduction of an additional POST within the second layer [118].

As shown in Fig. 4.10(a), all the 64 PREs are connected to each of the two POSTs, called POST #1 and POST #2, respectively, through a single excitatory synapse with 1T1R structure. Also, POST #1 and POST #2 are mutually connected by two non-plastic lateral inhibitory synapses playing a crucial role for the operation of this network. As POST #1 fires in response to the presentation of an input pattern, a spike is sent from POST #1 to POST #2 through an inhibitory synapse to decrease the internal potential of POST #2 by a certain fixed amount, thus preventing POST #2 to specialize on the pattern causing POST #1 fire. Similarly, as POST #2 fires in response to the submission of a pattern, a spike is sent from POST #2 to POST #1 by the other inhibitory synapse to reduce its internal potential by the same percentage. This mechanism, referred to as winner-takes-all (WTA), thus enables to maximize storage capacity of perceptron network inducing each POST to specialize on only one of submitted patterns [254].

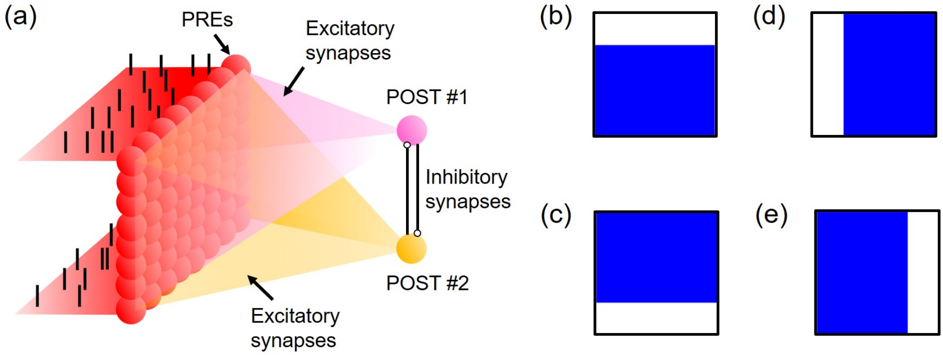


Figure 4.10: *(a) Schematic illustration of a perceptron network with 64 PREs fully connected to each of 2 POSTs by excitatory synapses. To maximize storage capacity of the network, the POSTs are mutually connected by two lateral inhibitory synapses implementing winner-takes-all (WTA) scheme. Illustration of 8x8 (b) top row and (c) bottom row visual patterns submitted to the network during first 1000 training epochs, and (c) left column and (d) right column presented in the following 1000 training epochs to achieve learning and recognition of multiple patterns.*

To achieve on-line learning of multiple patterns in simulation, the perceptron network in Fig. 4.10(a) was first trained for 1000 epochs providing two 8x8 visual patterns representing (b) top and (c) bottom rows, respectively, alternated with noise. After epoch 1000, top and bottom rows were replaced with (d) left and (e) right columns, which were submitted alternatively to noise for additional 1000 epochs. Also, WTA scheme was implemented setting inhibitory synapses such that the internal potential of POST #1 was reduced by amount of 60% at fires of POST #2 and vice versa.

Fig. 4.11 shows the color plots of calculated excitatory synaptic weights connecting (a) PREs to POST #1 and (b) PREs to POST #2 at epoch 0, epoch 1000, and epoch 2000, thus supporting the ability of the simulated network to learn separately the submitted patterns during each training phase via implementation of WTA scheme. Specifically, note that POST #1 and POST #2 can specialize on one or the other pattern with equal probability via WTA algorithm.

Fig. 4.11(c) and (d) also shows the evolution of calculated conductance of pattern and background synapses connecting PREs to POST #1 and POST #2, respectively, as a function of epochs evidencing the specialization of each POST on one of 2 submitted patterns in both training phases via selective potentiation at pattern synapses and noise-induced depression

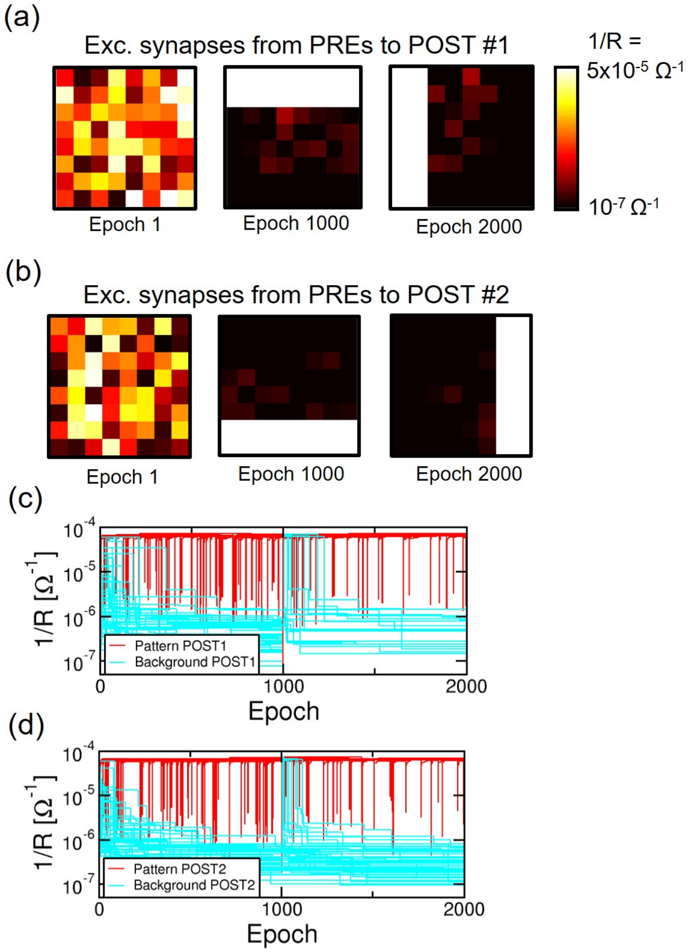


Figure 4.11: *Color code representation of calculated synaptic weights from PREs to (a) POST #1 and (b) POST #2 evidencing the network ability to learn the top/bottom row by POST #1/POST #2 within epoch 1000 and the left/right column by POST #1/POST #2 within epoch 2000 starting from random weight configurations. Time evolution of calculated conductance in pattern (red lines) and background (cyan lines) synapses during two sequential training phases for (c) POST #1 and (d) POST #2.*

at background synapses.

To validate these simulation results, this task was experimentally demonstrated by the hardware implementation of the network schematically illustrated in Fig. 4.12(a) in the case of 3x3 input patterns [251]. POST #1 and POST #2 are each connected to the 3x3 PRE layer via 9 1T1R synapses ca-

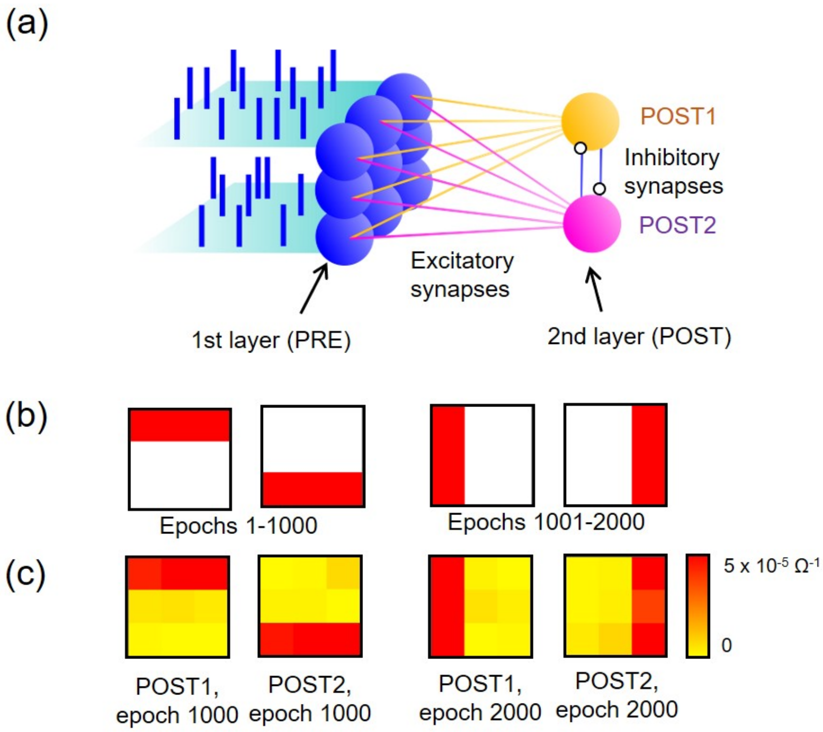


Figure 4.12: *(a) Schematic illustration of perceptron network with a 3x3 PRE layer and 2 POSTs used to demonstrate learning of multiple patterns in hardware. (b) Color plots of the pairs of 3x3 visual patterns sequentially submitted to train the network. (c) Color plots of measured synaptic weights connecting PREs to POST #1 and POST #2 at the end of first training phase (epoch 1000) and second training phase (epoch 2000) evidencing network ability to learn both submitted patterns during each phase. Copyright 2018 IEEE. Reprinted, with permission, from [251].*

pable of STDP. Also, to avoid learning of the same pattern by the 2 POSTs, POST #1 and POST #2 were controlled by the μC to implement the WTA optimization scheme. Fig. 4.12(b) shows two 3x3 patterns (top row and bottom row) submitted during the first 1000 training epochs and the following pair of 3x3 patterns (left column and right column) submitted in the following 1000 training epochs. Fig. 4.12(c) shows the color plots of measured synaptic conductances achieved at the end of each training phase evidencing the capability of hardware neural network to learn the first two submitted patterns within epoch 1000, with POST #1 and POST #2 specialized on top row and bottom row, respectively, and the following pair of presented patterns within epoch 2000, with POST #1 specialized on the left

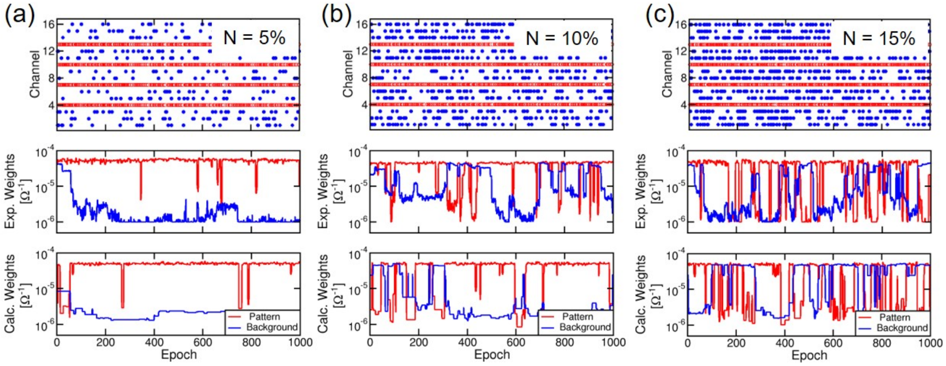


Figure 4.13: *(top) Raster plot of PRE spikes, (center) measured synaptic weights of pattern and background synapses and (bottom) calculated synaptic weights of pattern and background synapses during 1000-epoch-long learning process for noise densities (a) N = 5%, (b) N = 10%, and (c) N = 15% evidencing how too noise can make learning dynamics in the neural network strongly unstable. Copyright 2018 IEEE. Reprinted, with permission, from [251].*

column and POST # 2 on right column. These experimental results therefore support the feasibility of on-line learning of multiple patterns via WTA scheme in hardware neuromorphic networks equipped with 1T1R RRAM synapses.

4.5 Noise impact on pattern learning performance

As previously shown, noise presentation plays a crucial role to achieve unsupervised learning of visual patterns because it allows to implement the STDP depression condition ($\Delta t < 0$) within background synapses via the occurence of pattern-fire-noise sequences during learning process [127]. However, although the noise presentation is particularly beneficial for online learning because it enables the network to forget the previously learnt pattern, excessive noise can be detrimental for network performance. In fact, if the amount of noise submitted to the network is exaggerated, unwanted noise-fire-pattern sequences can occur with high probability during learning process inducing the depression of pattern synapses, thus making pattern learning dynamics unstable.

To investigate the impact of noise on learning performance, experiments and simulations were carried out using a 4x4 perceptron neural network [251]. Fig. 4.13 shows the sequence of PRE spikes submitted to the network (top), the time evolution of measured synaptic weights (center) and the time

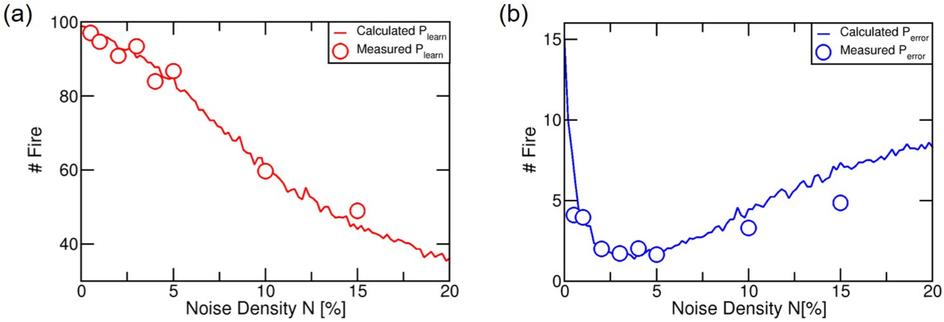


Figure 4.14: *(a) Measured and calculated P_{learn}, namely the probability to fire in response to the submitted pattern, and (b) P_{err}, namely the probability to fire in response to a noise submission, as a function of noise density N. Copyright 2018 IEEE. Reprinted, with permission, from [251].*

evolution of calculated synaptic weights (bottom) for three increasing values of noise density N, namely (a) N = 5%, (b) N = 10% and (c) N = 15%, indicating the ratio between the average number of on-pixels within a noise image and the total number of image pixels. These experimental results, supported by corresponding simulations, evidence that the stochastic presentation of random noise images of density N = 5% results in a stable learning dynamics whereas the presentation of random noise images with densities N = 10% and N = 15% leads to an increansingly unstable pattern learning because the submitted noise induces increasingly frequent noise-fire-pattern sequences leading to the abrupt depression of pattern synapses and potentiation of background synapses.

To capture the optimum value of N enabling stable pattern learning, learning probability P$_{learn}$ and error probability P$_{err}$ were measured and calculated for variable N training the network with a 4x4 diagonal pattern for 1000 epochs. While P$_{learn}$, which is defined as the probability of 'true fire', namely the probability of fire as a result of application of the diagonal pattern, was obtained for any N counting the number of true fires during whole training process, P$_{err}$, which is defined as the probability of 'false fire', namely the probability of fire as a result of application of any other 4x4 pattern different from the diagonal but with same pixel density of 25%, was obtained counting the number of false fires emitted by POST during whole training phase. As a result, Fig. 4.14 shows measured and calculated (a) P$_{learn}$ and (b) P$_{err}$ as a function of N ranging from 0% to 20%. On the one hand, increasingly low P$_{learn}$ values were achieved for increasing N supporting the results shown in Fig. 4.13. On the other hand, P$_{err}$ evi-

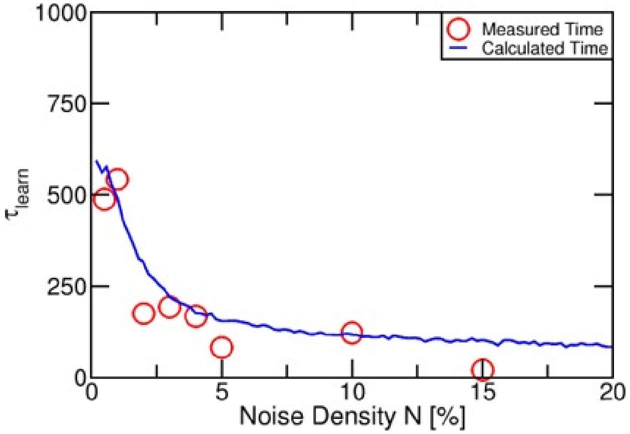


Figure 4.15: *Experimental data and calculated evolution of learning time τ_{learn}, defined as the time to depress the background synapses at a given resistance state, as a function of noise density N. Copyright 2018 IEEE. Reprinted, with permission, from [251].*

denced a minimum around N = N$_{opt}$ = 3%, thus indicating the best noise density value to achieve a stable pattern learning.

In addition, Fig. 4.15 shows the evolution of measured and calculated learning time τ_{learn} with increasing N. Since potentiation of pattern synapses has usually a very abrupt dynamics, τ_{learn} was defined as the time needed to depress background synapses at the resistance 66 kΩ. Both experimental and simulation results show that τ_{learn} decreases with increasing N according to a hyperbolic behavior, thus evidencing a trade-off with P$_{err}$ that unlike increases for N > 5%.

These results thus suggest that noise with a relatively low density capable of achieving a full background depression (not allowed for N = 0%) without inducing unstable learning is thus beneficial for optimizing pattern learning performance in perceptron neural networks.

4.6 Conclusions

In this chapter, learning and recognition of visual patterns has been achieved by simulations of 2-layer feedforward spiking neural networks using RRAM-based 1T1R synapses capable of STDP and experiments on a hardware perceptron network with 1T1R synapses. Both simulation and experimental results have evidenced the network ability to learn not only a single pattern but also more patterns submitted in sequence using a stochastic approach. Also, changing the network architecture by the introduction of another POST in the second layer, on-line learning of multiple patterns based on WTA scheme implemented via lateral inhibitory synapses between 2 POSTs has been demonstrated. Finally, a detailed study of noise effect on learning performance for a 4x4 perceptron network has been presented evidencing the need to suitably tune the density of submitted noise to achieve the best trade-off between learning stability and speed. These results thus support spiking neural networks equipped with hybrid CMOS/R-RAM synapses as promising building blocks for the development of scalable and energy-efficient hardware neuromorphic systems capable of brain-inspired computing.

CHAPTER 5

Feedforward spiking neural networks with 4T1R RRAM synapses for unsupervised pattern learning

5.1 Introduction

In recent years, the challenge to achieve brain-inspired cognitive functionalities in hardware has fueled a lot of curiosity toward the research field of neuromorphic engineering.

A key element in cognitive hardware systems is the ability to learn via bio-realistic plasticity rules, combined with the area scaling capability to enable integration of high-density neuron/synapse networks. To this purpose, RRAM devices have recently attracted a strong interest as potential synaptic elements in neuromorphic networks capable of replicating biologically plausible learning rules.

In this frame, motivated by Bienenstock-Cooper-Munro (BCM) theory [226] and biological experiments showing the dependence of synaptic plasticity on the rate of spikes emitted by pre- and post-synaptic neurons [133], spike-rate dependent plasticity (SRDP) learning rule was demonstrated at the level of single synapse using various types of RRAM devices in both

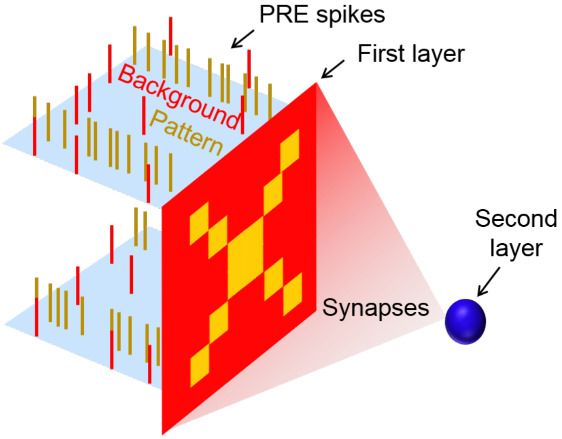


Figure 5.1: *Schematic illustration of a 2-layer perceptron neural network capable of pattern learning according to SRDP rule where high and low PRE spiking rates lead to pattern potentiation and background depression, respectively. Copyright 2018 IEEE. Reprinted, with permission, from [143].*

experiments and simulation [129, 138–140, 146, 231].

However, as already highlighted for STDP rule, to conclusively demonstrate brain-inspired unsupervised learning by SRDP, experiments and simulations of learning at the level of synaptic network are essential.

This chapter, which is based on the works [129, 143], covers the implementation of pattern learning at the level of network by 4T1R RRAM synapses capable of SRDP rule described in the Sec. 3.3. Pattern learning ability via SRDP is first tested by experiments and subsequently by simulation of a 2-layer feedforward neural network with 64 neurons for variable configuration of the initial synaptic weights. In addition to single pattern learning, on-line learning of 2 visual patterns submitted in sequence via SRDP is also addressed in simulation. Finally, the impact of noise frequency on learning efficiency is investigated for both pattern learning applications.

5.2 Experimental demonstration of pattern learning by SRDP

To prove the feasibility of unsupervised learning by SRDP at the level of synaptic network, SRDP synapses were used within a feed-forward perceptron neural network, where the input information is coded into the spiking frequency. Note however that the applicability of SRDP synapses is not restricted to a particular neuromorphic system or architecture. Indeed, SRDP

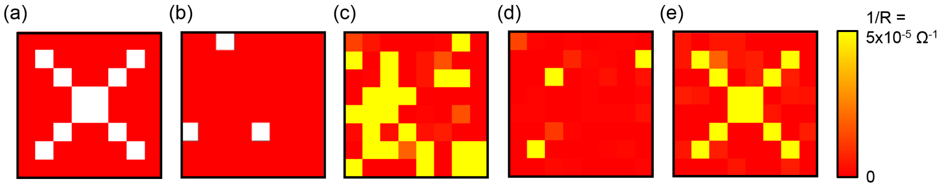


Figure 5.2: *Illustration of (a) the input pattern and (b) an example of random noise image submitted during the training process. Color plots of synaptic weights (c) initially prepared in a random state between LRS and HRS, (d) after LTD phase and (e) as a result of pattern presentation during the LTP phase of the training process. Copyright 2018 IEEE. Reprinted, with permission, from [143].*

synapses are generically suitable for the training of any spiking neural network, *e.g.*, feed-forward or recurrent networks, in presence of rate-coded spikes.

Fig. 5.1 depicts the perceptron network, where the PREs in the first layer generate spikes at high or low frequency, depending on their position being within or outside of a pattern, assumed to correspond to a reference image. The PRE spikes are submitted to a single POST in the second layer via SRDP synapses.

Thanks to the SRDP behavior, synapses in the pattern will experience LTP because of the high spiking frequency, whereas synapses in the background (*i.e.*, outside of the pattern) will undergo LTD due to the low PRE spiking frequency overwhelmed by random noise spiking. The SRDP algorithm was applied to integrated 2T1R structures used alternatively as LTD and LTP branches in the 4T1R synapse [143]. LTD and LTP operation schemes were applied for 1 s each on the same 2T1R structure. As a reference synaptic network, an array of 8x8 SRDP synapses that were initially prepared in a random state with resistance between LRS and HRS levels was adopted.

Fig. 5.2(a) shows the visual pattern that was considered as input for image learning demonstration. The training procedure consists of 2 phases: in the first phase (LTD), random noise images such as the one in Fig. 5.2(b) were submitted for 1 s to all synapses to achieve LTD.

Starting from the initial synaptic weight distribution in Fig. 5.2(c), the first training phase resulted in LTD as demonstrated by the HRS weights in Fig. 5.2(d). In the second phase, the LTP mode was adopted by stimulating background and pattern synapses with random spikes at low frequency (f_{PRE} = 5 Hz) and high frequency (f_{PRE} = 150 Hz), respectively,

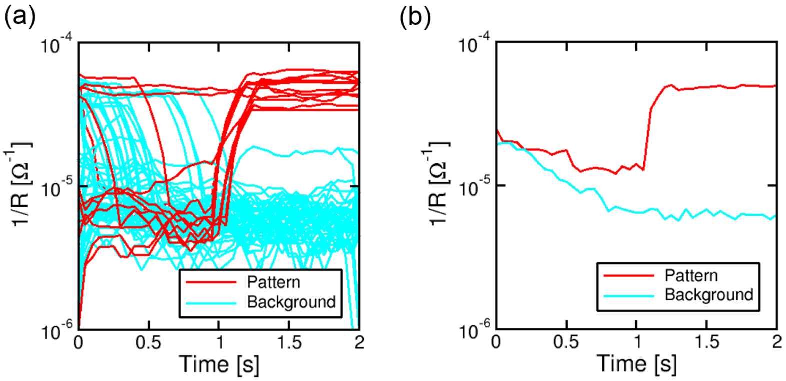


Figure 5.3: *(a) Time evolution of measured pattern (red) and background (cyan) conductance showing synaptic LTD within 1 s due to PRE and POST noise spiking and the selective potentiation of synapses in the pattern because of high frequency PRE stimulation during the following 1-s-long LTP phase. (b) Mean evolution of measured pattern and background synaptic weights as a function of time supporting background depression and pattern potentiation. Copyright 2018 IEEE. Reprinted, with permission, from [143].*

for 1 s. The final weight distribution in Fig. 5.2(e) demonstrates learning of the submitted pattern thanks to the spiking frequency being higher than $\Delta t_D{}^{-1}$, where Δt_D = 10 ms.

Fig. 5.3(a) shows the measured synaptic weights 1/R as a function of time during the 2 phases of training. In the first period, both pattern and background synapses approach low weight due to noise induced stochastic LTD. In the second period, synaptic weights in the pattern increase due to LTP process induced by SRDP, while background synapses remain at a low conductance due to low frequency spiking. Fig. 5.3(b) shows the corresponding average synaptic weights for the pattern and the background as a function of time, clearly indicating LTD and LTP phases.

5.3 Simulation study of pattern learning by SRDP

5.3.1 Single pattern learning

To further corroborate SRDP pattern learning by 4T1R synapse, the 2-layer perceptron network shown in Fig. 5.1 was simulated. The same 8x8 pattern of Fig. 5.2(a) was adopted for simplicity. Fig. 5.4(a) shows the sequence of spikes submitted at each of the 64 channels, evidencing different spiking frequencies at the pattern (f_{PRE} = 100 Hz) and background (f_{PRE} = 1 Hz).

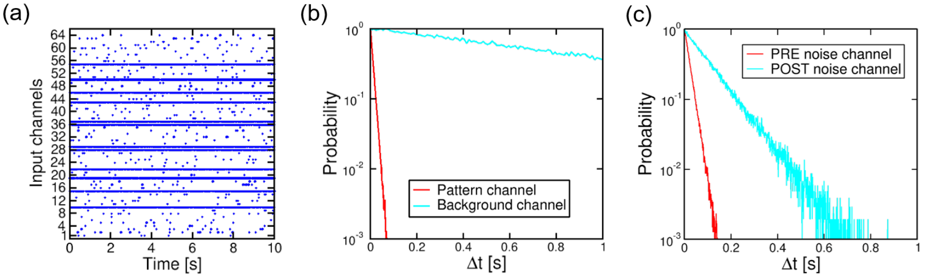


Figure 5.4: *(a) PRE spikes as a function of time showing high and low frequency stimulation for pattern and background input channels, respectively. Distributions of time intervals between two consecutive spikes for (b) pattern/background channels and (c) PRE/POST noise channels. Copyright 2018 IEEE. Reprinted, with permission, from [143].*

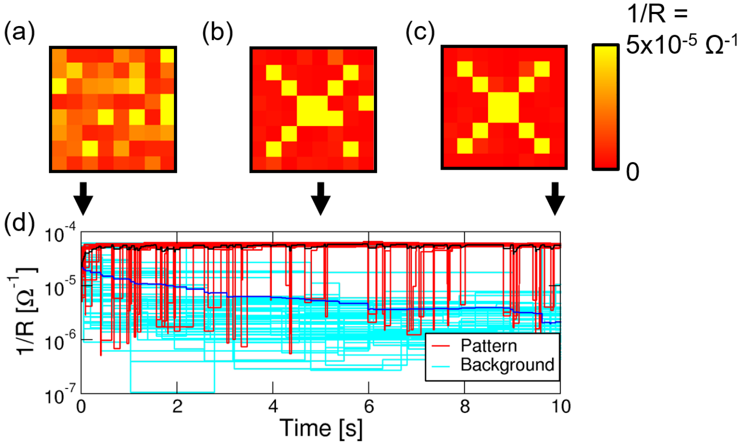


Figure 5.5: *Color plots of synaptic weights at times (a) t = 0 s, (b) t = 5 s and (c) t = 10 s. (d) Time evolution of calculated synaptic weights initialized in a random state between LRS and HRS levels. The evolution of conductances as a function of time evidences fast potentiation of pattern synapses (red) and a slower depression of background synapses (cyan). Black and blue lines indicate time evolution of mean pattern and background synapses, respectively. Copyright 2018 IEEE. Reprinted, with permission, from [143].*

Fig. 5.4(b) shows the distributions of time intervals between consecutive spikes for pattern and background, evidencing an exponential decrease with frequency which is typical of random Poissonian events. Fig. 5.4(c) shows the distribution of inter-spike times for PRE and POST noise spiking with rate f_3 = 50 Hz and f_4 = 10 Hz, respectively.

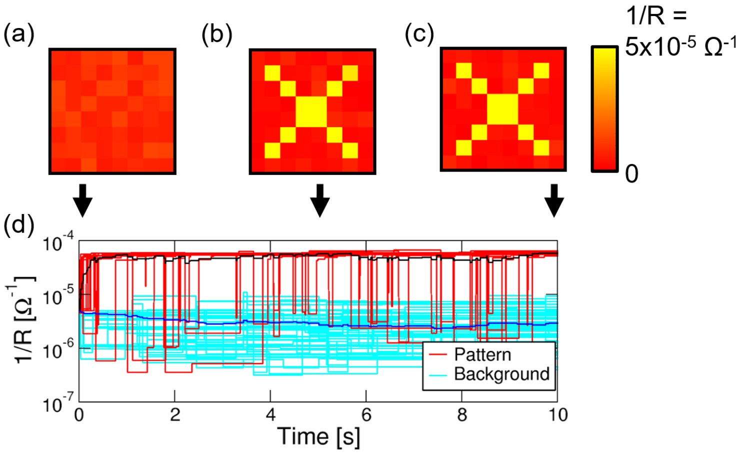


Figure 5.6: *Color plots of synaptic weights at times (a) t = 0 s, (b) t = 5 s and (c) t = 10 s. (d) Evolution of calculated synaptic weights as a function of time starting from initial HRS weights. Synaptic evolution reveals a very fast pattern learning since background is already fully depressed. Copyright 2018 IEEE. Reprinted, with permission, from [143].*

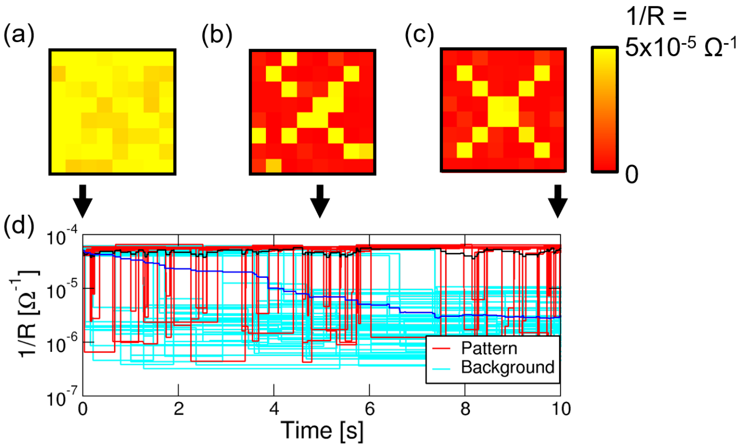


Figure 5.7: *Color plots of synaptic weights at times (a) t = 0 s, (b) t = 5 s and (c) t = 10 s. (d) Time evolution of calculated synaptic weights, which are initially prepared in LRS state, evidencing a slower pattern learning in comparison with the previous two cases because all background synapses need to be depressed. Copyright 2018 IEEE. Reprinted, with permission, from [143].*

Fig. 5.5 shows the calculated synaptic weights in a color plot at times (a) 0 s, (b) 5 s, (c) 10 s, and the detailed time evolution of the calculated 1/R during the whole training process. Initial weights are uniformly distributed between LRS and HRS. Pattern synapses are potentiated within about 1 s from the start of training, while background synapses approach low weight more slowly, as the noise spiking activity has lower frequency compared to f_{PRE} in the pattern. Note that pattern synapses may be temporarily disturbed from their high weight due to stochastic noise. This disturb was quantified in a probability of 1% for pattern synapses to have low weight during training, under the conditions of this simulation.

Also, synaptic weights were calculated as a function of time during training under the same conditions as Fig. 5.5, except the initial distribution being prepared in HRS (Fig. 5.6) or LRS (Fig. 5.7). In the first case, learning only requires LTP of pattern synapses, whereas in the second case complete learning requires LTD of the background synapses, thus requires longer time [143].

5.3.2 Impact of noise on learning efficiency

Noise plays a leading role in SRDP by inducing LTD. On the other hand, noise affects all synapses at the same extent, thus may also disturb pattern learning. To study the impact of noise on learning, the efficiency of perceptron network was evaluated as a function of PRE noise frequency f_3 and POST noise frequency f_4. The learning efficiency was evaluated by calculating the learning probability P_{learn}, defined as the probability of POST fire in response to the submission of the pattern after the training stage, and error probability P_{error}, defined as the probability of POST fire in response to the submission of an input random noise [117, 127]. The pattern in Fig. 5.2(a) was used for the training phase, which lasted 5000 epochs, equivalent to 5 s. Fig. 5.8 shows the calculated (a) P_{learn} and (b) P_{error} in a color plot as a function of f_3 and f_4. P_{learn} becomes very close to 1 as either f_3 or f_4 decreases, thus making noise disturbance negligible. As f_3 and f_4 increase, P_{learn} decreases because noise spikes make learning process strongly unstable. On the other hand, P_{error} shows the opposite behavior, as a low noise rate induces no LTD, thus any random noise may excite synapses in the LRS and cause false fire. A high noise frequency instead causes strong LTD and suppression of false fires, although true fires are also affected. The noise rates were identified for best tradeoff between efficient learning and low false fires, which can be found along the curve with constant geometric average $\sqrt{f_3 f_4}$ = 40 Hz.

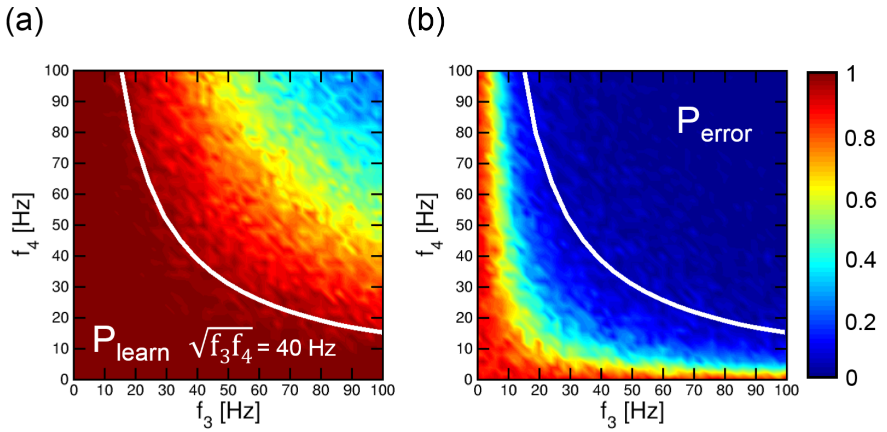


Figure 5.8: *Calculated color maps showing the effect of PRE and POST noise average frequencies f_3 and f_4 on (a) learning probability and (b) error probability of "X" pattern via a perceptron-like neural network with RRAM-based synapses capable of SRDP. Optimal performance is achieved if f_3 and f_4 satisfy the tradeoff relation described by indicated curve. Copyright 2018 IEEE. Reprinted, with permission, from [143].*

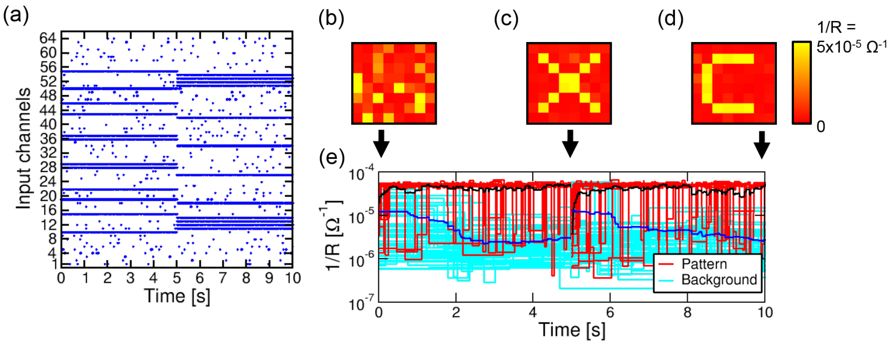


Figure 5.9: *(a) Raster plot of PRE spikes evidencing the change of input pattern at time 5 s. Color plots of synaptic weights at (b) t = 0 s, (c) t = 5 s and (d) t = 10 s during learning of a sequence of images with PRE and POST noise spiking rates equal to 50 Hz and 20 Hz, respectively. (e) Time evolution of synaptic weights showing a fast potentiation of "X" weights and a gradual depression of background synapses within 5 s. At 5 s, the "X" pattern is replaced with the "C" pattern and all weights adapt to new submitted pattern according to SRDP learning rule. Copyright 2018 IEEE. Reprinted, with permission, from [143].*

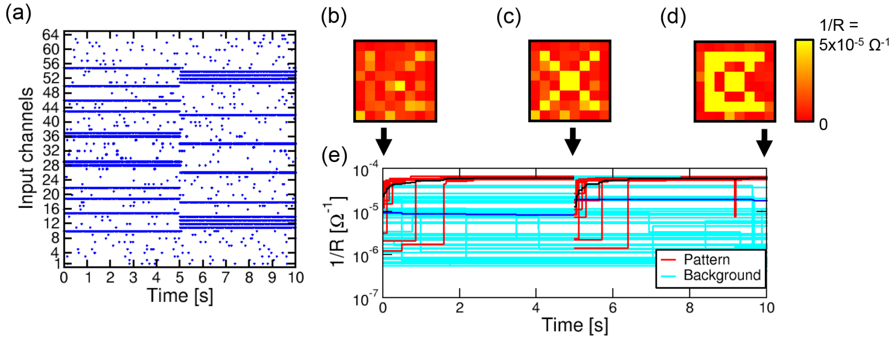


Figure 5.10: *(a) Raster plot of PRE input spikes due to sequential patterns. Color plots of synaptic weights at (b) t = 0 s, (c) t = 5 s and (d) t = 10 s during an online learning process with PRE and POST low frequency noise spiking at 10 Hz and 5 Hz, respectively. (e) Time evolution of synaptic weights evidencing final potentiation of synapses in both patterns since the first stored pattern "X" cannot be erased without sufficiently strong noise activity. Copyright 2018 IEEE. Reprinted, with permission, from [143].*

5.3.3 On-line learning of sequential patterns

One of the advantages of bidirectional SRDP, *i.e.*, the availability of both LTP and LTD, is on-line learning, where the synaptic network learns the currently submitted pattern and is capable of erasing, or forgetting, any previously stored pattern [118, 234]. To support the capability of on-line learning, the presentation of 2 different patterns in sequence to the perceptron network was simulated. Fig. 5.9(a) shows the spiking sequence submitted by the PRE layer, including a first phase with pattern "X" for 5 s, followed by a second phase where pattern "C" was submitted for 5 s. The figure also shows the color maps of 8x8 synaptic weights at times (b) 0 s, (c) 5 s, and (d) 10 s, evidencing accurate learning of the submitted patterns. Fig. 5.9(e) shows the synaptic weights as a function of time, indicating convergence to LRS or HRS of pattern synapses or background synapses, respectively, in each phase. In particular, as pattern "X" starts being excited at low frequency at 5 s, the corresponding synapses are depressed by PRE and POST random noise spiking activities at 50 Hz and 20 Hz, respectively. Therefore, as the input pattern is changed, the neural network is capable of forgetting the first pattern to adapt to the second one by SRDP plastic 4T1R synapses, by properly tuned noise spiking activity. However, if the online learning process was carried out with too low PRE and POST noise spike rates equal to 10 Hz and 5 Hz, respectively, the PRE input spike trains shown in Fig. 5.10(a) would lead from initial random weights to the simultaneous potentiation of synapses within both “X” and “C” pattern (Fig. 5.10(b-d)), thus preventing a selective online adaptation of synaptic weights to the visual patterns submitted in sequence to the first one.

5.4 Conclusions

In this chapter, unsupervised pattern learning at network level by 4T1R synapses capable of SRDP has been demonstrated by an extensive simulation study supporting experimental measurements. First, learning of a single 8x8 visual pattern by a perceptron network with 64 neurons has been demonstrated irrispective of initial distribution of synaptic weights and investigated for variable PRE and POST noise frequencies to find the best operating condition of the network. In addition, on-line learning of two sequential 8x8 visual patterns has also been demonstrated evidencing the ability of simulated perceptron network to recognize in real time pattern and background within submitted images, and the need to properly tune PRE and POST noise frequency to forget the previously learnt pattern. These results thus support hybrid CMOS/RRAM integrated circuits as building blocks for low-power hardware neuromorphic systems with memristive devices.

CHAPTER 6

Recurrent spiking neural networks with 1T1R RRAM synapses for associative learning

6.1 Introduction

The human brain is a very complex biological system capable of achieving a highly parallel and error-tolerant processing of sensory information with a very low power consumption of about 20 W [4, 82, 107] thanks to its massively distributed architecture and energy-efficient computing strategy based on spike events [6, 82, 107].

However, such a computing scheme cannot be efficiently reproduced in current digital computers based on classical von Neumann architecture because of the slow and energy-hungry data shuttle between working memory and CPU [4, 6].

To tackle this strong limitation, emerging non-volatile memory devices, also known as memristors [16, 17], such as PCM and RRAM devices have been extensively investigated in last decade for their ability to combine storage and computation [22, 23], which enabled to replicate biological rules believed to be the origin of learning capability such as STDP and

SRDP [107, 108, 113–115, 143, 229]. In addition, their nanoscale size can allow to achieve in hardware a synaptic density comparable to the biological one, which is approximately 10^4 synapses per neuron [82, 114].

Although synaptic density plays a crucial role in brain computation, this also depends on the arrangement of synaptic connectivity.

In this context, experimental studies have evidenced that the regions mostly involved in storage and retrieval of memories such as the CA3 region of the hippocampus, include large networks of neurons with a recurrent connectivity pattern processing information via forward and backward spikes propagating across excitatory and inhibitory synapses.

Based on experimental observations [255] and theoretical studies [256], recurrent neural networks, also referred to as attractor networks [256, 257], including the well-known Hopfield network [258, 259], have been implemented in both hardware via fully-CMOS [90, 99, 260, 261] and hybrid CMOS/memristive [262–264] circuits, and simulation [125, 265, 266].

In this chapter, which is based on the works [232, 265, 267, 268], Hopfield spiking neural networks equipped with 1T1R RRAM synapses capable of STDP are designed and simulated. First, learning, recall and stability of attractor states via external stimulation and recurrent cooperation and competition are demonstrated by simulations. Based on these results, fundamental brain-inspired primitives such as associative memory, pattern restoration and error correction are explored by simulations.

6.2 Hopfield spiking neural network with 1T1R RRAM synapses

Hopfield network is a well-known recurrent neural network capable of implementing brain-inspired primitives such as content-addressable memory [258, 259] or solving difficult optimization problems such as the Traveling-Salesman Problem (TSP) [269] and constraint-satisfaction problems such as the Sudoku puzzle [270] as a result of collective firing activity of large populations of elementary neuron units [258].

Fig. 6.1 shows a sketch of a Hopfield network with 6 fully connected neurons receiving external input stimuli X_i, with i = 1:6 [265].

To replicate this network scheme, the circuit implementation shown in Fig. 6.2 was designed and simulated [265]. Here, N = 6 neurons implemented by I&F blocks are fully connected by N·(N-1) = 30 RRAM-based 1T1R excitatory synapses (blue) and 30 inhibitory synapses (red) capable of STDP. Each neuron block N_i was designed with 2 inputs given by external asynchronous current spikes X_i and the sum of excitatory and inhibitory synaptic currents activated by other neurons, respectively, and 3 outputs,

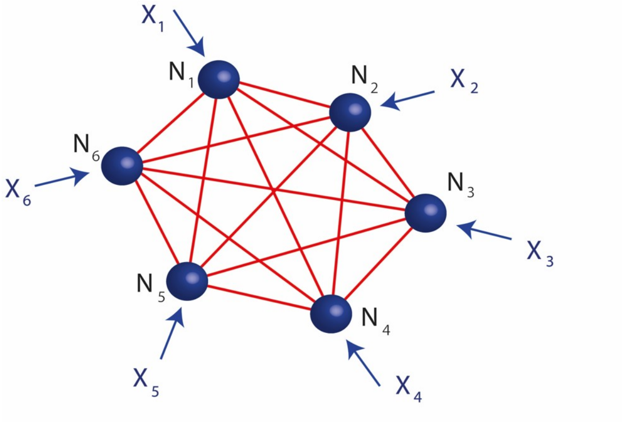


Figure 6.1: *Sketch of a Hopfield neural network comprising 6 neurons.*

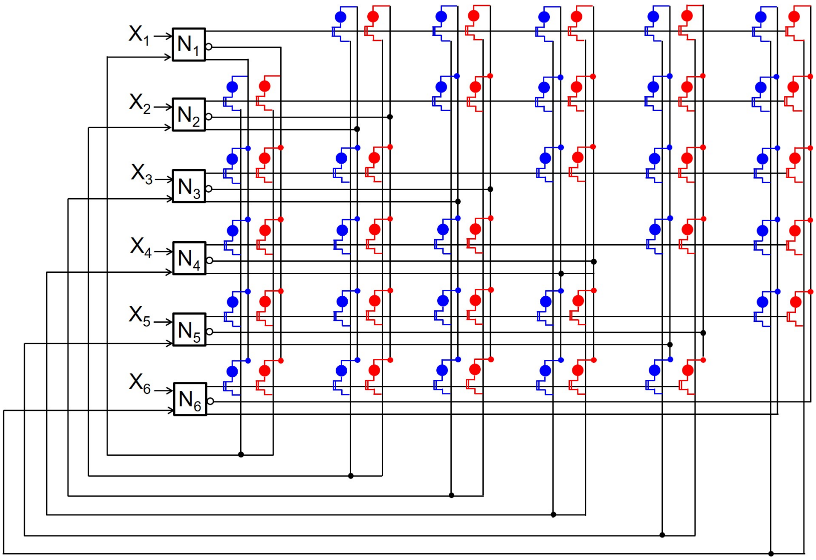


Figure 6.2: *Schematic illustration of the simulated Hopfield network with 6 fully connected neurons, 30 excitatory synapses (blue) and 30 inhibitory synapses (red). Diagonal synapses W_{ii} and W'_{ii} (i = 1 to 6) are absent to prevent self-feedback. Each neuron behaves as both pre-synaptic neuron controlling the gates of all row excitatory/inhibitory 1T1R synapses and post-synaptic neuron controlling the top electrodes of excitatory/inhibitory 1T1R synapses of corresponding column. Copyright 2017 IEEE. Reprinted, with permission, from [265].*

namely (i) G_i which is applied to the gate of 1T1R synapses of the i-th row, (ii) O'_i which is applied to the TE of inhibitory 1T1R synapses of i-th col-

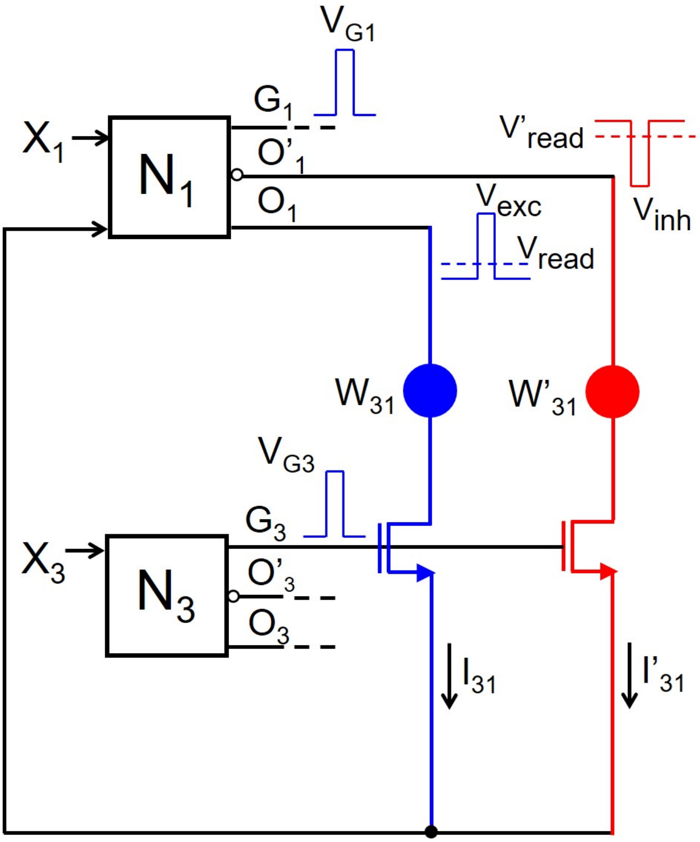


Figure 6.3: *Schematic illustration of the synaptic dynamics in the RRAM-based Hopfield network: as N_3 fires, the gate spike V_{G3} induces a positive (negative) current at synapse W_{31} (W'$_{31}$). If N_1 is coactive with N_3, the overlapping spikes lead to W_{31} potentiation and W'$_{31}$ depression. Copyright 2017 IEEE. Adapted, with permission, from [265].*

umn, and (iii) O_i which is applied to the TE of excitatory 1T1R synapses of i-th column. Thus, each N_i acts as both PRE and POST controlling the gate of i-th row excitatory/inhibitory 1T1R synapses and the TE of i-th column excitatory/inhibitory 1T1R synapses, respectively [265, 267]. Also, according to original Hopfield network model [259], note that the synapses along the diagonal of synaptic matrix are absent (W_{ii} = 0) to prevent that self-feedback leads the network to unstable states.

Fig. 6.3 describes the operating principle of this hybrid CMOS/RRAM Hopfield network to achieve the storage of a stable memory state or attractor state considering the pair of neurons N_1 and N_3 and their mutual excitatory and inhibitory synaptic connections W$_{31}$ and W'$_{31}$.

As N$_3$ fires, the gate voltage spike G$_3$ activates all the excitatory/inhibitory synaptic gates within the 3rd row, thus inducing excitatory and inhibitory currents, including the excitatory current I$_{31}$ and the inhibitory current I'$_{31}$, since the top electrodes of excitatory and inhibitory synapses are biased by read voltages V$_{read}$ and V'$_{read}$ = -V$_{read}$, respectively. These currents are recurrently fed to and integrated with the external current spike X$_1$ by N$_1$, eventually contributing to N$_1$ fire. If N$_1$ and N$_3$ fire at the same time, the overlap of voltage spikes G$_3$ and O$_1$, which has a positive amplitude V$_{exc}$ higher than V$_{set}$, causes the synaptic potentiation of the excitatory 1T1R synapse of weight W$_{31}$, while the overlap of voltage spikes G$_3$ and O'$_1$, which has a negative amplitude V$_{inh}$ exceeding V$_{reset}$, causes the synaptic depression of the inhibitory 1T1R synapse of weight W'$_{31}$.

Moreover, since N$_1$ and N$_3$ are both PRE and POST, the symmetric excitatory and inhibitory weights, namely W$_{13}$ and W'$_{13}$, are also themselves potentiated and depressed, respectively, thus leading to the typical symmetric configuration of synaptic weights featuring Hopfield neural networks [258].

In this implementation, the synaptic weight updates triggered by coactive neurons obey to the Hebbian postulate "*neurons that fire together wire together*" [271]. Specifically, potentiation of the excitatory synapses and depression of the inhibitory synapses between coactive neurons are achieved as simplified cases of the STDP in the 1T1R synapses described in Section 3.2, where the bipolar voltage pulse at the TE is replaced by a unipolar voltage pulse, with either positive voltage V$_{exc}$ or negative voltage V$_{inh}$.

6.3 Learning and recall of a single attractor

6.3.1 Learning of a single attractor

Based on network operation schematically described in Fig. 6.3, the hybrid CMOS/RRAM Hopfield network can settle into or learn an attractor state providing external spikes at high frequency to sub-populations of neurons within the network.

Fig. 6.4 shows the time evolution of attractor learning process activated by the external stimulation of neuron sub-population N$_1$, N$_2$ and N$_3$ via Poisson spike trains of amplitude I$_x$ = 10 μA and frequency f$_x$ = 200 Hz. Note that the read voltage and the threshold of I&F neurons used in this attractor learning simulation were V$_{read}$ = 0.3 V and I$_{th}$ = 30 μA, respectively.

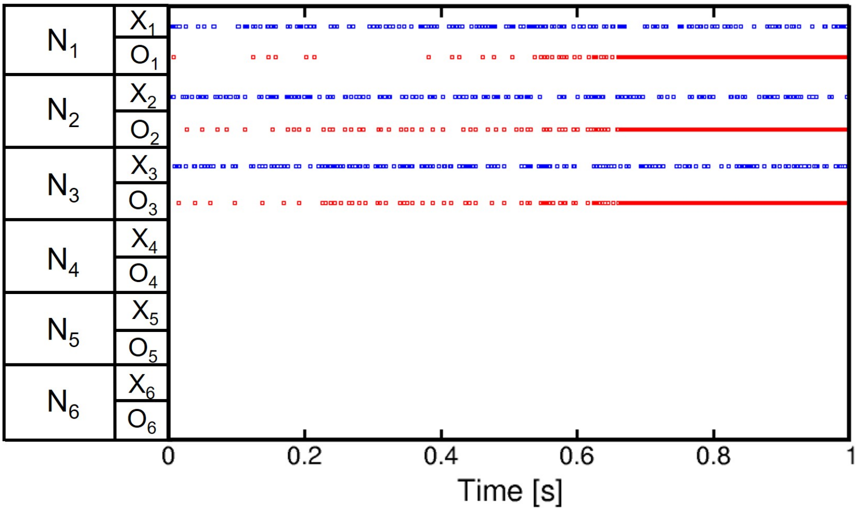
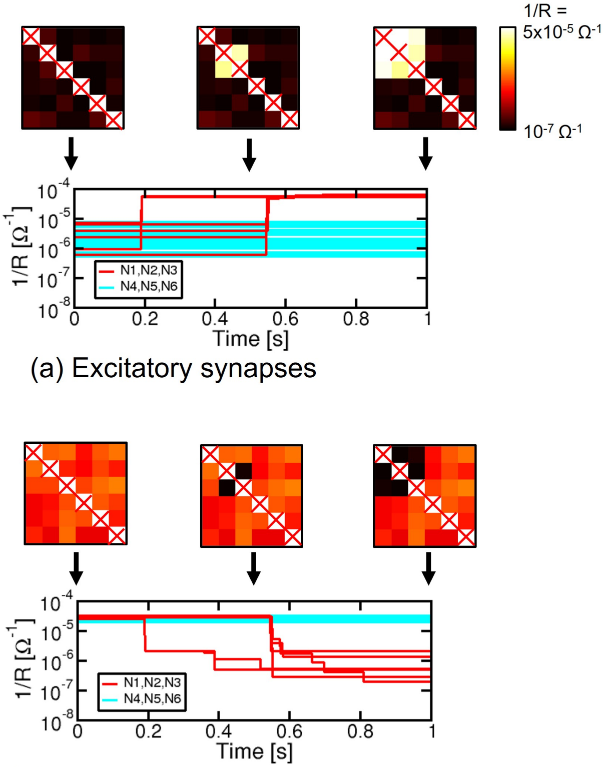


Figure 6.4: *Time evolution of calculated attractor learning in RRAM-based Hopfield neural network evidencing the spikes emitted by attractor neurons N_1, N_2 and N_3 (red dots) in response to an external stimulation (blue dots) at frequency f_x = 200 Hz. Copyright 2017 IEEE. Adapted, with permission, from [265].*

In the first epochs of training, the integration of external spikes provided to 3 neurons (blue dots) leads to the emission of output spikes (red dots) at low frequency because excitatory/inhibitory synaptic weights were initialized in a depressed/potentiated state. However, as the externally stimulated neurons fire at the same time, the potentiation of mutual excitatory synapses and the depression of corresponding inhibitory synapses are activated according to Hebbian plasticity rule inducing an increasing firing activity due to higher integrated synaptic currents, eventually leading to the formation of the attractor state [265]. This means that attractor learning in the Hopfield recurrent neural network results in the achievement of a stable high-frequency firing pattern for externally stimulated neurons [257].

Fig. 6.5 shows the evolution of both (a) excitatory and (b) inhibitory synaptic weights during attractor learning process described in Fig. 6.4, supported by color plots of weights at times t = 0, t = 0.5 s and t = 1 s. Starting from initial excitatory/inhibitory weights prepared in a depressed/potentiated state, the output spikes of externally-stimulated neurons activate set/reset transitions for the excitatory/inhibitory synapses shared by coactive neurons leading to the formation of the attractor state within t = 1 s [265]. Note that set transitions from HRS to LRS in 1T1R excitatory synapses and reset transitions from LRS to HRS in 1T1R inhibitory synapses



(a) Excitatory synapses

(b) Inhibitory synapses

Figure 6.5: *(a) Time evolution of the excitatory synaptic weights during attractor learning of sub-population N_1, N_2 and N_3. Starting from fully depressed excitatory weights, simultaneous spiking activity of the attractor neurons induces resistance transitions to LRS for mutual excitatory synapses until reaching attractor formation. (b) Time evolution of the inhibitory synaptic weights during attractor learning of sub-population N_1, N_2 and N_3. Starting from potentiated inhibitory weights, simultaneous spiking activity of attractor neurons induces resistance transitions to HRS for mutual inhibitory synapses until reaching attractor formation. Copyright 2017 IEEE. Adapted, with permission, from [265].*

were simulated using the stochastic Monte Carlo model for HfO_2 RRAM resistance distributions discussed in section 2.4.

6.3.2 Recall of a single attractor

After achieving single attractor learning, hybrid CMOS/RRAM Hopfield network was tested to demonstrate another key computational ability of recurrent neural networks, namely the recall of a stored attractor state by a

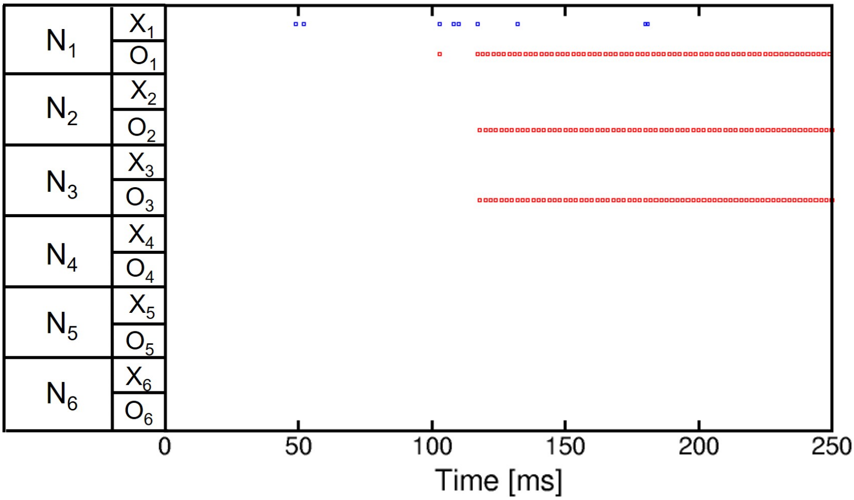


Figure 6.6: *Neuronal dynamics of the RRAM-based Hopfield network during recall process. The raster plot indicates that the external stimulation (blue dots) of a fraction of previously stored attractor, i.e. N_1, enables to reactivate the full attractor state (N_1, N_2, and N_3) and achieve a sustained spiking activity (red dots) even if the external stimulation is removed.*

partial external stimulation [257, 258].

To capture recall capability, the operation mode of network described for training in Fig. 6.3 is changed via the replacement of voltage pulses applied at the outputs of i-th neuron O_i and O'_i with the read voltages V_{read} and V'_{read}, respectively, thus preventing unwanted updates of the excitatory and inhibitory synaptic weights obtained by network training [267].

Fig. 6.6 shows the recall of the previously stored attractor in response to the application of a partial external stimulation at frequency f_x = 50 Hz only to one attractor neuron out of 3, namely N_1. The integration of external spikes X_1 of amplitude I_x = 10 μA (blue dots) leads N_1 to hit the threshold (I_{th} = 30 μA) and, consequently, emit output spikes (red dots) which, thanks to the attractor state, induce synaptic currents proportional to the LRS conductance of weights W_{12} and W_{13} at the input of N_2 and N_3, eventually leading to the attractor reactivation resulting in a sustained spiking activity despite the removal of external input at t = 200 ms.

To extensively investigate the recall of a single attractor state, Fig. 6.7(a) and (b) show the color plots of the recall probability as a function of f_x and I_{th} for I_x = I_{th}/5 in the case of one externally stimulated neuron and 2 externally stimulated neurons, respectively. Note that 100 1-s-long simulations

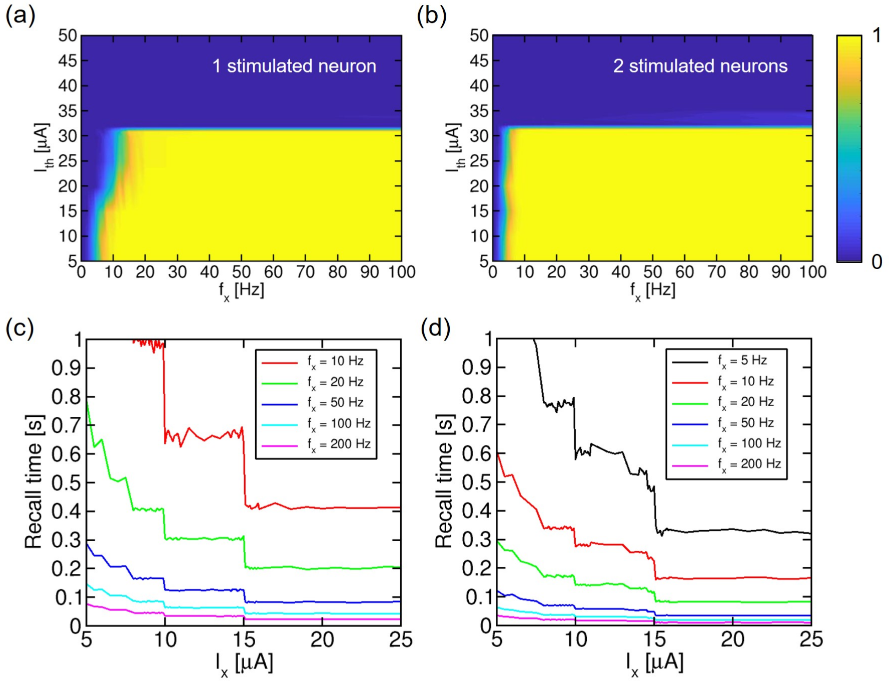


Figure 6.7: *Color plots of recall probability as a function of I_{th} and f_x for I_x set to $I_{th}/5$ in case of (a) 1 externally stimulated neuron and (b) 2 stimulated neurons. Average recall time as a function of I_x for variable f_x and I_{th} set to 30 µA in the case of (c) 1 externally stimulated neuron and (d) 2 externally stimulated neurons.*

were carried out for each pair of f_x and I_{th} values to obtain a sufficient statistics.

Simulation results show that attractor recall can be always achieved for $I_{th} \leq 30$ μA provided that $f_x \geq 10$ Hz in case of one stimulated neuron and $f_x \geq 5$ Hz in case of 2 stimulated neurons. On the one hand, neuron threshold has to be set within 30 μA because this is the maximum current that each neuron within attractor can integrate at each time to sustain the reverberation activity after removal of external input. On the other hand, the color plots evidence that the cooperation of two externally stimulated neurons can enable to achieve attractor recall even if the stimulation frequency is lower than one used in case of a single stimulated neuron. Also, note that if I_x was closer than I_{th}, the minimum frequency leading to attractor recall would decrease slightly since fewer external current spikes would be sufficient to reactivate spiking activity within the attractor.

In addition to probability to achieve attractor recall, average time needed to recall the attractor was also investigated. Fig. 6.7(c) and (d) show the calculated average recall time, namely the time required to reactivate a persistent spiking activity within the attractor, as a function of I$_x$ with increasing f$_x$ in the cases of one stimulated neuron and 2 stimulated neurons, respectively. The average recall time was studied assuming the maximum recall threshold, namely I$_{th}$ = 30 μA. Also, note that 1000 1-s-long simulations were carried out for each pair of I$_x$ and f$_x$ values to obtain a sufficient statistics.

Simulation results first evidence that calculated recall time decreases with increasing f$_x$ in both cases since a stronger external stimulation induces stimulated neuron/neurons to fire more rapidly, thus accelerating recall process. Also, observing recall time as a function of I$_x$ at each f$_x$, one can note that it decreases with transitions for specific I$_x$ indicating the need for one less external current spike to hit the neuron threshold and reactivate attractor. In particular, these jumps are abrupt in Fig. 6.7(c) because only one neuron channel is externally stimulated to achieve attractor recall, while they are more gradual in Fig. 6.7(d) since external spikes of 2 neurons can contribute to attractor recall. Finally, it should be noted that the external stimulation of 2 neurons accelerates recall process with respect to the case of 1 stimulated neuron, enabling to retrieve the attractor even at f$_x$ = 5 Hz, which, is not sufficient if only one neuron is externally stimulated.

6.4 Learning and recall of sequential attractors

6.4.1 Learning of two orthogonal attractors in sequence

In additon to the case of a single attractor, learning and recall capabilities of simulated Hopfield network were also tested for 2 non-overlapping or orthogonal sequential attractors [265].

To achieve sequential attractor learning, the 1-s-long external stimulation at f$_x$ = 200 Hz of neurons N$_1$, N$_2$ and N$_3$ was followed by a 1-s-long external stimulation of neurons N$_4$, N$_5$ and N$_6$ at the same frequency. Fig. 6.8 shows the color code representation of (a) excitatory and (b) inhibitory synaptic weights during sequential training. Starting from initially depressed/potentiated excitatory/inhibitory synaptic weights, the network is capable of learning the first attractor state in response to the external stimulation of N$_1$, N$_2$ and N$_3$ within t = 1 s by potentiation/depression of mutual excitatory/inhibitory synaptic weights, and the second attractor state in response to the external stimulation of N$_4$, N$_5$ and N$_6$ within t = 2 s by potentiation/depression of mutual excitatory/inhibitory synaptic weights.

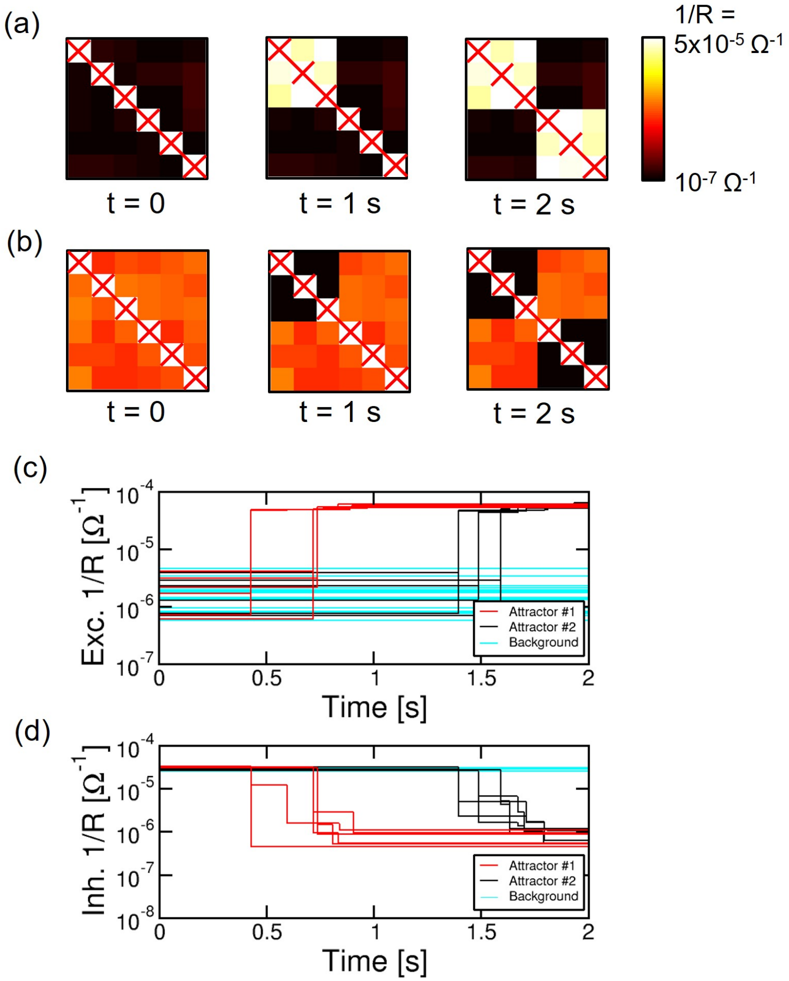


Figure 6.8: *Color code representation of calculated weights for (a) excitatory and (b) inhibitory synapses at times 0 s, 1 s, and 2 s during sequential learning process of two non-overlapping attractors. Time evolution of (c) excitatory and (d) inhibitory conductances evidencing formation of two orthogonal attractor states in sequence via resistance transitions of weights within each attractor to LRS and HRS, respectively. Copyright 2017 IEEE. Adapted, with permission, from [265].*

In addition, Fig. 6.8(c) and (d) show the time evolution of excitatory and inhibitory synaptic weights during sequential learning, respectively, evidencing the formation of each of two attractors via the resistance transitions to LRS/HRS of excitatory/inhibitory synapses within the first and second attractor, respectively. Note that in this simulation of sequential at-

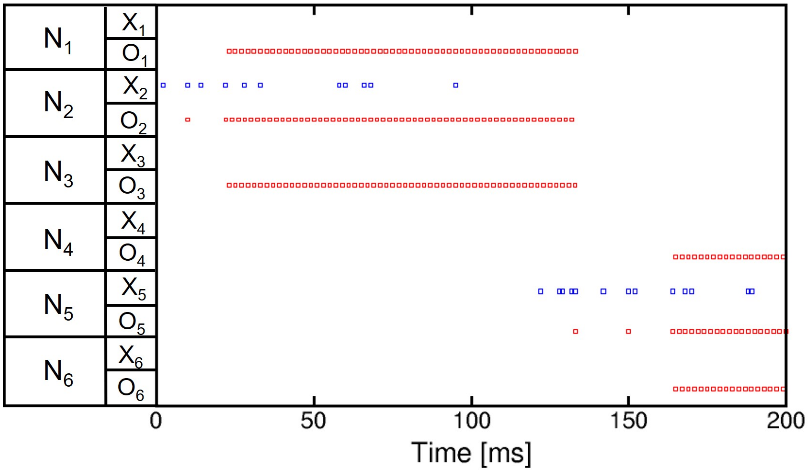


Figure 6.9: *Sequential recall of two orthogonal attractors, where the external stimulation provided to N_2 reactivates the first attractor (N_1, N_2, and N_3) in the first period within t = 100 ms, followed by stimulation of N_5 to recall second attractor (N_4, N_5, and N_6) in the second period within t = 200 ms.*

tractor learning I_x was set to 20 μA and the neuron threshold was set higher than 30 μA (I_{th} = 40 μA) to avoid that the strong spiking activity of first attractor after 1 s prevents the formation of the second attractor because of potentiated mutual inhibitory synapses.

6.4.2 Recall of two orthogonal attractors in sequence

Fig. 6.9 shows the spiking response of the network (red dots) induced by the partial stimulation of 2 previously stored attractors via external spikes (blue dots) sequentially provided to N_2 and N_5, respectively. The application of external spikes of amplitude I_x = 20 μA at frequency f_x = 100 Hz to N_2 leads it to cross the recall threshold I_{th} = 30 μA and emit output spikes activating synaptic currents across the potentiated excitatory synapses connecting N_2 to N_1 and N_3, eventually leading to the restoration of the first attractor by a high frequency spiking activity of sub-population N_1, N_2 and N_3 capable of sustaining even after t = 100 ms, namely when the external stimulation is switched from N_2 to N_5.

After t = 100 ms, the external current spikes provided to N_5 at frequency f_x cannot reactivate immediately the second attractor because they are reduced by inhibitory synaptic currents activated by output spikes of first attractor, thus preventing N_5 internal potential to cross rapidly the threshold.

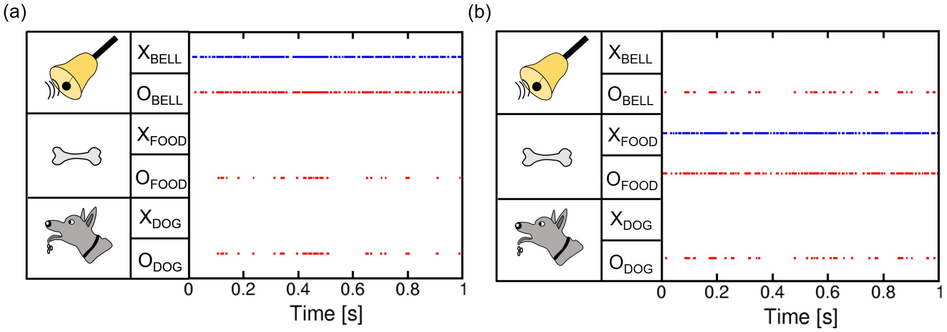


Figure 6.10: *Illustrative explanation of associative memory inspired to Pavlov's dog experiments. The external stimulation of a single neuron symbolized by the ring of a bell (a) or the vision of food (b) results in recalling of all neurons, namely bell, food and dog's salivation. Copyright 2017 IEEE. Reprinted, with permission, from [265].*

As the integration of these currents eventually leads N_5 to fire, the output spikes of N_5 activate inhibitory currents within the first attractor, thus leading to the switching from the first to the second attractor, as evidenced by activation of persistent spiking activity of N_4, N_5 and N_6 until t = 200 ms. Note that the switching from two attractors during sequential recall process can be achieved provided that inhibitory weights are initialized in a resistance state higher than LRS, for instance between 50 kΩ and 100 kΩ as in this simulation, such that the excitatory currents activated by second attractor can exceed the inhibitory currents activated by first attractor. This application thus highlights the need for inhibitory synapses to switch from one to another attractor in the simulated Hopfield network with 1T1R RRAM synapses.

6.5 Brain-inspired computing applications with RRAM-based Hopfield neural network

6.5.1 Associative memory

Learning and recall of attractor states discussed in the sections 6.3 and 6.4 pave the way to the replication of various human brain cognitive functions via RRAM-based Hopfield neural networks.

The recall of an attractor state by testing of the network with a partial or erroneous stimulus is first at the origin of a fundamental cognitive primitive in the mammalian brain known as associative learning, which received an intense theoretical and experimental investigation, as evidenced by the

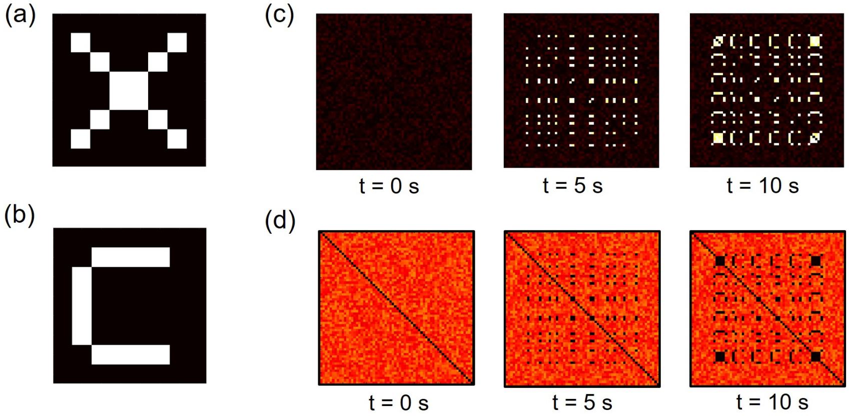


Figure 6.11: *(a) "X" and (b) "C" 8x8 input patterns used to train the simulated RRAM-based Hopfield neural network with 64 neurons. Color code representation of excitatory (c) and inhibitory (d) synaptic weights during sequential learning process at times t = 0 s, t = 5 s, and t = 10 s. Copyright 2018 IEEE. Adapted, with permission, from [267].*

Pavlov's dog experiments [272].

To illustrate the associative learning in the RRAM-based attractor neural network, Fig. 6.10 shows simulation results for recalling the attractor N_1, N_2, N_3, and its significance in terms of associative learning according to the Pavlov's dog experiments [265]. If the food presentation to the dog is always combined with the ringing of a bell, the "bell" and "food" concepts are associated, *i.e.*, an attractor linking bell and food is formed in the dog's brain. Consequently, whenever the dog hears the bell's ring alone, it resuscitates the concept of food and the stimulus to salivation (Fig. 6.10(a)), just as if the bone is directly presented to the dog (Fig. 6.10(b)).

This application is extremely significant since it provides a strong link between recurrent neural networks and the biophysics of the mammalian brain and consequently received a great interest resulting in many circuit level implementations by memristive devices [273–275].

6.5.2 Pattern completion

Pattern completion, namely the ability to reconstruct a previously learnt pattern by submission of an erroneous or incomplete stimulation, is one of the main brain-inspired computational tasks achieved in hardware [261, 263] and simulated recurrent neural networks [125, 266, 267].

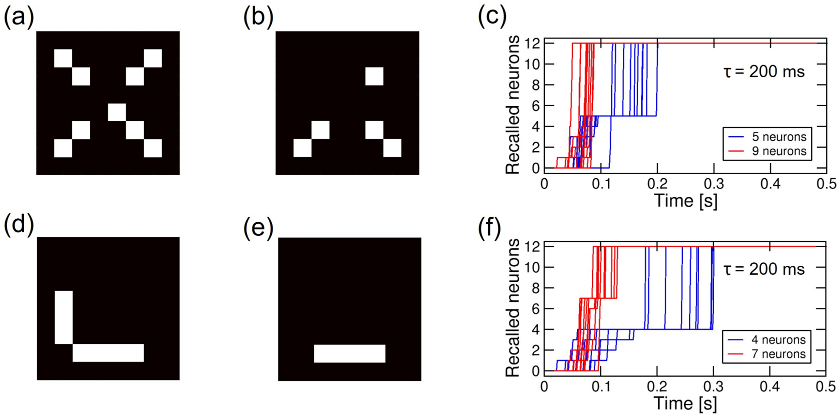


Figure 6.12: *Features of original pattern "X" with (a) 9 and (b) 5 active input channels inducing the reactivation of all the 12 attractor neurons within about 0.1 s and 0.2 s, respectively (c). Incomplete versions of the original pattern "C" with (d) 7 and (e) 4 active input channels leading to the reactivation of all the 12 attractor neurons within about 0.15 s and 0.3 s, respectively (f). Copyright 2018 IEEE. Adapted, with permission, from [267].*

To demonstrate pattern completion functionality, sequential learning of two orthogonal attractors associated to two 8x8 visual patterns representing a letter "X" (Fig. 6.11(a)) and a letter "C" (Fig. 6.11(b)), respectively, was first simulated by a RRAM-based Hopfield network with 64 leaky integrate-and-fire (LIF) neurons [267].

Fig. 6.11(c) and (d) show the color plots of calculated excitatory and inhibitory synaptic weights, respectively, during the sequential learning process of orthogonal attractors. Starting from initial excitatory synaptic weights in HRS and inhibitory synaptic weights in an intermediate resistive state (R = 50-100 kΩ), the presentation of pattern "X" to the network leads to the formation of the first attractor via potentiation/depression of excitatory/inhibitory synapses connecting "X" neurons within 5 s. After that, submitted pattern is changed from "X" to "C" leading to storage of the second attractor via potentiation/depression of excitatory/inhibitory synapses connecting "C" neurons within 10 s.

After achieving sequential learning of two attractors, pattern restoration was studied testing the Hopfield neural network with 64 neurons upon presentation of a fraction of the original patterns "X" and "C" [267].

Fig. 6.12(a) shows the input patterns that were submitted in the simulations, consisting of partial versions of the pattern "X" with only (a) 9 active

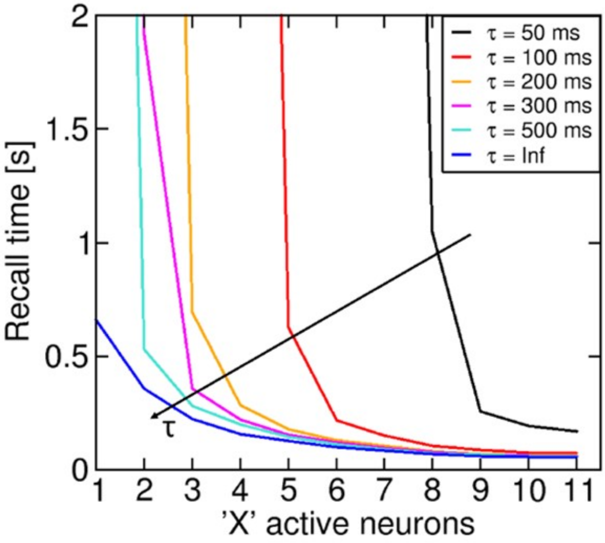


Figure 6.13: *Average recall time of attractor 'X' as a function of density of submitted partial pattern for variable τ values. Copyright 2018 IEEE. Adapted, with permission, from [267].*

channels or (b) 5 active channels and partial versions of the pattern "C" with only (c) 7 active channels or (d) 4 active channels.

Fig. 6.12(e) shows the simulation results of attractor recall with the partial patterns in (a) and (b), each case being simulated 10 times for statistical significance. The number of activated neurons increases with time during the submission of the partial pattern, eventually activating all the 12 neurons in the original pattern "X". Note that the average time required to retrieve the whole pattern "X" decreases as the number of externally stimulated neurons increases, as a result of the higher synaptic current feeding other unstimulated neurons within the selected attractor. Similarly to pattern "X", the stimulation of a fraction of attractor "C" leads to the restoration of all the 12 neurons in the attractor, as shown in Fig. 6.12(f). These results support error tolerant pattern recognition, where a pattern is recognized even in presence of a bare suggestion, or stimulation of only a fraction of the pattern.

In addition, Fig. 6.13 shows the calculated average recall time to retrieve "X" as a function of the number of attractor neurons stimulated by external spikes. In the simulations, various values of time constant τ, corresponding to various leakage during integration of incoming spikes, were assumed. The recall time was defined as the time the network takes to reactivate all

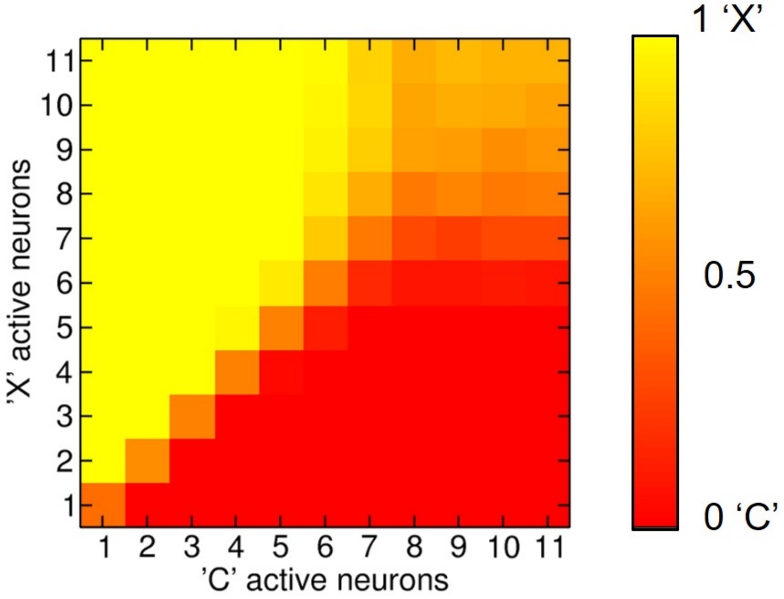


Figure 6.14: *Color map of probability P of recalling attractor "X", as a function of the number of externally activated neurons within pattern "X" and pattern "C". The probability of reactivating the attractor "C" can be obtained as 1-P. Copyright 2018 IEEE. Reprinted, with permission, from [267].*

neurons of a stored memory. As τ increases, the recall time for a given number of initially active neurons decreases because leakage impact on the firing activity of recalled attractor neurons gradually decreases. Similarly, the minimum number of active neurons to restore the full pattern increases for decreasing τ as the discharge of internal potential within each neuron becomes gradually faster by preventing attractor neurons to fire. Note that the same result would be also obtained for "C" since they features the same density of activated pixels [267].

6.5.3 Error correction

To explore the limits of the error tolerant recognition, and the possibility of confusion between competing patterns, Fig. 6.14 shows a color map of the calculated probability P of recognizing the pattern "X" after externally stimulating the Hopfield neural network with 64 neurons for 1 s [267]. The recognition probability is reported as a function of the number of externally stimulated neurons belonging to "X" or "C". The reported P is the average over 1000 simulations for each case. Note that all the simulations eventually led to recognition of either “X” or “C”, therefore the probability for recognizing pattern “C” is given by 1-P.

The results indicate that P increases as the number of stimulated X-neurons increases, and P decreases as the number of stimulated neurons within C increases. For similar number of X- and C- neurons being excited, the color plot shows random behavior of the simulated network with P of about 50%.

Finally, note that as the stimulated X- and C-neurons within the submitted test pattern are both above 7, P assumes intermediate values since such a high external excitation can activate either attractors with high probability, thus the recall process is mainly controlled by the stochastic Poisson input spike trains used to stimulate the attractor network.

These results corroborate the feasibility of error-tolerant brain-inspired Hopfield spiking neural network with RRAM-based 1T1R synapses capable of STDP.

6.6 Conclusions

In this chapter, a circuit implementation of Hopfield neural networks with 1T1R RRAM synapses capable of STDP has been described, enabling to achieve learning and recall of both a single attractor state and sequential attractor states. After demonstrating fundamental abilities for a recurrent neural network, attractive human brain primitives such as associative memory, pattern completion and error correction have been successfully achieved in simulation by the RRAM-based Hopfield neural network. These results thus pave the way for hardware implementation of brain-inspired associative learning in embedded low-power neuromorphic systems with resistive synapses.

Summary of results

A short summary of research achievements presented in this Ph.D. dissertation and the prospects of this work are reported in the following.

A stochastic Monte Carlo model for HfO_2 RRAM devices was first developed in order to capture stochastic learning processes in spiking neural networks with RRAM synapses thanks to its ability to accurately predict the experimental distributions of resistance obtained by characterization of set and reset transitions at variable applied voltage.

After that, two hybrid CMOS/RRAM circuits were designed and simulated enabling to achieve nanoscale resistive synapses capable of replicating two well-known bio-realistic learning rules as STDP and SRDP.

A RRAM-based synapse circuit with 1T1R structure enabled to implement STDP rule by application of a scheme using overlapping spikes. To support STDP functionality in hardware, this scheme was experimentally validated in 1T1R RRAM structures. Also, to support the dependence of STDP on initial state in 1T1R synapses, STDP characteristics calculated for variable initial resistance state by a previous analytical model of HfO_2 RRAM were compared with experimental counterparts evidencing a good agreement.

In addition to 1T1R synapse capable of STDP, another hybrid CMOS/RRAM synapse circuit comprising 4 transistors and one RRAM device was designed and simulated to implement SRDP learning rule.

Both synaptic potentiation for high frequency stimulation and synaptic depression for low frequency stimulation were successfully achieved by simulations, thus corroborating the experimental measurements carried out by 2T1R integrated structures.

After discussing the use of RRAM device at level of single synapse, simulations and experiments were also carried out at higher level of network for implementing neuromorphic computing tasks such as unsupervised learning and recognition of visual patterns.

In this frame, a two-layer feedforward spiking neural network with 1T1R synapses implementing STDP was designed and simulated achieving unsupervised learning and recognition of a visual pattern according to a stochastic approach. After demonstrating learning of a single visual pattern, on-line learning of sequential patterns and on-line learning of multiple patterns were also demonstrated in simulation. Also, to support the feasibility of these fundamental cognitive tasks in hardware, all calculated results were validated by experiments via a hardware neural network with 1T1R RRAM synapses built on a PCB. Furthermore, a detailed study based on experiments and simulations investigating pattern learning performance as a function of density of submitted random noise was carried out leading to find the noise density optimizing learning performance of neuromorphic network used in the experiments.

In addition to unsupervised pattern learning by STDP, unsupervised learning of a single pattern and on-line learning of sequential patterns were also demonstrated by design and simulation of a two-layer feedforward spiking neural network with 4T1R RRAM-based synapses implementing SRDP. To support simulation results, unsupervised pattern learning via SRDP rule was validated by experiments with 2T1R integrated structures, separately achieving background depression and pattern potentiation. Also, the impact of noise frequencies on learning efficiency of the neuromorphic network was investigated in simulation achieving the best trade-off.

Finally, a circuit implementation of a Hopfield recurrent neural network with 1T1R RRAM synapses capable of STDP was developed and simulated. RRAM-based Hopfield network first enabled to demonstrate in simulation learning and recall of both a single attractor state and two sequential orthogonal attractor states. Based on these results, attractive brain-inspired cognitive abilities such as associative memory, pattern restoration and error correction were achieved by simulation of RRAM-based Hopfield networks.

These results can open the way for the exploration and implementation at simulation and experimental level of other interesting neuromorphic computing applications in the coming years.

First, feedforward spiking neural networks with a higher number of layers fully connected by resistive synapses capable of bio-realistic plasticity can be developed and simulated to achieve the demonstration of cognitive tasks more complex than learning of a single visual pattern or sequential visual patterns such as the classification of images within very large datasets.

Also, taking inspiration from the great potential of Hopfield networks, Hopfield spiking neuromorphic network based on resistive synapses can be further investigated to demonstrate in simulation and hardware new cognitive functions such as decision making ability and unsupervised learning of temporal sequences, and the solution of hard computational problems such as constraint-satisfaction problems, *e.g.* Sudoku puzzle, or systems of linear equations.

Finally, the approach based on resistive switching synaptic devices can be very useful to also achieve the implementation of other fundamental types of neuromorphic networks that are gaining increasing interest in neuromorphic computing landscape such as Restricted Boltzmann Machines and Reservoir Computing networks.

List of publications

Book chapters

D. Ielmini and V. Milo, “Brain-inspired memristive neural networks for unsupervised learning”, in Handbook of Memristor Networks, G. Sirakoulis and L. Chua eds. (Springer, 2018).

International Refereed Journals

V. Milo, G. Pedretti, R. Carboni, A. Calderoni, N. Ramaswamy, and D. Ielmini, “A 4-transistors/1-resistor hybrid synapse based on resistive switching memory (RRAM) capable of spike-rate-dependent plasticity (SRDP)”, *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, 26(12), pp. 2806–2815, December 2018.

W. Wang, G. Pedretti, V. Milo, R. Carboni, A. Calderoni, N. Ramaswamy, A. S. Spinelli, and D. Ielmini, “Learning of spatio-temporal patterns in a spiking neural network with resistive switching synapses”, *Science Advances* 4(9):eaat4752, September 2018.

G. Pedretti, V. Milo, S. Ambrogio, R. Carboni, S. Bianchi, A. Calderoni, N. Ramaswamy, A. S. Spinelli, and D. Ielmini, “Stochastic learning in neuromorphic hardware via spike timing dependent plasticity with RRAM synapses”, *IEEE Journal of Emerging and Selected Topics in Circuits and Systems (JETCAS)* 8(1), pp. 77–85, March 2018.

Y. Ren, V. Milo, Z. Wang, H. Xu, X. Zhao, D. Ielmini, and Y. Liu “Analytical modeling of organic-inorganic $CH_3NH_3PbI_3$ perovskite resistive switching and its application for neuromorphic recognition”, *Adv. Theory Simul.*, 1700035, pp. 1–8, April 2018.

D. Ielmini and V. Milo, “Physics-based modeling approaches of resistive switching devices for memory and in-memory computing applications”, *Journal of Computational Electronics (JCEL)*, 16(4), pp. 1121–1143, November 2017.

G. Pedretti, V. Milo, S. Ambrogio, R. Carboni, S. Bianchi, A. Calderoni, N. Ramaswamy, A. S. Spinelli, and D. Ielmini “Memristive neural network for on-line learning and tracking with brain-inspired spike timing dependent plasticity”, *Scientific Reports*, 7:5288, July 2017.

S. Ambrogio, V. Milo, Z.-Q. Wang, S. Balatti, and D. Ielmini “Analytical modeling of current overshoot in oxide-based resistive switching memory (RRAM), *IEEE Electron Device Letters* 37(10), pp. 1268–1271, October 2016.

S. Balatti, S. Ambrogio, R. Carboni, V. Milo, Z.-Q. Wang, A. Calderoni, N. Ramaswamy, and D. Ielmini “Physical unbiased generation of random numbers with coupled resistive switching devices”, *IEEE Trans. Electron Devices* 63(5), pp. 2029–2035, May 2016.

S. Ambrogio, S. Balatti, V. Milo, R. Carboni, Z.-Q. Wang, A. Calderoni, N. Ramaswamy, and D. Ielmini, “Neuromorphic learning and recognition with one-transistor-one-resistor synapses and bistable metal oxide RRAM”, *IEEE Trans. Electron Devices* 63(4), pp. 1508–1515, March 2016.

S. Ambrogio, N. Ciocchini, M. Laudato, V. Milo, A. Pirovano, P. Fantini, and D. Ielmini, “Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses”, *Frontiers in Neuroscience* 10:56, March 2016.

International Conference Proceedings

W. Wang, G. Pedretti, V. Milo, R. Carboni, A. Calderoni, N. Ramaswamy, A. S. Spinelli, and D. Ielmini, “Computing of temporal information in spiking neural networks with ReRAM synapses”, *Faraday Discussions*, Aachen 15-17 October 2018.

V. Milo, E. Chicca, and D. Ielmini, “Brain-inspired recurrent neural network with plastic RRAM synapses”, *Int. Symp. on Circuits and Systems (ISCAS)*, Florence 27-30 May 2018.

V. Milo, G. Pedretti, M. Laudato, A. Bricalli, E. Ambrosi, S. Bianchi, E. Chicca, and D. Ielmini, “Resistive switching synapses for unsupervised learning in feed-forward and recurrent neural networks”, *Int. Symp. on Circuits and Systems (ISCAS)*, Florence 27-30 May 2018.

V. Milo, D. Ielmini, and E. Chicca “Attractor networks and associative memories with STDP learning in RRAM synapses”, *IEEE IEDM Tech. Dig.*, pp. 263–266, San Francisco (USA), 2-6 December 2017.

G. Pedretti, S. Bianchi, V. Milo, A. Calderoni, N. Ramaswamy, and D. Ielmini, “Modeling-based design of brain-inspired spiking neural networks with RRAM learning synapses”, *IEEE IEDM Tech. Dig.*, pp. 653–656, San Francisco (USA), 2-6 December 2017.

V. Milo, G. Pedretti, R. Carboni, A. Calderoni, N. Ramaswamy, S. Ambrogio, and D. Ielmini, “Demonstration of hybrid CMOS/RRAM neural networks with spike time/rate-dependent plasticity”, *IEEE IEDM Tech. Dig.*, pp. 440–443, San Francisco (USA), 3-7 December 2016.

S. Ambrogio, S. Balatti, V. Milo, R. Carboni, Z.-Q. Wang, A. Calderoni, N. Ramaswamy, and D. Ielmini, “Novel RRAM-enabled 1T1R synapse capable of low-power STDP via burst-mode communication and real-time unsupervised machine learning”, *IEEE Symposium on VLSI Technology*, pp. 196–197 (2016).

D. Ielmini, S. Ambrogio, V. Milo, S. Balatti, Z.-Q. Wang,“Neuromorphic computing with hybrid memristive/CMOS synapses for real-time learning”, *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1386–1389 (2016).

Awards and recognitions

IEEE Golden List of Reviewers for 2017, IEEE Trans. on Electron Devices, vol. 64, no. 12, December 2017

Acknowledgements

A conclusione di questo percorso di studi, desidero innanzitutto ringraziare il Prof. Daniele Ielmini per avermi dato la grande opportunità di svolgere il Dottorato di Ricerca sotto la sua supervisione e per la grande fiducia che egli ha riposto in me in questi 3 anni consentendomi di collaborare all'attività didattica e di partecipare a conferenze internazionali di grande prestigio.

Voglio inoltre ringraziare il Dr. Stefano Ambrogio per gli insegnamenti e il grande supporto che mi ha fornito durante il primo anno di dottorato, e l'intero gruppo di ricerca composto da Giacomo, Elia, Roberto, Alessandro, Mario, Nicola, Stefano, Irene, Octavian, Erika, Zhong e Wei per la grande disponibilità e amicizia che hanno dimostrato quotidianamente nei miei confronti.

Infine, un grazie speciale va ai miei genitori, a mio fratello e alla mia fidanzata per avermi sempre sostenuto con grande affetto in questa importante fase della mia vita. Senza di loro, il raggiungimento di questo obiettivo non sarebbe stato possibile.

Valerio

Bibliography

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, pp. 114–117, 1965.

[2] M. M. Waldrop, "The chips are down for Moore's law," *Nature*, vol. 530, no. 7589, pp. 144–147, 2016.

[3] M. Horowitz, "Computing's energy problem (and what we can do about it)," *IEEE Int. Solid-State Circuits Conference (ISSCC)*, pp. 10–14, 2014.

[4] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

[5] W. A. Wulf and S. A. McKee, "Hitting the memory wall: implications of the obvious," *ACM SIGARCH Computer Architecture News*, vol. 23, no. 1, pp. 20–24, 1995.

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

[7] R. S. Williams, "What's next?" *Computing in Science and Engineering*, vol. 19, no. 2, pp. 7–13, 2017.

[8] S. Salahuddin, K. Ni, and S. Datta, "The era of hyper-scaling in electronics," *Nature Electronics*, vol. 1, pp. 442–450, 2018.

[9] J. M. Shalf and R. Leland, "Computing beyond Moore's law," *Computer*, vol. 48, no. 12, pp. 14–23, 2015.

[10] M. Di Ventra and Y. V. Pershin, "The parallel approach," *Nature Physics*, vol. 9, pp. 200–202, 2013.

[11] K. J. Kuhn, "Considerations for ultimate CMOS scaling," *IEEE Trans. Electron Devices*, vol. 59, no. 7, pp. 1813–1828, 2012.

[12] A. M. Ionescu and H. Riel, "Tunnel field-effect transistors as energy-efficient electronic switches," *Nature*, vol. 479, no. 7373, pp. 329–337, 2011.

[13] S. Salahuddin and S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nano Lett.*, vol. 8, no. 2, pp. 405–410, 2008.

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

[15] S. Yu and P.-Y. Chen, "Emerging memory technologies: recent trends and prospects," *IEEE Solid-State Circuits Magazine*, vol. 8, no. 2, pp. 43–56, 2016.

[16] L. O. Chua, "Memristor – The missing circuit element," *IEEE Trans. Circ. Theory*, vol. 18, no. 5, pp. 507–519, 1971.

[17] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *Nature*, vol. 453, no. 7191, pp. 80–83, 2008.

[18] A. Kawahara, R. Azuma, Y. Ikeda, K. Kawai, Y. Katoh, K. Tanabe, T. Nakamura, Y. Sumimoto, N. Yamada, N. Nakai, S. Sakamoto, Y. Hayakawa, K. Tsuji, S. Yoneda, A. Himeno, K. Origasa, K. Shimakawa, T. Takagi, T. Mikawa, and K. Aono, "An 8 Mb multi-layered cross-point ReRAM macro with 443MB/s write throughput," *Proc. IEEE Int. Solid-State Circuits Conf.*, pp. 178–185, 2012.

[19] T. Y. Liu, T. H. Yan, R. Scheuerlein, Y. Chen, J. K. Lee, G. Balakrishnan, G. Yee, H. Zhang, A. Yap, J. Ouyang, T. Sasaki, S. Addepalli, A. Al-Shamma, C. Y. Chen, M. Gupta, G. Hilton, S. Joshi, A. Kathuria, V. Lai, D. Masiwal, and M. Matsumoto, "A 130.7 mm^2 2-layer 32Gb ReRAM memory device in 24nm technology," *Proc. IEEE Int. Solid-State Circuits Conf.*, pp. 210–211, 2013.

[20] "Intel and Micron produce breakthrough memory technology. (2015). [Online]. Available: http://newsroom.intel.com/community/intel_newsroom/blog 2015/07/28/intel-and-micron-produce-breakthrough-memory-technology."

[21] "Intel Optane Technology [Online]. http://www.intel.com/content/www/us/en/architecture-and-technology/inte-optane-technology.html."

[22] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature Electronics*, vol. 1, pp. 333–343, 2018.

[23] M. A. Zidan, J. P. Strachan, and W. D. Lu, "The future of electronics based on memristive systems," *Nature Electronics*, vol. 1, pp. 22–29, 2018.

[24] D. J. Wouters, R. Waser, and M. Wuttig, "Phase-change and Redox-based resistive switching memories," *Proceedings of the IEEE*, vol. 103, no. 8, pp. 1951–1970, 2015.

[25] D. Ielmini, "Resistive switching memories based on metal oxides: mechanisms, reliability and scaling," *Semicond. Sci. Technol.*, vol. 31, no. 6, p. 063002, 2016.

[26] S. W. Fong, C. M. Neumann, and H.-S. P. Wong, "Phase Change Memory – Towards a storage class memory," *IEEE Trans. Electron Devices*, vol. 64, no. 11, pp. 4374–4385, 2017.

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

[28] G. W. Burr, M. J. Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson, B. Kurdi, C. Lam, L. A. Lastras, A. Padilla, B. Rajendran, S. Raoux, and R. S. Shenoy, "Phase Change Memory technology," *Journal of Vacuum Science and Technology B*, vol. 28, no. 2, pp. 223–262, 2010.

[29] S. Raoux, W. Welnic, and D. Ielmini, "Phase change materials and their application to non-volatile memories," *Chem. Rev.*, vol. 110, no. 1, pp. 240–267, 2010.

[30] G. W. Burr, M. J. Brightsky, A. Sebastian, H.-Y. Cheng, J.-Y. Wu, S. Kim, N. E. Sosa, N. Papandreou, H.-L. Lung, H. Pozidis, E. Eleftheriou, and C. H. Lam, "Recent progress in phase-change memory technology," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 6, no. 2, pp. 146–162, 2016.

[31] A. Athmanathan, M. Stanisavljevic, N. Papandreou, H. Pozidis, and E. Eleftheriou, "Multilevel-cell phase-change memory: a viable technology," *IEEE Journal of Emerging Topics on Circuits and Systems (JETCAS)*, vol. 6, no. 1, pp. 87–100, 2016.

[32] A. Pirovano, A. L. Lacaita, A. Benvenuti, F. Pellizzer, S. Hudgens, and R. Bez, "Scaling analysis of phase-change memory technology," *IEEE IEDM Tech. Dig.*, pp. 29.6.1–29.6.4, 2004.

[33] 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," *IEDM Tech. Dig.*, pp. 939–942, 2007.

[34] M. Boniardi and D. Ielmini, "Physical origin of the resistance drift exponent in amorphous phase change materials," *Appl. Phys. Lett.*, vol. 98, no. 24, p. 243506, 2011.

[35] N. Papandreou, A. Sebastian, A. Pantazi, M. Breitwisch, C. Lam, H. Pozidis, and E. Elefteriou, "Drift-resilient cell-state metric for multilevel phase-change memory," *IEEE IEDM Tech. Dig.*, pp. 3.5.1–3.5.4, 2011.

[36] M. Stanisavljevic, A. Athmanathan, N. Papandreou, H. Pozidis, and E. Eleftheriou, "Phase-change memory: Feasibility of reliable multilevelcell storage and retention at elevated temperatures," *Int. Rel. Phys. Symp.*, pp. 5B.6.1–5B.6.6, 2015.

[37] A. Sawa, "Resistive switching in transition metal oxides," *Materials Today*, vol. 11, no. 6, pp. 28–36, 2008.

[38] C. Chappert, A. Fert, and F. N. V. Dau, "The emergence of spin electronics in data storage," *Nature Mater.*, vol. 6, pp. 813–823, 2007.

[39] J. J. Yang, D. B. Strukov, and D. R. Stewart, "Memristive devices for computing," *Nature Nanotechnol.*, vol. 8, no. 1, pp. 13–24, 2013.

[40] M.-J. Lee, C. B. Lee, D. Lee, S. R. Lee, M. Chang, J. H. Hur, Y.-B. Kim, C.-J. Kim, D.-H. Seo, U.-I. Chung, I.-K. Yoo, and K. Kim, "A fast, high-endurance and scalable non-volatile memory device made from asymmetric $\mathrm{Ta_2O}_{5-x}$/Tao_{2-x} bilayer structures," *Nature Mater.*, vol. 10, pp. 625–630, 2011.

[41] Q. Wang, Y. Itoh, T. Tsuruoka, M. Aono, and T. Hasegawa, "Ultra-low voltage and ultra-low power consumption nonvolatile operation of a three-terminal atomic switch," *Advanced Materials*, vol. 27, pp. 6029–6033, 2015.

[42] R. Fackenthal, M. Kitagawa, W. Otsuka, K. Prall, D. Mills, K. Tsutsui, J. Javanifard, K. Tedrow, T. Tsushima, Y. Shibahara, and G. Hush, "A 16 GB ReRAM with 200 MB/s write and 1 GB/s read in 27 nm technology," *ISSCC Tech. Dig.*, pp. 338–339, 2014.

[43] M. Ueki, K. Takeuchi, T. Yamamoto, A. Tanabe, N. Ikarashi, M. Saitoh, T. Nagumo, H. Sunamura, M. Narihiro, K. Uejima, K. Masuzaki, N. Furutake, S. Saito, Y. Yabe, A. Mitsuiki, K. Takeda, T. Hase, and Y. Hayashi, "Low-power embedded ReRAM technology for IoT applications," *Symposium on VLSI Technology*, pp. T108–T109, 2015.

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

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

[46] I. G. Baek, M. S. Lee, S. Seo, M. J. Lee, D. H. Se, D.-S. Suh, J. C. Park, S. O. Park, H. S. Kim, I. K. Yoo, U.-I. Chung, and I. T. Moon, "Highly scalable nonvolatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses," *IEEE IEDM Tech. Dig.*, pp. 587–590, 2004.

[47] S. Seo, M. J. Lee, D. H. Seo, E. J. Jeoung, D.-S. Suh, Y. S. Joung, I. K. Yoo, I. R. Hwang, S. H. Kim, I. S. Byun, J.-S. Kim, J. S. Choi, and B. H. Park, "Reproducible resistance switching in polycrystalline NiO films," *Appl. Phys. Lett.*, vol. 85, p. 5655, 2004.

[48] D. Ielmini, R. Bruchhaus, and R. Waser, "Thermochemical resistive switching: materials, mechanisms and scaling projections," *Phase Transition*, vol. 84, pp. 570–602, 2011.

[49] U. Russo, D. Ielmini, C. Cagli, and A. L. Lacaita, "Filament conduction and reset mechanism in NiO-based resistive-switching memory (RRAM) devices," *IEEE Trans. Electron Devices*, vol. 56, no. 2, pp. 186–192, 2009.

[50] ——, "Self-accelerated thermal dissolution model for reset programming in NiO-based resistive switching memory (RRAM) devices," *IEEE Trans. Electron Devices*, vol. 56, no. 2, pp. 193–200, 2009.

[51] D. Ielmini, "Modeling the universal set/reset characteristics of bipolar RRAM by field- and temperature-driven filament growth," *IEEE Trans. Electron Devices*, vol. 58, no. 12, pp. 4309–4317, 2011.

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

[53] F. Nardi, S. Balatti, S. Larentis, and D. Ielmini, "Complementary switching in metal oxides: toward diode-less cross-bar RRAMs," *IEEE IEDM Tech. Dig.*, pp. 709–712, 2011.

[54] I. Valov, R. Waser, J. R. Jameson, and M. N. Kozicki, "Electrochemical metallization memories-fundamentals, applications, prospects," *Nanotechnology*, vol. 22, p. 254003, 2011.

[55] M. N. Kozicki, M. Park, and M. Mitkova, "Nanoscale memory elements based on solid-state electrolytes," *IEEE Trans. Nanotechnol.*, vol. 4, no. 3, pp. 331–338, 2005.

[56] M. N. Kozicki, C. Gopalan, M. Balakrishnan, and M. Mitkova, "A low-power nonvolatile switching element based on copper-tungsten oxide solid electrolyte," *IEEE Trans. Nanotechnol.*, vol. 5, no. 5, pp. 535–544, 2006.

[57] I. Valov, "Redox-based resistive switching memories (ReRAMs): Electrochemical systems at the atomic scale," *ChemElectroChem*, vol. 1, pp. 26–36, 2014.

[58] C. Schindler, S. C. P. Thermadam, R. Waser, and M. N. Kozicki, "Bipolar and unipolar resistive switching in Cu-doped SiO_2," *IEEE Trans. Electron Devices*, vol. 54, no. 10, pp. 2762–2768, 2007.

[59] C. Schindler, M. Weides, M. N. Kozicki, and R. Waser, "Low current resistive switching in Cu-SiO_2 cells," *Appl. Phys. Lett.*, vol. 92, no. 12, p. 122910, 2008.

[60] U. Russo, D. Kalamanathan, D. Ielmini, A. L. Lacaita, and M. Kozicki, "Study of multilevel programming in programmable metallization cell (PMC) memory," *IEEE Trans. Electron Devices*, vol. 56, no. 5, pp. 1040–1047, 2009.

[61] S.-L. Li, D. S. Shang, J. Li, J. L. Gang, and D. N. Zheng, "Resistive switching properties in oxygen-deficient $Pr_{0.7}Ca_{0.3}MnO_3$ junctions with active al top electrodes," *J. Appl. Phys.*, vol. 105, p. 033710, 2009.

[62] Y.-F. Wang, Y. C. Lin, I.-T. Wang, T.-P. Lin, and T.-H. Hou, "Characterization and modeling of nonfilamentary Ta/TaO_x/TiO_2/Ti analog synaptic device," *Sci. Rep.*, vol. 5, p. 10150, 2015.

[63] C. W. Hsu, Y.-F. Wang, C.-C. Wan, I.-T. Wang, C.-T. Chou, W.-L. Lai, Y.-J. Lee, and T.-H. Hou, "Homogeneous barrier modulation of TaO_x/TiO_2 bilayers for ultra-high endurance three-dimensional storage-class memory," *Nanotechnology*, vol. 25, p. 165202, 2014.

[64] M. Hasan, R. Dong, H. J. Choi, D. S. Lee, D.-J. Seong, M. B. Pyun, and H. Hwang, "Uniform resistive switching with a thin reactive metal interface layer in Metal-$La_{0.7}Ca_{0.3}MnO_3$-Metal heterostructures," *Appl. Phys. Lett.*, vol. 92, p. 202102, 2008.

[65] H. Sim, H. Choi, D. Lee, M. Chang, D. Choi, Y. Son, E.-H. Lee, W. Kim, Y. Park, I.-K. Yoo, and H. Hwang, "Excellent resistance switching characteristics of Pt/$SrTiO_3$ Schottky junction for multi-bit nonvolatile memory application," *IEEE IEDM Tech. Dig.*, pp. 758–761, 2005.

[66] S. Ambrogio, S. Balatti, D. C. Gilmer, and D. Ielmini, "Analytical modeling of oxide-based bipolar resistive memories and complementary resistive switches," *IEEE Trans. Electron Devices*, vol. 61, no. 7, pp. 2378–2386, 2014.

[67] J. C. Slonczewski, "Current-driven excitation of magnetic multilayers," *J. Magn. Magn. Mater.*, vol. 159, no. 1-2, pp. L1–L7, 1996.

[68] L. Berger, "Emission of spin waves by a megnetic multilayer traversed by a current," *Phys. Rev. B*, vol. 54, pp. 9353–9358, 1996.

[69] M. Tsoi, A. G. M. Jansen, J. Bass, W.-C. Chiang, M. Seck, V. Tsoi, and P. Wyder, "Excitation of a magnetic multilayer by an electric current," *Phys. Rev. Lett.*, vol. 80, pp. 4281–4284, 1998.

[70] E. B. Myers, D. C. Ralph, J. A. Katine, R. N. Louie, and R. A. Buhrman, "Current-induced switching of domains in magnetic multilayer devices," *Science*, vol. 285, no. 5429, pp. 867–870, 1999.

[71] J. A. Katine, F. J. Albert, R. A. Buhrman, E. B. Myers, and D. Ralph, "Current-driven magnetization reversal and spin-wave excitations in Co/Cu/Co pillars," *Phys Rev Lett.*, vol. 84, no. 14, pp. 3149–3152, 2000.

[72] R. Carboni, S. Ambrogio, W. Chen, M. Siddik, J. Harms, A. Lyle, W. Kula, G. Sandhu, and D. Ielmini, "Understanding cycling endurance in perpendicular spin-transfer torque (p-STT) magnetic memory," *IEEE IEDM Tech. Dig.*, pp. 572–575, 2016.

[73] J. Kim, A. Paul, P. A. Crowell, S. J. Koester, S. S. Sapatnekar, J.-P. Wang, and C. H. Kim, "Spin-based computing: device concepts, current status. and a case study on a high-performance microprocessor," *Proc. IEEE*, vol. 103, no. 1, pp. 106–130, 2015.

[74] S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki, and K. Ando, "Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions," *Nature Mater.*, vol. 3, pp. 868–871, 2004.

[75] M. Julliere, "Tunneling between ferromagnetic-films," *Phys. Lett. A*, vol. 54, no. 3, pp. 225–226, 1975.

[76] A. D. Kent and D. C. Worledge, "A new spin on magnetic memories," *Nature Nanotechnology*, vol. 10, pp. 187–191, 2015.

[77] N. Locatelli, V. Cros, and J. Grollier, "Spin-torque building blocks," *Nature Mater.*, vol. 13, pp. 11–20, 2014.

[78] J. J. Kan, C. Park, C. Ching, J. Ahn, Y. Xie, M. Pakala, and S. H. Kang, "A study on practically unlimited endurance of STT-MRAM," *Trans. on Electron Devices*, vol. 64, no. 9, pp. 3639–3646, 2017.

[79] R. Carboni, S. Ambrogio, W. Chen, M. Siddik, J. Harms, A. Lyle, W. Kula, G. Sandhu, and D. Ielmini, "Modeling of breakdown-limited endurance in spin-transfer torque magnetic memory under pulsed cycling regime," *Trans. on Electron Devices*, vol. 65, no. 6, pp. 2470–2478, 2018.

[80] D. Saida, S. Kashiwada, M. Yakabe, T. Daibou, N. Hase, M. Fukumoto, S. Miwa, Y. Suzuki, H. Noguchi, S. Fujita, and J. Ito, "Sub-3 ns pulse with sub-100 μA switching of 1x-2x nm perpendicular MTJ for high-performance embedded STT-MRAM towards sub-20 nm CMOS," *IEEE Symp. on VLSI Technology*, pp. 1–2, 2016.

[81] J. J. Novak, R. P. Robertazzi, J. Z. Sun, G. Hu, J.-H. Park, J. Lee, A. J. Annunziata, G. P. Lauer, R. Kothandaraman, E. J. O'Sullivan, P. L. Trouilloud, Y. Kim, and D. C. Worledge, "Dependence of voltage and size on write error rates in spin-transfer torque magnetic random-access memory," *IEEE Magnetics Letters*, vol. 7, pp. 1–4, 2016.

[82] W. Maass, "To spike or not to spike: that is the question," *Proc. IEEE*, vol. 103, no. 12, pp. 2219–2224, 2015.

[83] C. Mead, "Analog VLSI and neural systems," *Addison-Wesley*, 1989.

[84] ——, "Neuromorphic electronic systems," *Proc. IEEE*, vol. 78, no. 10, pp. 1629–1636, 1990.

[85] M. Mahowald and R. Douglas, "A silicon neuron," *Nature*, vol. 354, pp. 515–518, 1991.

[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, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, and K. Boahen, "Neuromorphic silicon neuron circuits," *Front. Neurosci.*, vol. 5, p. 73, 2011.

[87] C. Bartolozzi and G. Indiveri, "Synaptic dynamics in analog VLSI," *Neural Comput.*, vol. 19, pp. 2581–2603, 2007.

[88] T. Yu and G. Cauwenberghs, "Log-domain time-multiplexed realization of dynamical conductance-based synapses," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2558–2561, 2010.

[89] M. Mahowald, "An analog VLSI system for stereoscopic vision," *Boston, MA: Kluwer*, 1994.

[90] E. Chicca, D. Badoni, V. Dante, M. D'Andreagiovanni, G. Salina, L. Carota, S. Fusi, and P. Del Giudice, "A VLSI recurrent network of integrate-and-fire neurons connected by plastic synapses with long-term memory," *IEEE Trans. Neural Netw.*, vol. 14, no. 5, pp. 1297–1307, 2003.

[91] J. V. Arthur and K. Boahen, "Recurrently connected silicon neurons with active dendrites for one-shot learning," *IEEE Int. Joint Conf. Neural Netw.*, vol. 3, pp. 1699–1704, 2004.

[92] G. Indiveri, E. Chicca, and R. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE Trans. on Neural Netw.*, vol. 17, no. 1, pp. 211–221, 2006.

[93] E. Chicca, A. M. Whatley, V. Dante, P. Lichtsteiner, T. Delbrück, P. Del Giudice, R. J. Douglas, and G. Indiveri, "A multi-chip pulse-based neuromorphic infrastructure and its application to a model of orientation selectivity," *IEEE Trans. Circuits Syst. I*, vol. 5, pp. 981–993, 2007.

[94] S. Mitra, S. Fusi, and G. Indiveri, "Real-time classification of complex patterns using spike-based learning in neuromorphic VLSI," *IEEE Trans. on Biomedical Circuits and Sytems*, vol. 3, no. 1, pp. 32–42, 2009.

[95] J. Seo, B. Brezzo, Y. Liu, B. D. Parker, S. K. Esser, R. K. Montoye, B. Rajendran, J. A. Tierno, L. Chang, D. S. Modha, and D. J. Friedman, "A 45 nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons," *Proc. IEEE Custom Integr. Circuits Conf.*, 2011.

[96] S. Sheik, M. Coath, G. Indiveri, S. L. Denham, T. Wennekers, and E. Chicca, "Emergent auditory feature tuning in a real-time neuromorphic VLSI system," *Front. Neurosci.*, vol. 6, p. 17, 2012.

[97] E. Neftci, J. Binas, U. Rutishauser, E. Chicca, G. Indiveri, and R. J. Douglas, "Synthesizing cognition in neuromorphic electronic systems," *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 37, pp. E3468–E3476, 2013.

[98] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri, "Neuromorphic electronic circuits for building autonomous cognitive systems," *Proceedings of IEEE*, vol. 102, no. 9, pp. 1367–1388, 2014.

[99] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, and G. Indiveri, "A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses," *Front. Neurosci.*, vol. 9, p. 141, 2015.

[100] S. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The Spinnaker project," *Proc. IEEE*, vol. 102, no. 5, pp. 652–665, 2014.

[101] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J.-M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen, "Neurogrid: a mixed-analog-digital multichip system for large-scale neural simuations," *Proc. IEEE*, vol. 102, no. 5, pp. 699–716, 2014.

[102] J. Schemmel, D. Briiderle, A. Griibland, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," *Proc. IEEE Int. Symp. Circuits Syst.*, pp. 1947–1950, 2010.

[103] S. Furber, "Large-scale neuromorphic computing systems," *J. Neural Eng.*, vol. 13, no. 5, p. 051001, 2016.

[104] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. N. Taba, M. Beakes, B. Brezzo, J. B. Kuang, R. Manohar, W. P. Risk, B. Jackson, and D. S. Modha, "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 10, pp. 1537–1557, 2015.

[105] J. M. Brader, W. Senn, and S. Fusi, "Learning real-world stimuli in a neural network with spike-driven synaptic dynamics," *Neural computation*, vol. 19, no. 11, pp. 2881–2912, 2007.

[106] N. Qiao and G. Indiveri, "Scaling mixed-signal neuromorphic processors to 28 nm FD-SOI technologies," *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, pp. 552–555, 2016.

[107] D. Kuzum, S. Yu, and H.-S. P. Wong, "Synaptic electronics: materials, devices and applications," *Nanotechnology*, vol. 24, no. 38, p. 382001, 2013.

[108] G. Indiveri, B. Linares-Barranco, R. Legenstein, G. Deligeorgis, and T. Prodromakis, "Integration of nanoscale memristor synapses in neuromorphic computing architectures," *Nanotechnology*, vol. 24, p. 384010, 2013.

[109] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. L. Gallo, K. Moon, J. Woo, H. Hwang, and Y. Leblebici, "Neuromorphic computing using non-volatile memory," *Advances in Physics: X*, vol. 2, no. 1, pp. 89–124, 2017.

[110] S. Yu, "Neuro-inspired computing with emerging nonvolatile memory," *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, 2018.

[111] G.-Q. Bi and M.-M. Poo, "Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and post-synaptic cell type," *J. Neurosci.*, vol. 18, no. 24, pp. 10 464–10 472, 1998.

[112] L. F. Abbott and S. B. Nelson, "Synaptic plasticity: taming the beast," *Nature Neurosci.*, vol. 3, no. 11, pp. 1178–1183, 2000.

[113] D. Kuzum, R. G. D. 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, 2012.

[114] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Lett.*, vol. 10, no. 4, pp. 1297–1301, 2010.

[115] 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," *IEEE Trans. on Electron Devices*, vol. 58, no. 8, pp. 2729–2737, 2011.

[116] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. De Salvo, "Phase change memory as synapse for ultra-dense neuromorphic systems: application to complex visual pattern extraction," *IEEE IEDM Tech. Dig.*, pp. 79–82, 2011.

[117] S. Ambrogio, N. Ciocchini, M. Laudato, V. Milo, A. Pirovano, P. Fantini, and D. Ielmini, "Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses," *Frontiers Neurosci.*, vol. 10, p. 56, 2016.

[118] G. Pedretti, V. Milo, S. Ambrogio, R. Carboni, S. Bianchi, A. Calderoni, N. Ramaswamy, A. S. Spinelli, and D. Ielmini, "Memristive neural network for on-line learning and tracking with brain-inspired spike timing dependent plasticity," *Sci. Rep.*, vol. 7, p. 5288, 2017.

[119] Z.-Q. Wang, S. Ambrogio, S. Balatti, and D. Ielmini, "A 2-transistor/1-resistor artificial synapse capable of communication and stochastic learning for neuromorphic systems," *Front. Neurosci.*, vol. 8, p. 438, 2015.

[120] S. Kim, M. Ishii, S. Lewis, T. Perri, M. BrightSky, W. Kim, R. Jordan, G. W. Burr, N. Sosa, A. Ray, J.-P. Han, C. Miller, K. Hosokawa, and C. Lam, "NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning," *IEEE IEDM Tech. Dig.*, pp. 443–446, 2015.

[121] M. V. Nair, L. K. Muller, and G. Indiveri, "A differential memristive synapse circuit for on-line learning in neuromorphic computing systems," *Nano Futures*, vol. 1, p. 035003, 2017.

[122] 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," *IEEE IEDM Tech. Dig.*, pp. 239–242, 2012.

[123] M. Suri, O. Bichler, D. Querlioz, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, and B. De Salvo, "CBRAM devices as binary synapses for low-power stochastic neuromorphic systems: Auditory (cochlea) and visual (retina) cognitive processing applications," *IEEE IEDM Tech. Dig.*, pp. 235–238, 2012.

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

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

[126] T. Serrano-Gotarredona, T. Masquelier, T. Prodromakis, G. Indiveri, and B. Linares-Barranco, "STDP and STDP variations with memristors for spiking neuromorphic learning systems," *Front. Neurosci.*, vol. 7, p. 2, 2013.

[127] S. Ambrogio, S. Balatti, V. Milo, R. Carboni, Z. Wang, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Neuromorphic learning and recognition with one-transistor-one-resistor synapses and bistable metal oxide RRAM," *IEEE Trans. Electron Devices*, vol. 63, no. 4, pp. 1508–1515, 2016.

[128] E. Covi, S. Brivio, A. Serb, T. Prodromakis, M. Fanciulli, and S. Spiga, “Analog memristive synapse in spiking networks implementing unsupervised learning,” *Front. Neurosci.*, vol. 10, p. 482, 2016.

[129] V. Milo, G. Pedretti, R. Carboni, A. Calderoni, N. Ramaswamy, S. Ambrogio, and D. Ielmini, “Demonstration of hybrid CMOS/RRAM neural networks with spike time/rate-dependent plasticity,” *IEEE IEDM Tech. Dig.*, pp. 440–443, 2016.

[130] T. Werner, E. Vianello, O. Bichler, D. Garbin, D. Cattaert, B. Yvert, B. De Salvo, and L. Perniola, “Spiking neural networks based on OxRAM synapses for real-time unsupervised spike sorting,” *Front. Neurosci.*, vol. 10, p. 474, 2016.

[131] A. Serb, J. Bill, A. Khiat, R. Berdan, R. Legenstein, and T. Prodromakis, “Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses,” *Nat. Commun.*, vol. 7, p. 12611, 2016.

[132] Z. Wang, S. Joshi, S. Savel’ev, W. Song, R. Midya, Y. Li, M. Rao, P. Yan, S. Asapu, Y. Zhuo, H. Jiang, P. Lin, C. Li, J. H. Yoon, N. K. Upadhyay, J. Zhang, M. Hu, J. P. Strachan, M. Barnell, Q. Wu, H. Wu, R. S. Williams, Q. Xia, and J. J. Yang, “Fully memristive neural networks for pattern classification with unsupervised learning,” *Nat. Electronics*, vol. 1, pp. 137–145, 2018.

[133] P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, “Rate, timing, and cooperativity jointly determine cortical synaptic plasticity,” *Neuron*, vol. 32, pp. 1149–1164, 2001.

[134] J.-P. Pfister and W. Gerstner, “Triplets of spikes in a model of spike timing-dependent plasticity,” *The Journal of Neuroscience*, vol. 26, no. 38, pp. 9673–9682, 2006.

[135] J. Gjorgjieva, C. Clopath, J. Audet, and J.-P. Pfister, “A triplet spike-timing-dependent plasticity model generalizes the Bienenstock-Cooper-Munro rule to higher-order spatiotemporal correlations,” *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 48, pp. 19 383–19 388, 2011.

[136] H. Z. Shouval, “What is the appropriate description level for synaptic plasticity?” *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 48, pp. 19 103–19 104, 2011.

[137] K. Seo, I. Kim, S. Jung, M. Jo, S. Park, J. Park, J. Shin, K. Biju, J. Kong, K. Lee, B. Lee, and H. Hwang, “Analog memory and spike-timing-dependent plasticity characteristics of a nanoscale titanium oxide bilayer resistive switching device,” *Nanotechnology*, vol. 22, p. 254023, 2011.

[138] Y. Li, Y. Zhong, J. Zhang, L. Xu, Q. Wang, H. Sun, H. Tong, X. Cheng, and X. Miao, “Activity-dependent synaptic plasticity of a chalcogenide electronic synapse for neuromorphic systems,” *Sci. Rep.*, vol. 4, p. 4906, 2014.

[139] W. He, K. Huang, N. Ning, K. Ramanathan, G. Li, Y. Jiang, J. Y. Sze, L. Shi, R. Zhao, and J. Pei, “Enabling an integrated rate-temporal learning scheme on memristor,” *Sci. Rep.*, vol. 4, p. 4755, 2014.

[140] M. Ziegler, C. Riggert, M. Hansen, T. Bartsch, and H. Kohlstedt, “Memristive Hebbian plasticity model: device requirements for the emulation of Hebbian plasticity based on memristive devices,” *IEEE Trans. on Biomedical Circuits and Systems*, vol. 9, no. 2, pp. 197–206, 2015.

[141] M. Hansen, F. Zahari, M. Ziegler, and H. Kohlstedt, “Double-barrier memristive devices for unsupervised learning and pattern recognition,” *Front. Neurosci.*, vol. 11, p. 91, 2017.

[142] S. Aghnout, G. Karimi, and M. R. Azghadi, “Modeling triplet spike-timing-dependent plasticity using memristive devices,” *J. Comput. Electron.*, vol. 16, pp. 401–410, 2017.

[143] V. Milo, G. Pedretti, R. Carboni, A. Calderoni, N. Ramaswamy, S. Ambrogio, and D. Ielmini, “A 4-transistors/1-resistor hybrid synapse based on resistive switching memory (RRAM) capable of spike-rate-dependent plasticity (SRDP),” *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 12, pp. 2806–2815, 2018.

[144] E. Covi, R. George, J. Frascaroli, S. Brivio, C. Mayr, H. Mostafa, G. Indiveri, and S. Spiga, "Spike-driven threshold-based learning with memristive synapses and neuromorphic silicon neurons," *Journal of Physics D: Applied Physics*, vol. 51, no. 34, p. 344003, 2018.

[145] S. La Barbera, D. Vuillaume, and F. Alibart, "Filamentary switching: synaptic plasticity through device volatility," *ACS Nano*, vol. 9, no. 1, pp. 941–949, 2015.

[146] Z. Wang, S. Joshi, S. E. Savel'ev, H. Jiang, R. Midya, P. Lin, M. Hu, N. Ge, J. P. Strachan, Z. Li, Q. Wu, M. Barnell, G.-L. Li, H. L. Xin, R. S. Williams, Q. Xia, and J. J. Yang, "Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing," *Nat. Mater.*, vol. 16, pp. 101–108, 2017.

[147] M. D. Pickett, G. Medeiros-Ribeiro, and R. S. Williams, "A scalable neuristor built with Mott memristors," *Nat. Mater.*, vol. 12, pp. 114–117, 2012.

[148] C. D. Wright, P. Hosseini, and J. A. Vazquez Diosdado, "Beyond von-Neumann computing with nanoscale phase-change memory devices," *Adv. Funct. Mater.*, vol. 23, pp. 2248–2254, 2013.

[149] T. Tuma, A. Pantazi, M. Le Gallo, A. Sebastian, and E. Eleftheriou, "Stochastic phase-change neurons," *Nat. Nanotechnology*, vol. 11, pp. 693–699, 2016.

[150] D. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

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

[152] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[153] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Proc. Advances in Neural Information Processing Systems*, vol. 25, pp. 1090–1098, 2012.

[154] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[155] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Proc. Advances in Neural Information Processing Systems*, vol. 27, pp. 3104–3112, 2014.

[156] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuska, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, 2011.

[157] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.

[158] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–489, 2016.

[159] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, pp. 354–359, 2017.

[160] H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainerand,

N. Jojic, S. W. Scherer, B. J. Blencowe, and B. J. Frey, “The human splicing code reveals new insights into the genetic determinants of disease,” *Science*, vol. 347, no. 6218, p. 1254806, 2015.

[161] S. Park, M. Chu, J. Kim, J. Noh, M. Jeon, B. H. Lee, H. Hwang, B. Lee, and B.-G. Lee, “Electronic system with memristive synapses for pattern recognition,” *Sci. Rep.*, vol. 5, p. 10123, 2015.

[162] P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang, W. Zhang, Q. Zhang, N. Deng, L. Shi, H.-S. P. Wong, and H. Qian, “Face classification using electronic synapses,” *Nature Commun.*, vol. 8, p. 15199, 2017.

[163] S. Yu, Z. Li, P.-Y. Chen, H. Wu, B. Gao, D. Wang, W. Wu, and H. Qian, “Binary neural network with 16 Mb RRAM macro chip for classification and online training,” *IEEE IEDM Tech. Dig.*, pp. 416–419, 2016.

[164] G. W. Burr, R. M. Shelby, C. D. Nolfo, J. W. Jang, R. S. Shenoy, and P. Narayanan, “Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element,” *IEEE IEDM Tech. Dig.*, pp. 29.5.1–29.5.4, 2014.

[165] G. W. Burr, R. M. Shelby, S. Sidler, C. di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. 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,” *IEEE Trans. on Electron Devices*, vol. 62, no. 11, pp. 3498–3507, 2015.

[166] “MNIST handwritten digits dataset. [Online]. Available: http://yann.lecun.com/exdb/mnist/.”

[167] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. P. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, and G. W. Burr, “Equivalent-accuracy accelerated neural-network training using analogue memory,” *Nature*, vol. 558, no. 7708, pp. 60–67, 2018.

[168] A. Krizhevsky, “Learning of multiple layers of features from tiny images. (Ch 3), https://www.cs.toronto.edu/ kriz/cifar.html,” 2009.

[169] B. Chen, F. Cai, J. Zhou, W. Ma, P. Sheridan, and W. D. Lu, “Efficient in-memory computing architecture based on crossbar arrays,” *IEEE IEDM Tech. Dig.*, pp. 17.5.1–17.5.4, 2015.

[170] J. Borghetti, G. S. Snider, P. J. Kuekes, J. J. Yang, D. R. Stewart, and R. S. Williams, “’Memristive’ switches enable ’stateful’ logic operations via material implication,” *Nature*, vol. 464, pp. 873–876, 2010.

[171] S. Balatti, S. Ambrogio, and D. Ielmini, “Normally-off logic based on resistive switches – Part I: Logic gates,” *IEEE Trans. Electron Devices*, vol. 62, no. 6, pp. 1831–1838, 2015.

[172] ——, “Normally-off logic based on resistive switches – Part II: Logic circuits,” *IEEE Trans. Electron Devices*, vol. 62, no. 6, pp. 1839–1847, 2015.

[173] M. Cassinerio, N. Ciocchini, and D. Ielmini, “Logic computation in phase change materials by threshold and memory switching,” *Adv. Mater.*, vol. 25, pp. 5975–5980, 2013.

[174] P. Hosseini, A. Sebastian, N. Papandreou, C. D. Wright, and H. Bhaskaran, “Accumulation-based computing using phase-change memories with FET access devices,” *IEEE Electron Device Lett.*, vol. 36, no. 9, pp. 975–977, 2015.

[175] A. Sebastian, T. Tuma, N. Papandreou, M. Le Gallo, L. Kull, T. Parnell, and E. Eleftheriou, “Temporal correlation detection using computational phase-change memory,” *Nat. Commun.*, vol. 8, p. 1115, 2017.

[176] S. N. Truong and K.-S. Min, “New memristor-based crossbar array architecture with 50-% area reduction and 48-% power saving for matrix-vector multiplication of analog neuromorphic computing,” *J. Semicond. Technol. Sci.*, vol. 14, pp. 356–363, 2014.

[177] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Davila, C. E. Graves, Z. Li, J. P. Strachan, P. Lin, Z. Wang, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia, “Analogue signal and image processing with large memristor crossbars,” *Nat. Electron.*, vol. 1, pp. 52–59, 2018.

[178] R. P. Feynman, “Simulating physics with computers,” *International Journal of Theoretical Physics*, vol. 21, no. 6/7, pp. 467–488, 1982.

[179] J. Preskill, “Quantum computing: pro and con,” *Proc. R. Soc. Lond. A*, vol. 454, pp. 469–486, 1998.

[180] J. N. Eckstein and J. Levy, “Materials issues for quantum computation,” *MRS Bulletin*, vol. 38, no. 10, pp. 783–789, 2013.

[181] T. D. Ladd, F. Jelezko, R. Laflamme, Y. Nakamura, C. Monroe, and J. L. O’Brien, “Quantum computers,” *Nature*, vol. 464, pp. 45–53, 2010.

[182] P. W. Shor, “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer,” *SIAM J. Comput.*, vol. 26, no. 5, pp. 1484–1509, 1997.

[183] L. Grover, “A fast quantum mechanical algorithm for database search,” *Proceedings on 28th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 212–219, 1996.

[184] W. Maass, “Noise as resource for computation and learning in networks of spiking neurons,” *Proc. IEEE*, vol. 102, no. 5, pp. 860–880, 2014.

[185] J. Preskill, “Quantum computing in the NISQ era and beyond,” *Quantum*, vol. 2, p. 79, 2018.

[186] E. Prati, “Quantum neuromorphic hardware for quantum artificial intelligence,” *Journal of Physics: Conference Series*, vol. 880, no. 1, pp. 1–6, 2017.

[187] C. J. Ballance, T. P. Harty, N. M. Linke, M. A. Sepiol, and D. M. Lucas, “High-fidelity quantum logic gates using trapped-ion hyperfine qubits,” *Phys. Rev. Lett.*, vol. 117, p. 060504, 2016.

[188] V. M. Schäfer, C. J. Ballance, K. Thirumalai, L. J. Stephenson, T. G. Ballance, A. M. Steane, and D. M. Lucas, “Fast quantum logic gates with trapped-ion qubits,” *Nature*, vol. 555, pp. 75–78, 2018.

[189] L. Childress and R. Hanson, “Diamond NV centers for quantum computing and quantum networks,” *MRS Bulletin*, vol. 38, no. 2, pp. 134–138, Feb. 2013.

[190] V. Acosta and P. Hemmer, “Nitrogen-vacancy centers: physics and applications,” *MRS Bulletin*, vol. 38, no. 2, pp. 127–130, Feb. 2013.

[191] P. C. Maurer, G. Kucsko, C. Latta, L. Jiang, N. Y. Yao, S. D. Bennett, F. Pastawski, D. Hunger, N. Chisholm, M. Markham, D. J. Twitchen, J. J. Cirac, and M. D. Lukin, “Room-temperature quantum bit memory exceeding one second,” *Science*, vol. 336, no. 6086, pp. 1283–1286, 2012.

[192] M. A. Eriksson, S. N. Coppersmith, and M. G. Lagally, “Semiconductor quantum dot qubits,” *MRS Bulletin*, vol. 38, no. 10, pp. 794–801, Oct. 2013.

[193] W. D. Oliver and P. B. Welander, “Materials in superconducting quantum bits,” *MRS Bulletin*, vol. 38, no. 10, pp. 816–825, Oct. 2013.

[194] C. G. Almudever, L. Lao, X. Fu, N. Khammassi, I. Ashraf, D. Iorga, S. Varsamopoulos, C. Eichler, A. Wallraff, L.Geck, A. Kruth, J. Knoch, H. Bluhm, and K. Bertels, “The engineering challenges in quantum computing,” *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 837–845, 2017.

[195] V. Mourik, K. Zuo, S. M. Frolov, S. R. Plissard, E. P. A. M. Bakkers, and L. P. Kouwenhoven, “Signatures of Majorana fermions in hybrid superconductor-semiconductor nanowire devices,” *Science*, vol. 336, no. 6084, pp. 1003–1007, 2012.

[196] A. Stern and N. H. Lindner, “Topological quantum computation–from basic concepts to first experiments,” *Science*, vol. 339, no. 6124, pp. 1179–1184, 2013.

[197] “https://quantumexperience.ng.bluemix.net/qx/experience.”

[198] “https://spectrum.ieee.org/tech-talk/computing/hardware/ibmedges-closer-to-quantum-supremacy-with-50qubit-processor.”

[199] “https://spectrum.ieee.org/tech-talk/computing/hardware/intels-49qubit-chip-aims-for-quantum-supremacy.”

[200] “https://ai.googleblog.com/2018/03/a-preview-of-bristlecone-googles-new.html.”

[201] R. Srivastava, I. Choi, and T. Cook, “The commercial prospects of quantum computing. Technical Report. NQIT,” 2016.

[202] A. Gheorghiu, A. Datta, C. Wade, D. Nadlinger, D. O’Brien, E. Geurtsen, E. Kassa, F. Sweeney, H. Rowlands, I. Walmsley, I. Choi, J. Joo, J. Smith, J. Becker, N. Walk, N. de Beaudrap, P. Leek, P. Wallden, P. Inglesant, R. Deshmukh, R. Srivastava, S. Benjamin, W. Zhang, W. Hensinger, and X. Yuan, “NQIT Annual Report 2018,” 2018.

[203] D. J. Bernstein and T. Lange, “Post-quantum cryptography,” *Nature*, vol. 549, pp. 188–194, 2017.

[204] M.-M. Poo and L. Wang, “Andrew Chi-Chih Yao: the future of quantum computing,” *National Science Review*, vol. 5, pp. 598–602, 2018.

[205] D. Ielmini and V. Milo, “Physics-based modeling approaches of resistive switching devices for memory and in-memory computing applications,” *J Comput Electron*, vol. 16, pp. 1121–1143, 2017.

[206] 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,” *IEEE Trans. Electron Devices*, vol. 59, no. 9, pp. 2468–2475, 2012.

[207] S. Kim, S.-J. Kim, K. M. Kim, S. R. Lee, M. Chang, E. Cho, Y.-B. Kim, C. J. Kim, U.-I. Chung, and I.-K. Yoo, “Physical electro-thermal model of resistive switching in bi-layered resistance change memory,” *Sci. Rep.*, vol. 3, p. 1680, 2013.

[208] S. Kim, S. H. Choi, and W. Lu, “Comprehensive physical model of dynamic resistive switching in an oxide memristor,” *ACS Nano*, vol. 8, no. 3, pp. 2369–2376, 2014.

[209] P. Huang, X. Liu, B. Chen, H. Li, Y. Wang, Y. X. Deng, K. L. Wei, L. Zeng, B. Gao, G. Du, X. Zhang, and J. F. Kang, “A physics-based compact model of metal-oxide-based RRAM DC and AC operations,” *IEEE Trans. Electron Devices*, vol. 60, no. 12, pp. 4090–4097, 2013.

[210] L. Larcher, F. M. 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,” *IEEE Trans. Electron Devices*, vol. 61, no. 8, pp. 2668–2673, 2014.

[211] H. Li, P. Huang, B. Gao, B. Chen, X. Liu, and J. Kang, “A SPICE model of resistive random access memory for large-scale memory array simulation,” *IEEE Electron Device Lett.*, vol. 35, no. 2, pp. 211–213, 2014.

[212] P. Huang, D. Zhu, S. Chen, Z. Zhou, Z. Chen, B. Gao, L. Liu, X. Liu, and J. F. Kang, “Compact model of HfO$_x$-based electronic synaptic devices for neuromorphic computing,” *IEEE Trans. Electron Devices*, vol. 64, no. 2, pp. 614–621, 2017.

[213] H. D. Lee, B. Magyari-Köpe, and Y. Nishi, "Model of metallic filament formation and rupture in NiO for unipolar switching," *Phys. Rev. B*, vol. 81, p. 193202, 2010.

[214] S.-G. Park, B. Magyari-Köpe, and Y. Nishi, "Impact of oxygen vacancy ordering on the formation of a conductive filament in TiO_2 for resistive switching memory," *IEEE Electron Device Lett.*, vol. 32, no. 2, pp. 197–199, 2011.

[215] D. Ielmini, F. Nardi, and C. Cagli, "Physical models of size-dependent nanofilament formation and rupture in NiO resistive switching memories," *Nanotechnology*, vol. 22, no. 25, p. 254022, 2011.

[216] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Statistical fluctuations in HfO_x resistive-switching memory: Part I – Set/reset variability," *IEEE Trans. Electron Devices*, vol. 61, no. 8, pp. 2912–2919, 2014.

[217] D. Ielmini, S. Larentis, and S. Balatti, "Physical modeling of voltage-driven resistive switching in oxide RRAM," *IEEE International Integrated Reliability Workshop (IIRW) Final Report*, pp. 9–15, 2012.

[218] S. Clima, Y. Chen, A. Fantini, L. Goux, R. Degraeve, B. Govoreanu, G. Pourtois, and M. Jurczak, "Intrinsic tailing of resistive states distributions in amorphous HfO_x and TaO_x based resistive random access memories," *IEEE Electron Device Lett.*, vol. 36, no. 8, pp. 769–771, 2015.

[219] S. Ambrogio, V. Milo, Z. Wang, S. Balatti, and D. Ielmini, "Analytical modeling of current overshoot in oxide-based resistive switching memory (RRAM)," *IEEE Electron Device Lett.*, vol. 37, no. 10, pp. 1268–1271, 2016.

[220] S. Balatti, S. Ambrogio, D. C. Gilmer, and D. Ielmini, "Set variability and failure induced by complementary switching in bipolar RRAM," *IEEE Electron Device Lett.*, vol. 34, no. 7, pp. 861–863, 2013.

[221] S. Balatti, S. Ambrogio, Z.-Q. Wang, S. Sills, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Understanding pulsed-cycling variability and endurance in HfO_x RRAM," *IEEE International Reliability Physics Symposium (IRPS)*, pp. 5B.3.1–5B.3.6, 2015.

[222] S. Balatti, S. Ambrogio, Z. Wang, and D. Ielmini, "True random number generation by variability of resistive switching in oxide-based devices," *IEEE Journal of Emerging and Selected Topics in Circuits and Systems (JETCAS)*, vol. 5, no. 2, pp. 214–221, 2015.

[223] S. Balatti, S. Ambrogio, R. Carboni, V. Milo, Z. Wang, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Physical unbiased generation of random numbers with coupled resistive switching devices," *IEEE Trans. Electron Devices*, vol. 63, no. 5, pp. 2029–2035, 2016.

[224] T. Masquelier and S. J. Thorpe, "Unsupervised learning of visual features through spike timing dependent plasticity," *PLoS Comput Biol*, vol. 3, no. 2, p. e31, 2007.

[225] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers Comput. Neurosci.*, vol. 9, p. 99, 2015.

[226] E. L. Bienenstock, L. N. Cooper, and P. W. Munro, "Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex," *J. Neurosci.*, vol. 2, no. 1, pp. 32–48, 1982.

[227] M. F. Bear, "A synaptic basis for memory storage in the cerebral cortex," *Proc. Natl. Acad. Sci. USA*, vol. 93, no. 24, pp. 13 453–13 459, 1996.

[228] G. Indiveri, F. Corradi, and N. Qiao, "Neuromorphic architectures for spiking deep neural networks," *IEEE IEDM Tech. Dig.*, pp. 68–71, 2015.

[229] 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, p. 384012, 2013.

[230] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H.-S. P. Wong, "A low energy oxide-based electronic synaptic device for neuromorphic visual systems with tolerance to device variation," *Adv. Mater.*, vol. 25, no. 12, pp. 1774–1779, 2013.

[231] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. K. Gimzewski, and M. Aono, "Short-term plasticity and long-term potentiation mimicked in single inorganic synapses," *Nat. Mater.*, vol. 10, pp. 591–595, 2011.

[232] D. Ielmini and V. Milo, "Brain-inspired memristive neural networks for unsupervised learning," *Handbook of Memristor Networks, G. Sirakoulis and L. Chua eds., Springer*, 2018.

[233] D. Ielmini, S. Ambrogio, V. Milo, S. Balatti, and Z.-Q. Wang, "Neuromorphic computing with hybrid memristive/CMOS synapses for real-time learning," *IEEE Int. Symp. on Circuits and Systems (ISCAS)*, pp. 1386–1389, 2016.

[234] S. Ambrogio, S. Balatti, V. Milo, R. Carboni, Z. Wang, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Novel RRAM-enabled 1T1R synapse capable of low-power STDP via burst-mode communication and real-time unsupervised machine learning," *IEEE Symposium on VLSI Technology*, pp. 196–197, 2016.

[235] W. S. McCulloch and W. H. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math Biophys.*, vol. 5, pp. 115–133, 1943.

[236] M. Prezioso, F. Merrikh-Bayat, B. Hoskins, K. Likharev, and D. Strukov, "Self-adaptive spike-time-dependent plasticity of metal-oxide memristors," *Sci. Rep.*, vol. 6, p. 21331, 2016.

[237] A. Calderoni, S. Sills, and N. Ramaswamy, "Performance comparison of O-based and Cu-based ReRAM for high-density applications," *Proc. Int. Memory Workshop (IMW)*, pp. 1–4, 2014.

[238] Z. Wei, Y. Katoh, S. Ogasahara, Y. Yoshimoto, K. Kawai, Y. Ikeda, K. Eriguchi, K. Ohmori, and S. Yoneda, "True random number generator using current difference based on a fractional stochastic model in 40-nm embedded ReRAM," *IEEE IEDM Tech. Dig.*, pp. 107–110, 2016.

[239] C.-Y. Huang, W. C. Shen, Y.-H. Tseng, Y.-C. King, and C.-J. Lin, "A contact-resistive random-access-memory based true random number generator," *IEEE Electron Device Lett.*, vol. 33, no. 8, pp. 1108–1110, 2012.

[240] S. Gaba, P. Sheridan, J. Zhou, S. Choi, and W. Lu, "Stochastic memristive devices for computing and neuromorphic applications," *Nanoscale*, vol. 5, no. 13, pp. 5872–5878, 2013.

[241] S. Balatti, S. Ambrogio, Z. Wang, S. Sills, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Voltage-controlled cycling endurance of HfO_x-based resistive-switching memory (RRAM)," *IEEE Trans. Electron Devices*, vol. 62, no. 10, pp. 3365–3372, 2015.

[242] Z. Wang, S. Ambrogio, S. Balatti, S. Sills, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Postcycling degradation in metal-oxide bipolar resistive switching memory," *IEEE Trans. Electron Devices*, vol. 63, no. 11, pp. 4279–4287, 2016.

[243] J.-W. Jang, S. Park, G. W. Burr, H. Hwang, and Y.-H. Jeong, "Optimization of conductance change in $Pr_{1-x}Ca_xMnO_3$-based synaptic devices for neuromorphic systems," *IEEE Electron Device Lett.*, vol. 36, no. 5, pp. 457–459, 2015.

[244] 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," *IEEE IEDM Tech. Dig.*, pp. 665–668, 2014.

[245] A. C. Torrezan, J. P. Strachan, G. Medeiros-Ribeiro, and R. S. Williams, "Sub-nanosecond switching of a tantalum oxide memristor," *Nanotechnology*, vol. 22, no. 48, p. 485203, 2011.

[246] H. Y. Lee, Y. S. Chen, P. S. Chen, P. Y. Gu, C. W. Chen, W. P. Lin, H. W. Liu, Y. Y. Hsu, S. S. Sheu, P. C. Chiang, W. S. Chen, F. T. Chen, C. H. Lien, and M.-J. Tsai, "Evidence and solution of over-reset problem for HfO_x based resistive memory with sub-ns switching speed and high endurance," *IEEE IEDM Tech. Dig.*, pp. 460–463, 2010.

[247] T. Chouard and L. Venema, "Machine intelligence," *Nature*, vol. 521, no. 7553, p. 435, 2015.

[248] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[249] M. L. Minsky and S. A. Papert, "Perceptrons," *Cambridge, MA: MIT Press.*, 1969.

[250] P. Sheridan, W. Ma, and W. Lu, "Pattern recognition with memristor networks," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1078–1081, 2014.

[251] G. Pedretti, V. Milo, S. Ambrogio, R. Carboni, S. Bianchi, A. Calderoni, N. Ramaswamy, A. S. Spinelli, and D. Ielmini, "Stochastic learning in neuromorphic hardware via spike timing dependent plasticity with RRAM synapses," *IEEE Journal of Emerging Topics on Circuits and Systems (JETCAS)*, vol. 8, no. 1, pp. 77–85, 2018.

[252] N. Caporale and Y. Dan, "Spike timing-dependent plasticity: A Hebbian learning rule," *Annu. Rev. Neurosci.*, vol. 31, pp. 25–46, 2008.

[253] G. Pedretti, S. Bianchi, V. Milo, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Modeling-based design of brain-inspired spiking neural networks with RRAM learning synapses," *IEEE IEDM Tech. Dig.*, pp. 653–656, 2017.

[254] T. Masquelier, R. Guyonneau, and S. J. Thorpe, "Competitive STDP-based spike pattern learning," *Neural Computation*, vol. 21, no. 5, pp. 1259–1276, 2009.

[255] K. Nakazawa, T. J. McHugh, M. A. Wilson, and S. Tonegawa, "NMDA receptors, place cells and hippocampal spatial memory," *Nature Reviews Neuroscience*, vol. 5, pp. 361–372, 2004.

[256] D. J. Amit, "Modeling brain function: The world of attractor neural networks," *Cambridge University Press*, 1992.

[257] E. Rolls, "Attractor networks," *Wiley Interdisciplinary Rev. Cogn. Sci.*, vol. 1, no. 1, pp. 119–134, 2010.

[258] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Natl. Acad. Sci. USA (PNAS)*, vol. 79, pp. 2554–2558, 1982.

[259] ——, "Neurons with graded response have collective computational properties like those of two-state neurons," *Proc. Natl. Acad. Sci. USA (PNAS)*, vol. 81, pp. 3088–3092, 1984.

[260] M. Giulioni, P. Camilleri, M. Mattia, V. Dante, J. Braun, and P. Del Giudice, "Robust working memory in an asynchronously spiking neural network realized in neuromorphic VLSI," *Front. Neurosci.*, vol. 5, p. 149, 2012.

[261] M. Giulioni, F. Corradi, V. Dante, and P. Del Giudice, "Real time unsupervised learning of visual stimuli in neuromorphic VLSI systems," *Sci. Rep.*, vol. 5, p. 14730, 2015.

[262] S. G. Hu, Y. Liu, Z. Liu, T. P. Chen, J. J. Wang, Q. Yu, L. J. Deng, Y. Yin, and S. Hosaka, "Associative memory realized by a reconfigurable memristive Hopfield neural network," *Nat. Commun.*, vol. 6, p. 7522, 2015.

[263] S. B. Eryilmaz, D. Kuzum, R. Jeyasingh, S. B. Kim, M. BrightSky, C. Lam, and H.-S. P. Wong, "Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array," *Front. Neurosci.*, vol. 8, p. 205, 2014.

[264] X. Guo, F. Merrikh-Bayat, L. Gao, B. D. Hoskins, B. Linares-Barranco, F. Alibart, L. Theogarajan, C. Teuscher, and D. B. Strukov, "Modeling and experimental demonstration of a Hopfield network analog-to-digital converter with hybrid CMOS/memristor circuits," *Front. Neurosci.*, vol. 9, p. 488, 2015.

[265] V. Milo, D. Ielmini, and E. Chicca, "Attractor networks and associative memories with STDP learning in RRAM synapses," *IEEE IEDM Tech. Dig.*, pp. 263–266, 2017.

[266] N. Diederich, T. Bartsch, H. Kohlstedt, and M. Ziegler, "A memristive plasticity model of voltage-based STDP suitable for recurrent bidirectional neural networks in the hippocampus," *Sci. Rep.*, vol. 8, p. 9367, 2018.

[267] V. Milo, E. Chicca, and D. Ielmini, "Brain-inspired recurrent neural network with plastic RRAM synapses," *IEEE Int. Symp. on Circuits and Systems (ISCAS)*, pp. 1–5, 2018.

[268] V. Milo, G. Pedretti, M. Laudato, A. Bricalli, E. Ambrosi, S. Bianchi, E. Chicca, and D. Ielmini, "Resistive switching synapses for unsupervised learning in feed-forward and recurrent neural networks," *IEEE Int. Symp. on Circuits and Systems (ISCAS)*, pp. 1–5, 2018.

[269] J. J. Hopfield and D. W. Tank, "Neural computation of decisions in optimization problems," *Biol. Cybern.*, vol. 52, pp. 141–152, 1985.

[270] J. J. Hopfield, "Searching for memories, Sudoku, implicit check bits, and the iterative use of not-always-correct rapid neural computation," *Neural Computation*, vol. 20, pp. 1119–1164, 2008.

[271] D. O. Hebb, "The organization of behavior: A neuropsychological theory," *John Wiley, New York*, 1949.

[272] I. P. Pavlov, "Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex," *Oxford University Press, London*, 1927.

[273] Y. V. Pershin and M. Di Ventra, "Experimental demonstration of associative memory with memristive neural networks," *Neural Netw.*, vol. 23, no. 7, pp. 881–886, 2010.

[274] M. Ziegler, R. Soni, T. Patelczyk, M. Ignatov, T. Bartsch, P. Meuffels, and H. Kolhstedt, "An electronic version of Pavlov's dog," *Adv. Funct. Mater.*, vol. 22, no. 13, pp. 2744–2749, 2012.

[275] Z.-H. Tan, X.-B. Yin, R. Yang, S.-B. Mi, C.-L. Jia, and X. Guo, "Pavlovian conditioning demonstrated with neuromorphic memristive devices," *Sci. Rep.*, vol. 7, p. 713, 2017.