TOPICAL REVIEW • OPEN ACCESS

HfO_2-based resistive switching memory devices for neuromorphic computing

To cite this article: S Brivio *et al* 2022 *Neuromorph. Comput. Eng.* **2** 042001

View the article online for updates and enhancements.

NEUROMORPHIC
Computing and Engineering

OPEN ACCESS

RECEIVED
2 February 2022

REVISED
3 August 2022

ACCEPTED FOR PUBLICATION
7 September 2022

PUBLISHED
21 October 2022

TOPICAL REVIEW

HfO_2-based resistive switching memory devices for neuromorphic computing

S Brivio[1], S Spiga[1,*] and D Ielmini[2,*]

[1] CNR-IMM, Unit of Agrate Brianza, via C Olivetti 2, 20041 Agrate Brianza, Italy
[2] Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano and IUNET, piazza L da Vinci 32, 20133 Milano, Italy
[*] Authors to whom any correspondence should be addressed.

E-mail: sabina.spiga@mdm.imm.cnr.it and daniele.ielmini@polimi.it

Keywords: hafnium oxide, RRAM, neuromorphic computing, artificial neural networks, spiking neural networks, in-memory computing, volatile and nonvolatile memory

Abstract

HfO_2-based resistive switching memory (RRAM) combines several outstanding properties, such as high scalability, fast switching speed, low power, compatibility with complementary metal-oxide-semiconductor technology, with possible high-density or three-dimensional integration. Therefore, today, HfO_2 RRAMs have attracted a strong interest for applications in neuromorphic engineering, in particular for the development of artificial synapses in neural networks. This review provides an overview of the structure, the properties and the applications of HfO_2-based RRAM in neuromorphic computing. Both widely investigated applications of nonvolatile devices and pioneering works about volatile devices are reviewed. The RRAM device is first introduced, describing the switching mechanisms associated to filamentary path of HfO_2 defects such as oxygen vacancies. The RRAM programming algorithms are described for high-precision multilevel operation, analog weight update in synaptic applications and for exploiting the resistance dynamics of volatile devices. Finally, the neuromorphic applications are presented, illustrating both artificial neural networks with supervised training and with multilevel, binary or stochastic weights. Spiking neural networks are then presented for applications ranging from unsupervised training to spatio-temporal recognition. From this overview, HfO_2-based RRAM appears as a mature technology for a broad range of neuromorphic computing systems.

1. Introduction

A major challenge in neuromorphic engineering is the design and development of novel devices which mimic the behavior of biological elements of a neural network, such as spiking neurons and learning synapses [1–3]. In this regard, the class of resistive (or memristive) devices, such as the resistive switching random access memory (RRAM) has attracted a good deal of interest for the simple structure, the low-power operation and the easy integration with the complementary metal-oxide semiconductor (CMOS) process flow [4–7]. The ability of controlling the device conductance by electrical stimuli, similar to the neuronal spikes causing potentiation and depression of a biological synapse, has spurred the development of artificial synapses based on RRAM devices [8–10].

The neuromorphic research has focused on two main directions, namely (i) the development of artificial neural networks (ANNs) aiming at high-accurate recognition of image, video and audio data [11], and (ii) the engineering of spiking neural networks (SNNs) to closely mimic the adaptability and high-energy efficiency of the human brain [12].

The good scaling behavior of RRAM devices in terms of both device area [13] and 3D integration [14] enables the implementation of high density of synapses needed in both deep learning architectures and brain-inspired circuits with high connectivity between neurons and synapses. As biological synapses weight the communication among neurons, in the same way, resistance states of nonvolatile RRAMs can modulate the connection among artificial neurons. Furthermore, RRAM devices can play the role of both memory and computing devices, thus allowing in-memory computing (IMC) where data are processed *in situ* within the memory array [15]. This is in line with the co-location of neurons and synapses in the human brain, which is a fundamental asset to achieve ultra low power consumption [16, 17]. More pioneering is the employment of volatile RRAM devices to implement, through their switching transients, the various dynamical components typically present in neural, in SNNs and in human brain [16, 18].

A significant challenge, however, is the identification of a mature RRAM technology which combines high flexibility of neuromorphic functions, e.g., spike integration in a neuron or analog potentiation/depression in a synapse, with good compatibility with the CMOS process flow to enable large scale integration of neuromorphic systems.

HfO_2 has been soon identified as a promising material for RRAMs, due to the controllable resistance switching arising from the oxygen exchange between the metallic electrode and the metal oxide layer [19]. HfO_2 based stacks have been proposed and optimized toward nonvolatile RRAM since more than 15 years [4, 6, 13, 19]. In addition, HfO_2 possesses a set of key properties, including (i) high dielectric constant for scalable CMOS technology, both for logic and charge trap memory devices [20, 21], (ii) controllable ferroelectricity/antiferroelectricity behavior for high density memory and for synaptic devices [22, 23], and (iii) high compatibility with the CMOS process. HfO_2 thus appears as a dielectric material of choice for microelectronic systems encompassing logic, memory and neuromorphic functions on the same chip. For this reason, HfO_2-based RRAM has been one of the most popular technology for early demonstrator of neuromorphic devices and circuits [8]. More recently, also HfO_2 based ferroelectric memories are gaining interest for novel computation schemes [22, 24]. It is worth mentioning that despite the use of HfO_2 for both technologies, there are substantial differences between HfO_2-based RRAM and ferroelectric memories in terms of real implemented complete material stacks and thermal budget needed during fabrication, integration strategy, as well as the programming conditions and working principle. Therefore, also depending on future advancements, they can be used to target different applications or functions, in different areas of the same chip.

This review article provides an overview of HfO_2-based RRAM devices and circuits for neuromorphic computing. The review is organized as follows: section 2 introduces the RRAM device structure and operation, addressing the defect formation mechanism, the electrically-induced switching and the impact of the electrode material in determining the nonvolatile or volatile behavior of the device. Section 3 illustrates the programming schemes for both off-line training, where high precision weights are mapped into the synaptic memory array, and on-line training for what concerns nonvolatile devices. Examples of use of volatile devices are reported in the same section. Section 4 addresses the computing schemes, including various types of ANNs and SNNs where HfO_2-based RRAMs serve the role of neuromorphic synapses. Section 5 discusses the current challenges, possible solutions and perspectives of HfO_2-based RRAM.

2. RRAM device operation and material stacks

2.1. Nonvolatile devices

Nonvolatile RRAMs are two terminals devices whose resistance can be reversibly changed between two or more values by application of electrical stimuli [6, 27]. The programmed resistance states are stable for a long retention time after the stimulus has been released. Both 1 resistor (1R) or 1 transistor–1 resistor (1T1R) configurations (see figure 1(a)) are used for characterization and implementation into memory or computing systems. The memory cell (1R) is based on a two terminal configuration where a switching medium (also composed by various materials) is sandwiched between two electrodes. The switching event can occur locally in a filament region shorting the two electrodes (filamentary switching) or over the entire cross-section of the cell (interface-type switching) [27, 28]. The switching mechanism of HfO_2 based RRAMs is usually reported as filamentary switching and the resistance change is ascribed to ionic movement and local redox reactions [28–31].

Regarding device operation, a preliminary forming operation is often required to activate the reversible switching in a fresh device, i.e. to create the filament shorting the two electrodes. The forming voltage is usually larger than the following switching voltages. The switching from high resistance (HRS) to low resistance state (LRS) is called set transition, while the reverse operation is called reset transition. The device resistance state can be read at voltages lower than those typically required for switching, so that the device resistance is not modified. In set and forming operations (usually performed at the same polarity), the current undergoes a rapid increase that may irreversibly damage the device. For this reason, a current limitation is applied through the measurement system or through the transistor of the 1T1R configuration operated in its saturation regime. The value of the maximum current allowed during the set operation determines the value of the reached LRS, enabling multilevel storage, as shown by different colors in figures 1(b) and (c) for the 1R and 1T1R configurations, respectively. It must be mentioned that any increase of the current compliance results

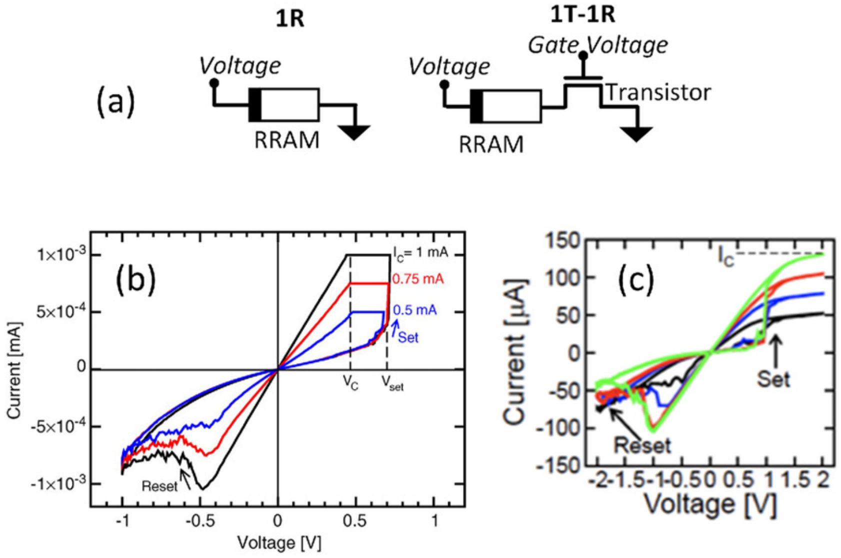

Figure 1. Types of RRAM devices and their operation. (a) Schematic for the 1R and 1T1R configuration. (b) Examples of set and reset operations in a 1R TiN–20 nm HfO$_x$–TiN RRAM device as a function of current compliance applied through the semiconductor device parameter analyzer. (c) Examples of set and reset operations in a 1T1R TiN/Ti/HfO$_2$/TiN-based RRAM cell. The current compliance I_c is applied by varying the gate voltage of the select transistor. (b) © 2012 IEEE. Reprinted, with permission, from [25]. (c) © 2021 IEEE. Reprinted, with permission, from [26].

in an increase of the energy consumption per programming operation. Moreover, the use of the 1T1R configuration (especially using an integrated transistor during device fabrication) provides a better control of the current compliance during forming and set operations, due to the limitation of parasitic capacitance effects [32], and thus allowing low power operation with respect to the use of 1R configuration only. In addition, the 1T1R configuration is beneficial for large array integration density since the transistor provide a selector limiting sneak path problems. Therefore the use of 1T1R configuration for HfO$_2$ RRAM has lead to the best device performance. Recently, nonvolatile HfO$_2$ RRAM have shown large endurance [33], retention up to 10 years [34, 35], multilevel operation [26, 36], excellent scalability down to 10 nm or less [13] demonstrated at single device level, integration in 1T1R large arrays and in combination with scaled CMOS technology nodes [26, 35, 37, 38], and possibility of integration in 3D arrays [7, 39].

HfO$_2$ RRAM devices have also been optimized through material engineering of the stack with the aim of improving electrical performances and, more recently, with the aim of achieving analog switching. Regarding the used material stack for HfO$_2$ based RRAM devices, the basic structure usually consists of a HfO$_2$ layer as switching medium (with thickness in the 5–30 nm range), and bottom and top electrodes formed by metals or nitrides, such as Pt, Au, W, Ni, TiN, TaN. One of the most used optimized stack includes the us of a Ti (or Hf) scavenging layer, between the HfO$_2$ layer grown by atomic layer deposition (ALD) and the top TiN electrode [13, 19, 33, 40, 41]. As shown in figure 2 in the case of Ti, the metal layer is easily oxidized by oxygen exchange between Ti and HfO$_2$, possibly also promoted by an additional post-deposition annealing. This phenomenon leads to the formation of a TiO$_x$ layer which serves as oxygen exchange layer during the following switching operation, as well as to the formation of a sub-stochiometric HfO$_x$ layer close to the top electrode and an asymmetrical oxygen vacancy profile in the oxide layer. Various scavenging metal layers have been studied in combination with HfO$_2$ grown by ALD. Ti or Hf scavenging layers and the created asymmetric oxygen vacancy profile in the oxide layer are beneficial in terms of reduction of the forming voltage [13, 42, 43], as shown in figure 2(c). Independently from the RRAM cell size (1 μm or 40 nm) the increase of Hf cap thickness leads to a decrease of the forming voltage.

Moreover, the insertion of Ti and Hf interlayers is also beneficial for the subsequent switching properties. In particular, the use of TiN/Ti/HfO$_2$/TiN or TiN/Hf/HfO$_2$/TiN (top electrode/scavenging metal/oxide/bottom electrode) stacks leads to the best results in terms of device endurance and retention. For instance, Chen *et al* [33] reported 10^{10} pulse endurance in TiN/10 nm Hf/5 nm HfO$_2$/TiN 1T1R RRAM devices. The energetic of oxygen vacancy creation and migration can be modified by HfO$_2$ doping with a trivalent metal [45, 46]. For instance, Al doping has been reported to influence uniformity, retention and resistance fluctuations [34, 47, 48]. All these effects may help implementing high precision multilevel storage useful for some kind of neuromorphic computing schemes as we will discuss in sections 3 and 4 [49–51].

Overall, the attempted materials science solution are many but a clear framework to optimize RRAM device toward a specific application has not been devised yet. We can notice that, recently, the use of multilayers

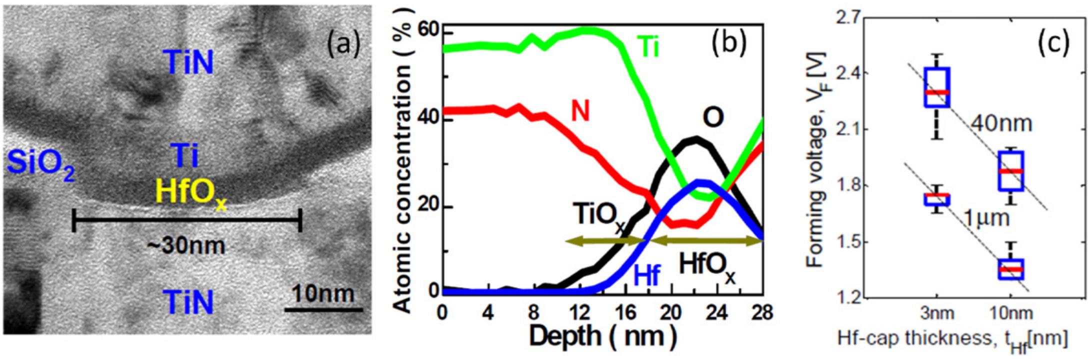

Figure 2. Effect of the inclusion of a scavenging layer at one metal/HfO_2 interface. (a) Transmission electron microscope image and (b) XPS depth profile of TiN/Ti/HfO_2/TiN stack. (c) Relationship between the forming voltage and the thickness of the Hf scavenging layer for TiN/Hf/5 nm HfO_2/TiN RRAM cells. The Hf layer is inserted between the HfO_2 film and the top TiN electrode. (a) © 2009 IEEE. Reprinted, with permission, from [44]. (b) © 2008 IEEE. Reprinted, with permission, from [19]. (c) © 20011 IEEE. Reprinted, with permission, from [13].

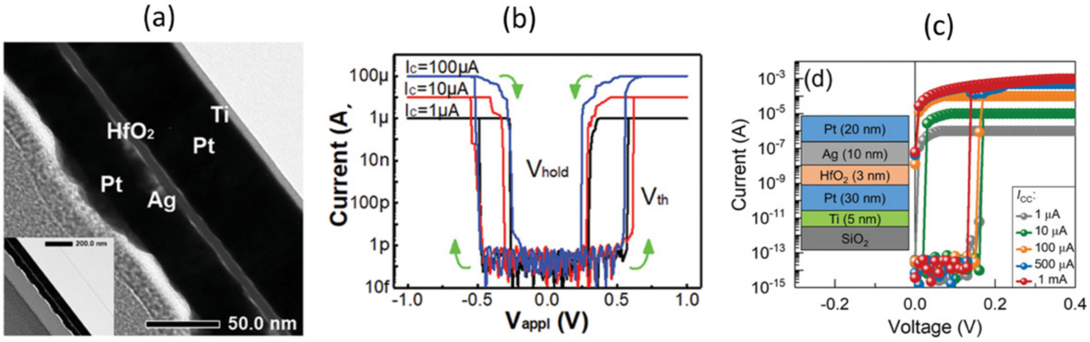

Figure 3. Volatile RRAM device scheme and operation. (a) Schematic of the Pt/Ag nanodots (top electrode)/HfO_2/Pt (bottom electrode) device, and (b) its representative measured DC *IV* characteristics. (c) Ag (top electrode)/HfO_2/Pt (bottom electrode) 1R volatile device: *IV* characteristics as a function of the current compliance. (a) and (b) reproduced from [66] CC BY 4.0. (c) reproduced from [55] CC BY 4.0.

has empirically demonstrated successful for the improvement of the analog modulation of the conductance in several works [11, 52–54]. In particular, some examples of successful realization of multilevel of analog RRAMs make use of multilayers stacks like Pt/HfO_x/TiO_x/HfO_x/TiO_x/TiN [52], Al/AlO_x/HfO_2/Ti/TiN [53], TiN/$HfAlO_x$/TaO_x/TiN [11] or TiN/TaO_x/HfO_x/TiN [54].

2.2. Volatile devices

Another class of RRAM devices, named also volatile memristors or diffusive memristors, is the one for which the retention of the LRS can span various time scales from ultra short time (ns) up to tens of ms or seconds [18, 55]. These devices are based on cation migration [28] and the device stacks are usually based on one active electrode (Ag or Cu), a solid electrolyte as switching medium (HfO_2, TiO_2, TaO_x, SiO_x) and an inert electrode (Pt, Au, W, TiN, Pd, carbon) [18, 56–59]. The use of symmetrical stacks like Ag/electrolyte/Ag [60, 61] or Pt/Ag-doped material/Pt [62] have been also proposed. Even if HfO_2 is not the only materials of choice for volatile devices, HfO_2-based volatile RRAM devices have been recently demonstrated by various groups and proposed for neuromorphic and computing applications [18, 55, 57, 61, 63].

Regarding the device operation, initially the cell is in the pristine HRS and the application of a voltage to the active electrode leads to the formation of a Ag (or Cu) conductive filament connecting the two electrode (LRS) thanks to the injection of cation from the active electrode material into the solid electrolyte. The filament self-dissolve once the applied voltage falls below a *hold* value [64, 65]. After the initial formation of a filament, which may be associate to a forming process, the volatile switching is achieved by applying a voltage over a *threshold* value (usually lower than forming voltage) for filament formation.

Figure 3(a) shows a transmission electron microscopy section image of Pt/Ag nanodots (top electrode)/HfO_2/Pt (bottom electrode) stack, while figure 3(b) reports the corresponding measured current-*vs*-voltage (*IV*) curves. Another example of volatile switching is reported in figure 3(c) for the Ag (top electrode)/HfO_2/Pt (bottom electrode) device. The *IV* curves show an abrupt and volatile switching, and the final LRS value can be controlled by the imposed current compliance. Usually symmetric stack structures (Ag/switching medium/Ag) show bi-directional volatile switching, while non-symmetric stack structure (Ag/switching medium/Pt) shows uni-directional volatile switching. On the other hand, bi-directional switching has been observed also for non-symmetric structures [55, 56, 66] and ascribed to residual Ag filament close to the inert electrode after forming operation. An example for the latter case is indeed reported in figure 3(b). Finally, it is possible to observe both volatile and nonvolatile switching in Ag- or Cu-based RRAMs especially if the device is operated at large compliance currents leading to the formation of a large and stable filament [57].

3. Programming schemes for nonvolatile and volatile RRAMs

The main usage of nonvolatile RRAMs in neuromorphic hardware is as synaptic weight. Weights can be stored in multiple conductance levels, described by non-overlapping distribution of values. In this case, we speak about multilevel operation. We refer to analog devices, instead, in case their conductance can be modulated through a continuum of values without identifying levels and corresponding distinct distributions. The programming methodology of nonvolatile synaptic weights is different in case the training is performed either on-line or off-line. Indeed, in case the training is performed off-line, weights have to be uploaded to the synaptic array as conductance values with high accuracy. In particular, for off-line training, it is possible to take advantage of program verify schemes, especially for 1T1R structures. On the contrary, in case of on-line learning, program verify schemes are hardly implemented. Furthermore, most of the training protocols and learning rules do not indicate absolute weight values (i.e. conductance values), but, in turn, prescribe weight changes relatively to the current value. Therefore, for on-line training, programming schemes devised to apply relative conductance changes instead of absolute conductance values are needed and they will be referred to as *weight or conductance update*, in the following. Both multilevel and analog programming can be implemented as weight updates. Despite multilevel or analog conductance modulations are generally considered as the best options for highly accurate ANN or SNN implementations, neural networks with binary weight demonstrate great computing potential and ease of implementation as discussed in details in the next section. For on-line training, binary weight can be also programmed in a stochastic manner.

For what concerns volatile devices, the research about their implementation in neuromorphic systems is still at its infancy. In the following, some possible uses are described which take advantage of short term internal or resistance evolution of the devices to emulate dynamical features of neural network or dynamical elements present in human brain.

3.1. Multilevel programming

For ANN with off-line training, accurate program-verify techniques are needed to store high-precision weights for running the network. In fact, the application of a single programming pulse typically results in a relatively large variation of conductance [9, 68, 69]. In addition, the conductance can also change after the programming pulse due to random telegraph noise (RTN) [34, 47, 70, 71], $1/f$ noise [72] and random walk effects [73], all contributing to the broadening of the conductance distribution. Finally, the weights stored in the array might be affected by a device-to-device variation, due to the physical differences in the device structures and geometries [74]. As a result, single-pulse programming operations are not suitable for synaptic weight storage in RRAM arrays. In general, 1R devices can be programmed in a multilevel fashion, as well, for instance by modulating the voltage applied either during the RESET or during SET operation in case of devices that do not require current limitation [10, 75, 76]. However, such programming method results in a higher variability than the programming through the current compliance provided by an integrated transistor. Therefore the use if 1T1R configuration is usually employed for an efficient multilevel programming as discussed in the following.

Figure 4 shows two program-verify techniques for RRAM devices with 1T1R structures [67]. In the incremental step program-verify algorithm (ISPVA, figure 4(a)), voltage pulses with incremental amplitude are applied at the top electrode terminal of the 1T1R structure, while the gate voltage is maintained fixed to control the compliance current [77]. As a result, the device can be gradually set to the desired level, the latter being selected via the gate voltage. On the other hand, the top electrode voltage is kept to a constant value in the incremental gate-voltage verify algorithm (IGVVA), while the gate voltage, hence the compliance current, is increased until the desired conductance level is reached [26]. Figure 4(c) shows the measured standard deviation σ_G of conductance as a function of the average conductance, $\langle G \rangle$. The programming variation σ_G is significantly decreased by the IGVVA technique with relatively small incremental voltage $\Delta V_G = 10$ mV, namely the IGVVA10 technique, which results in a σ_G between 2 μS and 5 μS for LRS. Further mitigation of the cycle-to-cycle and device-to-device variations of conductance can be achieved by redundancy and bit slicing

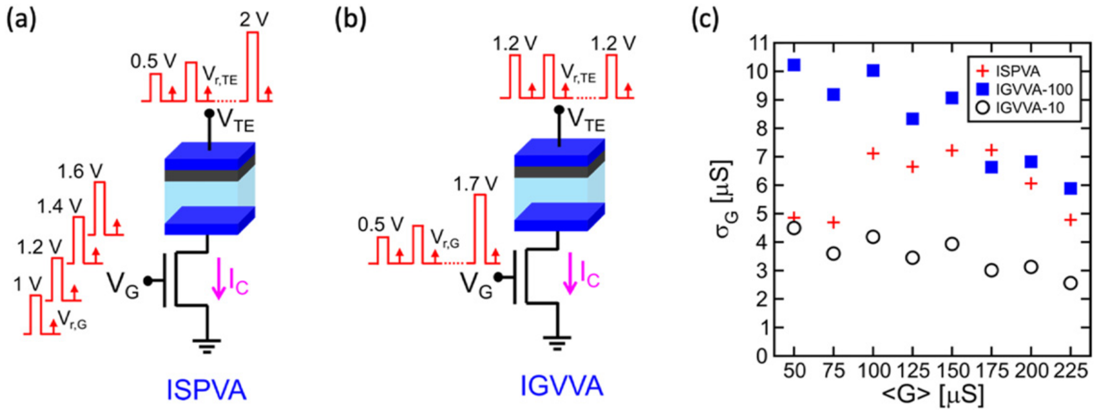

Figure 4. Sketches of (a) ISPVA and (b) IGVVA. (c) reports the standard deviation of the conductance value variation as a function of the average conductance for different correction code algorithms. © 2021 IEEE. Reprinted, with permission, from [67].

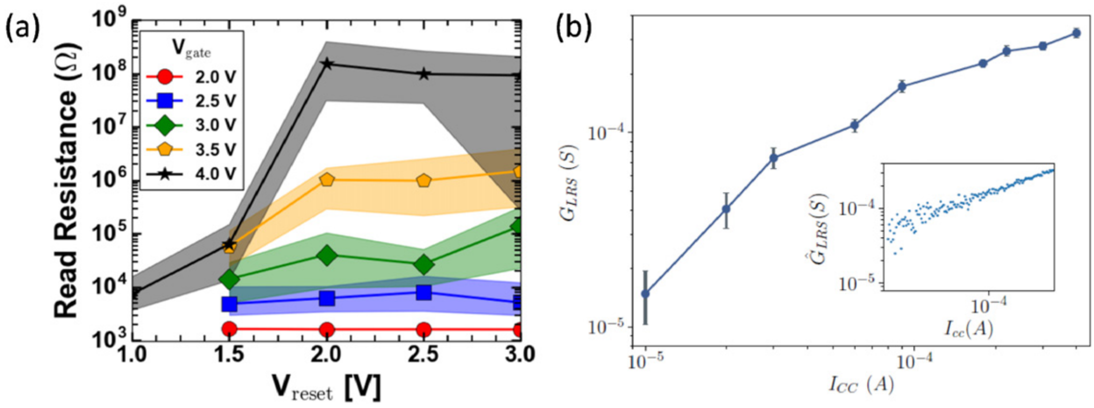

Figure 5. RRAM multilevel programming. (a) Multilevel programming through reset operations with different applied gate voltages in 1T1R structures. (b) Multilevel programming through set operations with different applied gate voltages through the use of a weight update protocol. (a) reproduced from [79] CC-BY 4.0. (b) © 2020 IEEE. Reprinted, with permission, from [80].

techniques in multiple arrays, with an overhead in terms of memory area and associated power consumption and latency [78].

The previous programming algorithms refer to set operation, which is conventionally controlled through a current compliance by a transistor. Recently, the use of the compliance current has been investigated also during the reset operation [79]. Usually the reset operation is performed by applying a transistor gate voltage that is larger than those applied during the set, so that the transistor do not act as a current limiter and the voltage drop across the transistor is minimized. On the contrary, if a gate voltage smaller than that used for set is applied for reset, the voltage divider allows programming intermediate states, even though with limited reliability. Dalgaty *et al* [79] report this programming strategy for TiN/Ti/HfO_2/TiN 1T1R devices as shown in figure 5(a) and apply it to Bayesian neural network as described in the next section.

The use of the transistor current limitation allows setting with some precision a absolute resistance value. To turn such programming scheme into a weight update, i.e. applying relative resistance changes, or steps, dedicated protocols must be elaborated. Payvand *et al* [80] propose a programming algorithm and circuits to exploit current compliance control to set HfO_2-1T1R devices in several resistance state (figure 5(b)). The working principle consists in a continuous reset of the device to the highest resistance state and a following set into the desired intermediate state. Furthermore, before applying the reset and set operation, the resistance of the device is read and the new current compliance value and transistor gate voltage are evaluated on the basis of the desired resistance change.

3.2. Analog weight update

The programming scheme described in this subsection is a rather unconventional programming investigated in the last years and it consists in the stimulation of the devices through train of identical *weak* pulses to

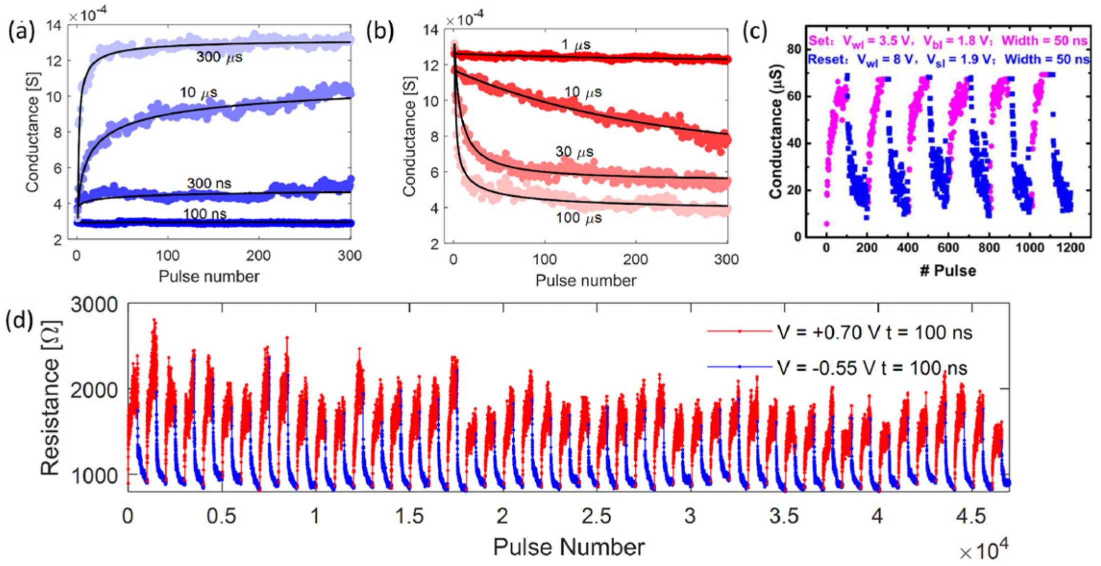

Figure 6. RRAM analog weight update. (a) and (b) 1R TiN/Ti/HfO_2/TiN RRAM cell. Analog programming stimulated by trains of identical set [(a) −0.7 V] and reset [(b), 0.9 V] pulses with various pulse widths. (c) Example of analog programming in the TiN/$HfAlO_x$/TaO_x/TiN 1T1R devices. Six set/reset weight update cycles are shown. (d) Example of endurance test for analog weight updates achieved in the same device of (a) and (b). (a) and (b) reproduced from [81] CC-BY 4.0. (c) reproduced from [11] CC-BY 4.0.

get an analog weight update, sometimes called *gradual* programming. Figures 6(a), (b) and (d) report the analog weight update in the TiN/Ti/HfO_2TiN device stimulated by trains of negative (set) and positive (reset) pulses, respectively. In the figures 6(a) and (b), we can observe that very short pulse widths at equal voltage produce no conductance change. On the contrary, few long pulses are sufficient to drive a high conductance change.

In general, conductance change of a RRAM device over time is fast or slow on the base of the strength of the programming conditions. Strong (weak) programming conditions are achieved by a combination of relatively high (low) voltages and/or long (short) pulses. A train of weak pulses, each of which induces slow switching change, is able to produce a gradual conductance transition useful for analog weight update operations. Despite a clear understanding of which are the factors allowing the implementation of an analog weight update is still lacking, several works evidenced that interlayers at the metal/oxide interfaces and interface switching enable such analog weight update [81, 82]. In other works, the doping of HfO_2 has been engineered to obtain analog conductance updates [50]. Materials engineering has been oriented on two aspects of the conductance update. First of all, a linear evolution of conductance as a function of the number of pulses is highly desirable to have a proper implementation of the training algorithms based on the back-propagation of the error. For instance Woo *et al* [53] used a AlO_x/HfO_2 bilayer to demonstrate a linear conductance dynamics. Obviously, the number of states that can be programmed is of primary importance for neuromorphic computing. However, it is not straightforward to define a value for the number of levels or states especially when the conductance evolution as a function of the number of pulses is not linear (see figures 6(a) and (b)), i.e. in case possible resistance state are not evenly spaced. For this reason, a precise comparison among literature results is not straightforward. However, best literature results may corresponds to few hundreds of effective resistance states, most of them reported for HfO_2-based devices [54, 83]. A mathematical definition of the effective number of states given a certain conductance dynamics has been proposed in [84], which can be useful for a quantitative assessment and device engineering.

Another aspect that requires a further device improvement for analog conductance update regards the memory window, which is usually quite limited for HfO_2 [28, 81]. As a matter of fact, resistance windows of an order of magnitude have been only reported for either set [82] or reset operation [85], while the reverse operation occurs always abruptly. Another evidenced issue of analog weight update is what has been named switching noise [86] or stimulated telegraph noise [87]. Such noise is ubiquitous in all filamentary devices and it was studied with reference to HfO_2-based devices [84]. Such noise is particularly relevant for reset and at high resistance values and results from a dynamical equilibrium between the processes of drift and diffusion of oxygen vacancies [84]. Anyway, despite such analog weight update is unconventional with respect to the standard memory programming, interestingly, it has already been demonstrated in TiN/Ti/HfO_2/TiN devices wire-connected to 350 nm CMOS technology node neurons [69] and in TiN/TaO_x/$HfAl_yO_x$/TiN 1T1R 1 kbit array [11]. An example of repeated analog set and reset updates for the latter devices is reported in

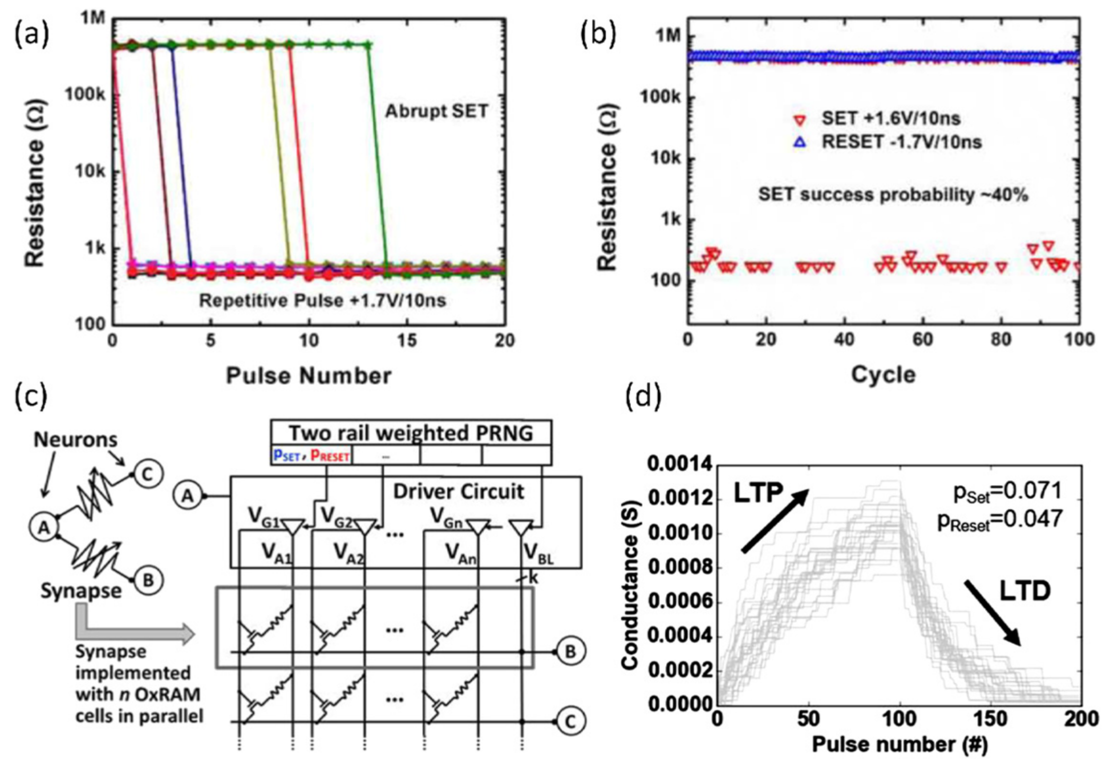

Figure 7. Stochastic RRAM programming. (a) Resistance as a function of the pulse number for stochastic programming. Different runs (different colors) show the switching from high to low resistance after different number of pulses. (b) Endurance test for weak programming conditions for set operation. (c) Circuital scheme of two synapses (B and C) constituted by several 1T1R blocks in a line. (d) Gradual set (or long term potentiation, LTP) and reset (or long term depression, LTD) operations obtained from compound synapses each of them composed by 20 1T1R RRAM devices in parallel. Each of the 20 light gray lines represents a measurement of one compound synapse. (a) and (b) reproduced from [85] CC-BY 3.0. (c) and (d) reproduced from [88] CC-BY 4.0.

figure 6(c). As a matter of fact, the use of weak programming pulse is expected to have a beneficial effect on device endurance. As an example figure 6(d) reports repeated analog set/reset update cycles up to some tens of thousand pulses in TiN/Ti/HfO_2/TiN devices.

3.3. Stochastic programming

In some devices or programming conditions, the switching events are so fast that weak programming conditions do not result in any gradual resistance modulation. Conversely, the switching can be considered stochastic in the sense that only two distinct levels are obtained and weak programming pulses produce the switching from one to the other and vice versa with some probability [28]. Stochastic programming has been employed with success in neural networks as an alternative to multilevel or analog weight storage [9, 85, 89]. As stated above, stochastic programming needs the use of weak programming pulses as in the case of gradual programming. Yu *et al* [85] report a stochastic set operation for devices composed of a HfO_2/TiO_2 multilayer. Figure 7(a) reports various set experiments (different colors) in which the switching occurs after different number of weak programming pulses. Figure 7(b) reports an endurance experiment using weak programming conditions for the set operation. Garbin *et al* [9] use the stochastic switching to obtain a gradual resistance change of the parallel of many HfO_2-based 1T1R devices, as shown in the circuital scheme of figure 7(c). The resulting set and reset dynamics as a function of the number of pulses is reported in figure 7(d).

3.4. Programming and uses of volatile RRAMs

Volatile RRAM devices have been mainly explored as selectors for memory crossbars [60, 61], as well as for hardware security [90]. Such applications will not be dealt with in the present review which will only concentrate the use of volatile devices enabling actual neuromorphic functionalities. In particular, the filament self-dissolution after a programming event from HRS to LRS can extend from μs to seconds. The relaxation is easily tracked by measuring the resistance of the devices by low applied voltages and opens the possibility to emulate short term dynamical elements in neuromorphic chips. The longest reported relaxation times in literature for Ag or Cu-based volatile RRAM are in the range of tens of ms up to seconds [18, 55, 56]. Figures 8(a)–(c) show some examples of relaxation dynamic in HfO_2- or SiOx-based RRAMs and how the relaxation time (t_R)

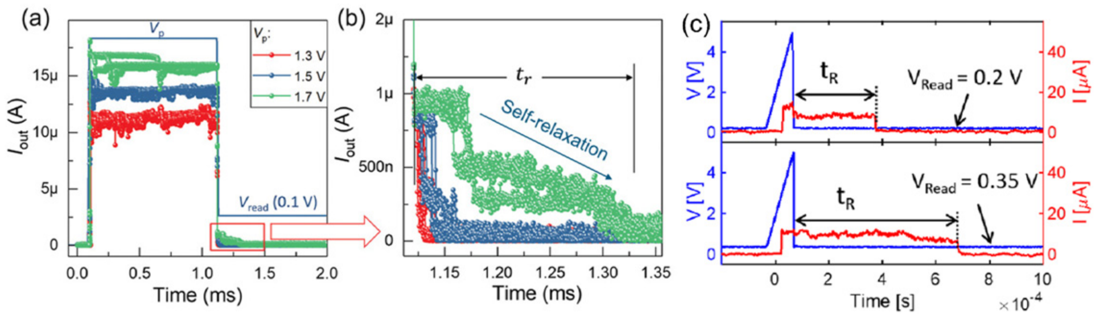

Figure 8. Relaxation dynamics of volatile RRAMs. (a) Current response of the Ag/HfO$_2$/Pt 1R RRAM device under applied pulses at various voltage amplitudes V_p and (b) corresponding relaxation behavior by reading the current at 0.1 V. t_R is the relaxation time. (c) Examples of the possibility to control the relaxation time (t_R) with reading voltage V_read, achieved in the 1T1R Ag/SiO$_x$/C device. (a) and (b) reproduced from [55] CC-BY 4.0. (c) reproduced from [56] CC-BY 4.0.

can be controlled by pulse voltage amplitude (figure 8(b)) and reading voltage (figure 8(c)). In figure 8(a), programming pulses with different amplitudes, V_p, are applied to a Pt/HfO$_2$/Ag device. The device switches to a LRS whose resistance is related to V_p. In general the final current value is related to the programming voltage amplitude (figure 8(a)) or time [55, 56]. After the pulse end, the low current states of the device is read at 0.1 V to monitor the self-relaxation process. The device current is progressively restored to the initial value during a relaxation time which is longer for higher initial current values (larger applied V_p) (figure 8(b)). Conversely, figure 8(c) shows that the relaxation time, t_R, is also controlled by the reading voltage [56]. The latter results are achieved in the Ag/SiO$_x$/C RRAM stack. As a matter of fact, Chekol *et al* [91] and Covi *et al* [56] showed that the relaxation time can be controlled to some extent by changing the programming conditions, like pulse width/amplitude and current compliance in 1T1R structures, respectively.

The existence of a relaxation dynamics enables the possibility of reaching a *cumulative* or *integrative* effect to repeated pulses. In particular, in case pulses are repeated with a period shorter than the typical relaxation time, the effect of each pulse is summed up to that of the previous pulses, thus leading to a LRS value lower that what achieved by single pulse [57, 62, 92], possibly leading to an extension of the relaxation time with continuous pulse stimulation. This effect is exploited to emulate the so called *pulse paired facilitation* observed in biological synapses [57, 62, 92] and the *synaptic metaplasticity* [93].

4. Computing schemes

Thanks to *in situ* data processing where data movement is virtually suppressed, IMC can accelerate a broad range of computing processes in both the digital and analog domains. These may include ANNs for deep learning [94] and SNNs which aim at mimicking the human brain particularly regarding its ability of learning and adaptation [12]. Both fully-connected networks [11, 77, 95] and convolutional neural networks (CNNs) [9, 54, 96] have been implemented with HfO$_2$-based RRAMs. Synaptic weights are generally quantized to a certain number of levels [26, 77] including the extreme case of binary weights (e.g. LRS and HRS) in binarized [97, 98] and ternary neural networks (TNNs) [99]. In addition to ANNs, other types of networks have been considered, e.g., decision trees and random forests implemented in ternary content addressable memory (TCAM) arrays. [100] Various type of SNNs have been implemented with HfO$_2$ RRAM devices, with the aim of supporting spike-based learning [101] and spatio-temporal recognition [63].

The following is a summary of the main demonstrations of IMC primitives with HfO$_2$ RRAM devices for machine learning and SNNs.

4.1. Acceleration of machine learning algorithms

A strong benefit of IMC derives from the one-step, parallel matrix vector multiplication (MVM) operation which provides the backbone of the fully-connected ANN in figure 9(a) [77]. Here, each neuron input contains the summation of the input signals multiplied by a synaptic weight W_{ij}, which is readily expressed by the MVM operation. Figure 9(b) shows the crosspoint array which can execute the computation of the MVM in one step: the application of a voltage V_i at the ith row of the array results in a current $I_j = \Sigma W_{ij} \cdot V_i$, which is equivalent to the MVM operation. In particular, the synaptic weight W_{ij} is implemented as the difference between two conductance values, namely

$$W_{ij} = G_{ij}^{+} - G_{ij}^{-}, \tag{1}$$

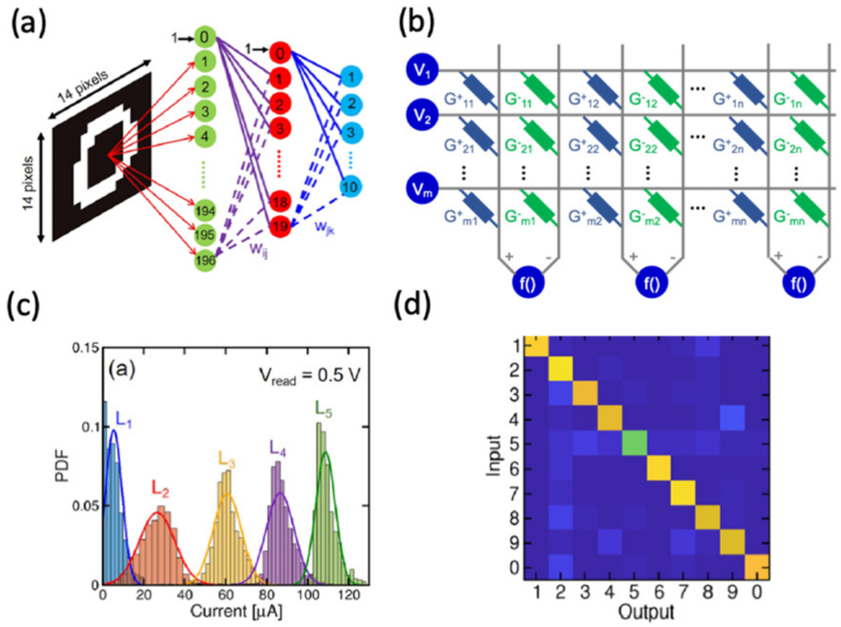

Figure 9. Implementation of neural network layers by HfO_2-based RRAM arrays. (a) Schematic illustration of a two-layer neural network for MNIST recognition. (b) Implementation of a single layer on a crosspoint array of synaptic memory devices, where each synaptic weight is represented by a differential pair of devices. (c) Distribution of measured read currents for HfO_2 RRAM devices programmed with multilevel states L_1, L_2, L_3, L_4 and L_5. (d) Confusion matrix for RRAM-based recognition of the MNIST dataset. (a), (c) and (d) reproduced from [77] CC-BY 4.0. (b) reproduced from [102] CC-BY 3.0.

where the minus sign can be obtained by subtraction of the currents in the two adjacent columns in figure 9(b) [102]. Figure 9(c) shows the distributions of individual device currents measured at $V_{\text{read}} = 0.5$ V in a 4 kb array of HfO_2-based RRAM [77]. The conductance G_{ij} of each device in the array was obtained by the ISPVA program-verify technique, as described in the previous section [77]. A total number of 5 levels were programmed in the array, including the HRS (L1) and four LRS levels (L2 to L5) with increasing conductance. These quantized conductance levels were used in equation (1) to describe the synaptic weights calculated from the back-propagation algorithm, which is a typical supervised off-line training technique [94]. Figure 9(d) shows the confusion matrix of the implemented hardware two-layer fully-connected ANN, namely the probability of a certain output response by the network as a function of the class of the input pattern. While on average the network gives a correct response, the accuracy is only around 83% compared to a software-level accuracy of 92%. This is due to two main limitations of the conductance matrix, namely (i) quantization of the weights with on only 5 levels, and (ii) stochastic variations of conductance with significant spread around the ideal value in figure 9(c).

To improve the accuracy of the ANN, more advanced program-verify techniques can be adopted, such as the IGVVA with small incremental gate voltage, e.g., 10 mV [26]. The improved programming precision allows to reduce the standard deviation of conductance, thus increasing the number of levels and reducing the quantization error. The improved precision of conductance, combined with quantization-aware algorithm for off-line training, allows for a substantial increase of recognition accuracy approaching the equivalent software performance [67]. Alternative error correction codes have been developed that take advantage of encoding/decoding strategies with the support of a hardware encoder device matrix, which impacts the area and energy requirements [103].

In addition to inference accelerators, *in situ* training was demonstrated in IMC hardware with 1T1R arrays of RRAM devices with Ta/HfO_2/Pt stack [95]. The on-line training was achieved by the stochastic gradient descent algorithm, where the synaptic update was executed directly on the device by adjusting the gate voltage, similar to the IGVVA approach [95]. Similar results were obtained for a TiN/TaO_x/$HfAl_yO_z$/TiN stack by applying a train of equal pulses with constant gate and top electrode voltage [11]. With a similar RRAM stack, a fully-memristive hardware implementation of CNN with 32 levels of conductance was demonstrated in [54]. Other CNN implementations with HfO_2-based RRAM were reported in [9, 96].

To reduce the complexity of precise RRAM programming for multiple-level operation, binarized neural networks (BNNs) [104] and TNNs [99] were developed with HfO_2-based RRAM. In a BNN, both neuron states and synaptic weights have binary values, such as +1 and −1, which strongly simplifies the computation and hardware implementation. In fact, in contrast to analog MVM implementations, all products and

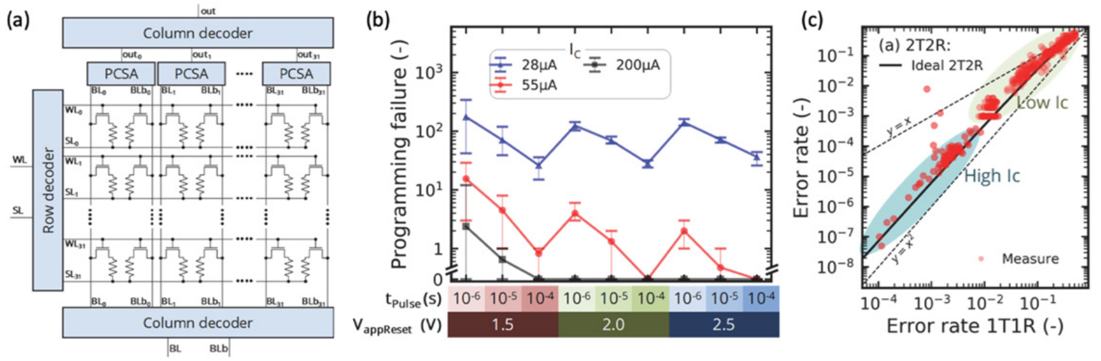

Figure 10. Sketch and performance of a BNN based on HfO_2 RRAM. (a) Schematic illustration of the 2T2R array, where the differential state of the synaptic element is read by a PCSA circuit. (b) Programming failure as a function of the reset pulse width and voltage amplitude and for various compliance currents I_c controlling the LRS. (c) Correlation plot of the 2T2R error rate as a function of the corresponding error rate of the 1T1R structure. The slope of 2 on the bilogarithmic plot indicates a better reliability with respect to programming variations of LRS and HRS. © 2018 IEEE. Reprinted, with permission, from [104].

summation are carried out in the digital domain, with binary product being implemented by a XNOR operation, while summation is replaced by the POPCOUNT operation which counts all output signals equal to one. A BNN was implemented in hardware with HfO_2-based RRAM arranged in the 2T2R structure shown in figure 10(a). Here, the two RRAM devices are programmed in a complementary state (HRS/LRS or LRS/HRS) and the synaptic weight is represented by the difference between the two RRAM currents, which is sensed by a precharge sense amplifier (PCSA) [104]. The 2T2R structure allows for a better immunity to errors resulting from tails of the distributions of the LRS and HRS conductance. As shown in figure 10(b), these errors can be minimized by increasing the compliance current, which reduces the LRS tails, and increasing the reset voltage and pulse width, which reduces the HRS tails. Figure 10(c) shows that write errors in the 2T2R structure increases quadratically with the single-bit error, as a result of the LRS and HRS occurring independently in the memory array [104].

The BNN concept was further demonstrated for learning, by taking advantage of the gradual potentiation and depression of RRAM devices, although with relaxed requirements about the symmetry and linearity of weight update with respect to on-line training with back-propagation algorithm [97]. The main advantages of the BNN are the resilience to conductance variation and the fully-digital approach within the computing hardware, where analog–digital converters are no more needed. However, the full parallelism of the analog domain MVM cannot be simply achieved within hardware BNN. Extension to TNN was also reported by using the same 2T2R synaptic architecture, which allows for a slight increase in recognition accuracy for the same network size [99]. Binary RRAMs were also adopted for TCAMs [105], which find extensive applications in storing synaptic tags for spike routing in multi-core SNNs [106] and decision trees for machine learning [107]. TCAMs are usually implemented by static random-access memories (SRAMs), however require relatively large silicon area due to the six-transistor structure of SRAMs. TCAMs with nonvolatile RRAMs can be reduced to a smaller 2T2R structure where the three states can be obtained by the configurations LRS/HRS (state 1), HRS/LRS (state 0) and HRS/HRS (state X, or 'do not care') [105].

A key limitation of RRAM devices for hardware ANN implementation is the limited precision due to the program/read variations [68, 70]. Such variations can be turned into a precious feature in stochastic computing circuits, e.g., true random number generators [109, 110], Bayesian neural networks [108] and Monte Carlo Markov chains [111]. In a Bayesian neural network (figure 11(a)), synaptic parameters usually consist of random variables, which well match the random nature of RRAM conductance obtained without program-verify algorithms [108]. Figure 11(a) shows the methodology for describing a given probability distribution of weights with stochastic RRAMs: the distribution can be approximated by a combination of a number of Gaussian distributions, each obtained by programming RRAM devices with a fixed pulsed amplitude and time, without verify. To represent a single probability distribution, a relatively large number of devices (e.g., 1024) would be required, as opposed to the single RRAM device (or two RRAM devices in the case of a differential synapse) required for describing a fixed, non-probabilistic weight in an ANN. The output neuron distribution can be obtained in figure 11(c) by sampling multiple output results from predefined sub-sets of the RRAM synapses [108]. Similarly, probabilistic Monte Carlo Markov chains were demonstrated by harnessing the stochastic distributions of programmed RRAM conductance, thus taking advantage of the cycle-to-cycle and device-to-device variations of RRAM [111].

The reported examples show that, depending on the RRAM multilevel precision, various computing schemes can be implemented which target different applications. The requirement on RRAM precision and

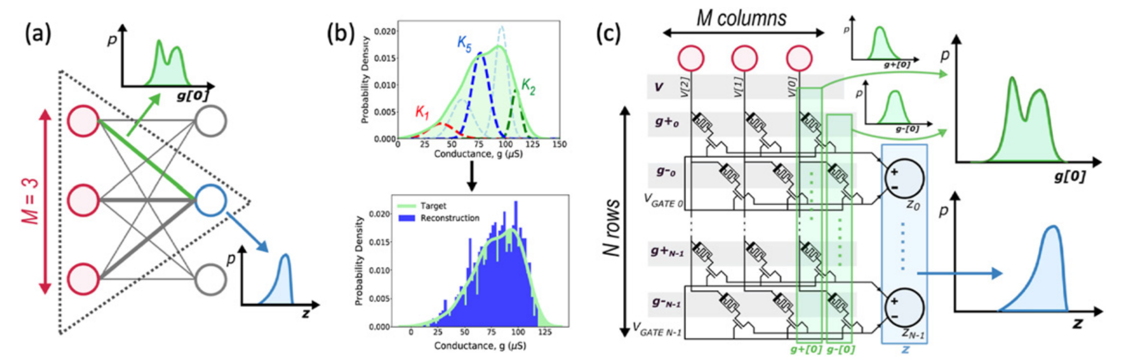

Figure 11. Implementation of Bayesian inference in a stochastic HfO$_2$ RRAM array. (a) Illustration of a 3 × 3 Bayesian neural network where each synapse and neuron are described by probability distributions. (b) Implementation of a target probability distribution by composition of individual normal distributions of HfO$_2$ RRAM conductance. (c) Circuit realization of the 3 × 1 network in the dashed triangle in (a), where the random conductance distribution of each column of the $N \times M$ RRAM array corresponds to the distribution of a synaptic parameter. As a result of the applied input voltage, the row currents are fed to the neuron circuit for subtraction and evaluation of the activation function. Reproduced from [108] CC-BY 4.0.

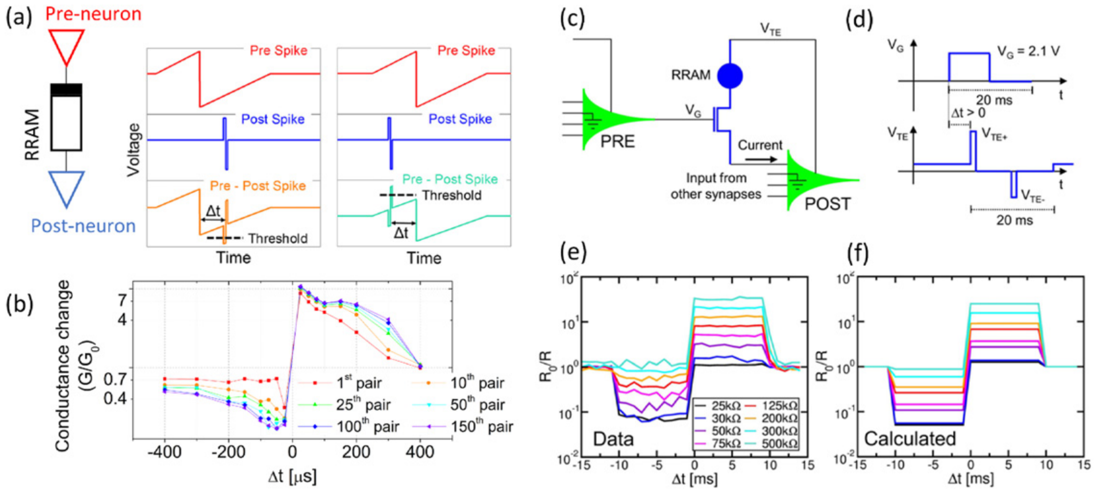

Figure 12. STDP synapses with HfO$_2$ RRAM devices. (a) Sketch of the 1R synapse for STDP based on the overlap between pre-neuron and post-neuron spikes. When Δt is negative, the resulting pulse shows a negative peak exceeding the threshold for reset transition, thus resulting in a depression update. When Δt is positive, the resulting pulse shows a positive peak exceeding the threshold for set transition, thus resulting in a potentiation update. (b) Resulting conductance change as a function of time delay Δt. (c) Sketch of the 1T1R synapse for STDP. (d) Illustration of the pre-spike (top) and post-spike (bottom) for positive Δt. (e) Resulting ratio of initial to final resistance as a function of Δt for increasing initial resistance. (f) Simulation results, showing predictable STDP based on time delay and initial synaptic weight. (a) and (b) reproduced from [10] CC-BY 4.0. (c)–(f) reproduced from [113] CC-BY 4.0.

consequent computation accuracy can be relaxed in favor of the reduction of system complexity and cost or in favor of ultimate low power operation in case the computing system is used, for instance, as a tool for the pre-processing or filtering of sensor data in the so called intelligent edge computing concept [112].

4.2. Spiking computing schemes

While ANNs show excellent performance in terms of image, speech and object classification, they also have several key weaknesses such as the catastrophic forgetting and the limited learning capability. To overcome these limits, SNNs directly mimic the information processing in the brain to gain a better performance in terms of learning, adaptation, real-time interaction with the environment and energy efficiency [12]. Both nonvolatile and volatile HfO$_2$-based RRAMs have been explored to study and demonstrate SNN concepts.

Nonvolatile devices are mostly used, as in the case of ANNs, to store static synaptic weights as conductance values. In addition, SNN training is generally driven by local learning protocols rather then the minimization of global error functions like in the case of ANNs. This fact renders the implementation of online training protocols easier for SNNs than for ANNs. Indeed, online learning is often investigated in SNNs. A large amount of publications have been dealing with the implementation of the so-called spike-timing dependent plasticity

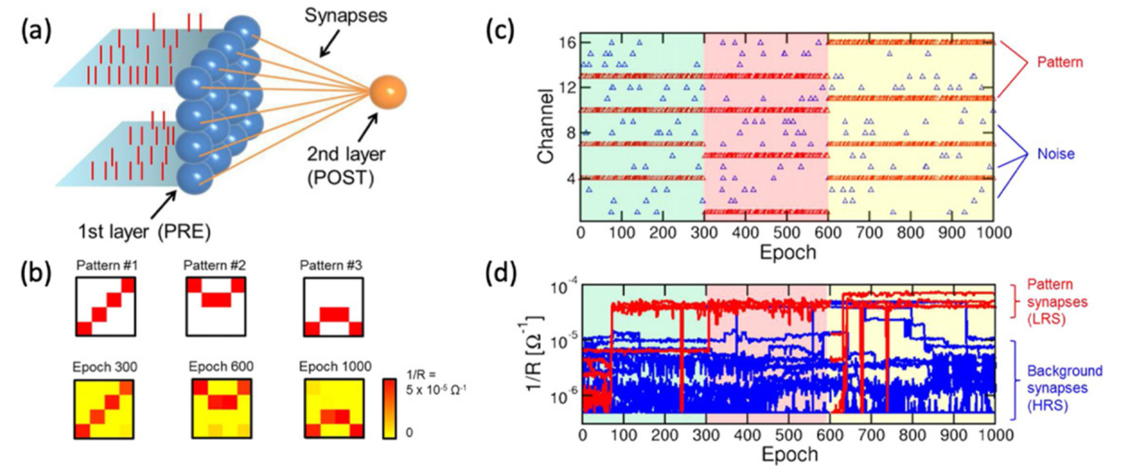

Figure 13. Unsupervised learning by STDP in a SNN. (a) Sketch of the SNN with 16 input neurons, 16 synapses and one output neuron. (b) input pattern presented to the RRAM-based synapses during phase 1, 2 and 3 (top), and measured configuration of the synaptic weights at the end of each phase (c). Input spikes as a function of time, including noise spikes to enable random STDP depression to forget previous patterns. (d) Measured conductance for each synapse in the network. Pattern synapses spontaneously converge to the LRS while background synapses converge to HRS due to the STDP weight update. Reproduced from [101] CC-BY 4.0.

(STDP) learning protocol in a synaptic device that mediates the communication between a pre-synaptic and a post-synaptic neuron. In biological STDP, the delay time, Δt, between the pre-synaptic and post-synaptic spikes dictates the synaptic weight update [114]. In particular, long-term potentiation takes place for pre-synaptic spike preceding the post-synaptic spike, while long-term depression takes place for the opposite spike sequence.

Analog STDP weight modulation qualitatively similar to the biological one has been reproduced by stimulating 1R devices at their two terminals by properly shaped overlapping pulses as shown in figure 12(a). Given the triangular shaped pre-spikes, the resulting voltage drop on the device depends on the relative timing of pre- and post-spikes and the obtained conductance change results in the typical asymmetric STDP shape reported in figure 12(b). The reported results refer to TiN/HfO_2/Ti/TiN devices properly optimized to give analog operation in response to train of pulses with increasing amplitude, as well as, in response to train of identical pulses [10]. Alternative shapes or even binary versions of STDP curves can be obtained by designing the shape of the programming pulses driving HfO_2-based devices, as attested by several publications [52, 85, 88, 115].

A temporal overlap scheme has been also proposed for 1T1R synapse structure [113, 116]. Figure 12(c) illustrates the concept of 1T1R synapse for STDP [113]. The pre-synaptic spike is applied to the gate of the select transistor, while the post-synaptic spike, also called feedback spike, is applied to the top electrode of the RRAM device. Under pre-synaptic stimulation, the synaptic current, which is proportional to the RRAM conductance, is injected from the transistor source to the post-synaptic neuron. Under post-synaptic stimulation, a set transition takes place across the RRAM device in case of a small positive delay Δt, where the pre-synaptic spike overlaps with the positive pulse of the post-synaptic spike (figure 12(d)). On the other hand, a reset transition takes place in the case of negative delay, where the pre-synaptic spike overlaps with the negative side of the post-synaptic spike. Figures 12(e) and (f) show the measured and calculated STDP characteristics, respectively, supporting the effects of potentiation and depression for positive and negative delay, respectively [113]. Improved STDP characteristics with exponentially decaying weight update as a function of increasing positive/negative delay can be obtained by properly shaping the post/pre-synaptic spikes and introducing a second select transistor in the two-transistor/one-resistor (2T1R) structure [117].

In all these works, the STDP protocol is implemented by including the temporal information in long lasting pulses, which could complicate the management of a high number of devices in an array. VLSI neuromorphic chips are able to implement various STDP variants by encoding the temporal information in the discharge of capacitors in synapse and neuron CMOS circuits [118, 119]. On one side, this solution avoids the use of long overlapping pulses [119], on the other side, the feasible time constants are limited by the physical dimension of the capacitors [120]. The programming of RRAMs according to a generalized and biologically plausible version of STDP, akin to the Bienenstok–Cooper–Munro theory [121], was demonstrated by connecting a TiN/Ti/HfO_2/TiN device in between two CMOS neurons realized in 350 nm technology node. Further, the proposal of a six-transistor/one-resistor (6T1R) HfO_2-based synapse in connection with the same CMOS neuron circuits was validated through system level simulations against the hand-written classification task [83]. Supervised training protocols can also be implemented in SNNs. For instance, the so called *delta* rule which

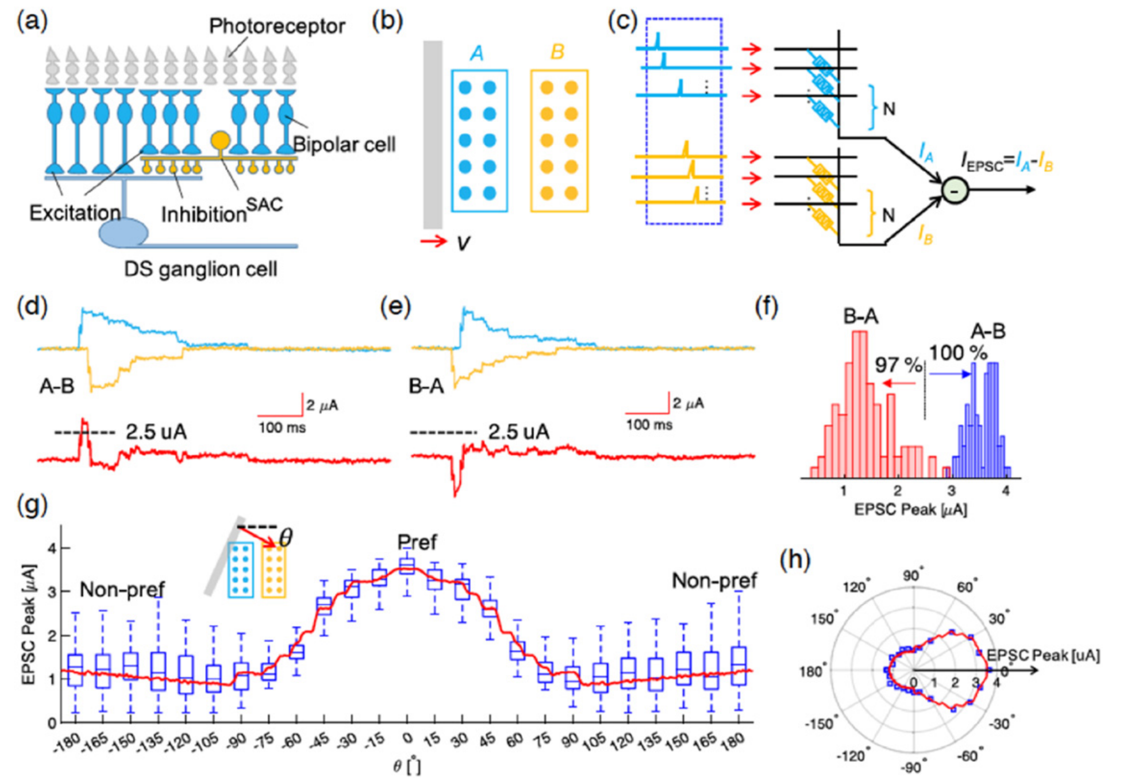

Figure 14. Direction sensitivity by spatio-temporal pattern recognition in volatile HfO_2 RRAM. (a) Sketch of the receptive fields of DS ganglion cell. (b) Artificial receptive fields with excitatory (blue) and inhibitory (yellow) branches, each consisting of several HfO_2 RRAM for statistical averaging of retention times. (c) Circuit for spike stimulation of the receptive fields, where the EPSC is obtained as the difference between the excitatory and inhibitory response currents. (d) Current responses of excitatory and inhibitory branches (top), and resulting EPSC (bottom) for an image moving from left to right. (e) Same as (d) but for an image moving from right to left. (f) Distribution of EPSC peaks for preferred direction, equivalent to sequence A–B (b), and non-preferred direction, equivalent to sequence B–A (c). (g) EPSC peak as a function of the angle between non-preferred direction (180°) and preferred direction (0°). (h) Polar plot of the EPSC peak. Reproduced from [63] CC-BY 4.0.

is a spike-based version of the gradient descent has been validated by simulation based on experimental data obtained from Pt/HfO_2/TiO_x/Ti RRAM devices [122].

The investigation of memristive SNNs has gradually moved from system-level simulations to mixed hardware/software experiments, to fully hardware realizations. Fully-memristive neural networks implementing STDP synapses were implemented for the demonstration of unsupervised learning [101]. Figure 13(a) shows a sketch of the one-layer neural network that was used for the unsupervised learning of a 4 × 4 image pattern. The 1T1R synapses were stimulated by spiking signals, either containing the image pattern (figure 13(b)) or noisy, sparse images. The synapses were initialized with a random configuration, resulting in a stochastic spiking of the single output neuron, with a higher probability of fire under presentation of the input image pattern. As a result, the presentation of the image pattern preferentially causes the pre-post sequence, hence potentiation of the stimulated synapses, whereas random noise stochastically causes a post-pre sequence leading to depression. The stochastic STDP dynamics thus allows for sequential image learning, where each submitted image is learnt by the synaptic array, then replaced by the newly arrived pattern, as shown in figures 13(c) and (d) [101]. A full-hardware SNN with integrate-and-fire neurons and 1T1R RRAM-based synapses was integrated in the 130 nm CMOS node, demonstrating inference accuracy of 84% with binarized weights for a simplified MNIST dataset [123]. Similar SNNs were developed by adopting other types of RRAM materials, such as SiO_x [124].

A different usage of nonvolatile devices is the one proposed by Dalgaty *et al* [79]. They used 1T1R HfO_2-based RRAMs as programmable resistors in RC circuits for the tuning and diversification of synaptic and neuronal time constants. As a matter of fact, in fully CMOS analog spiking chips, neurons and synaptic dynamics are implemented with the charge and discharge of capacitors in order to match the signal timescales. Capacitors occupy large silicon footprint and are, therefore, shared among all neurons and synapses of a chip. Therefore, the solution proposed by Dalgaty *et al* [79] expands the tunability and diversification possibilities of purely CMOS neuromorphic chips through nanoscaled nonvolatile programmable RRAM resistors.

On the other side, time constants and synaptic/neuronal dynamics can in principle implemented by taking advantage of device physics [1, 125]. From this standpoint, volatile RRAMs-based on HfO_2 can mimic short term memory in the human brain, thus supporting various cognitive processes such as the spatio-temporal sequence recognition. Although nonvolatile RRAM based on HfO_2 were also shown to learn and recognize spatio-temporal patterns [126], volatile RRAMs can directly mimic transient effects, such as the excitatory post-synaptic current (EPSC), thus serving as an ideal hardware parallel for short-term memory effects. Figure 14(a) illustrates the concept for in-memory sensing and processing capable of direction sensitivity similar to the human retina [63]. The direction sensitive (DS) ganglion cell in figure 14(a) serves in this role by collecting the EPSCs from excitatory and inhibitory synapses stimulated by the photoreceptors in the retina. Due to their space configuration, excitatory and inhibitory synapses are stimulated at different times by a moving light image. For instance, an image moving from left to right in figure 14(b) stimulates excitatory synapse (A) followed by inhibitory synapses (B). This can be replicated in hardware by the neural network in figure 14(c), where excitatory and inhibitory synapses are mimicked by volatile RRAM devices, each contributing a transient current for a limited retention time t_{ret}. The difference between excitatory and inhibitory currents yields EPSC, which consists of a positive current peak for the preferred sequence A–B (image moving from left to right, figure 14(d) or negative peak for the non-preferred sequence B–A (image moving from right to left, figure 14(e). Figure 14(f) shows the probability distributions of the maximum measured EPSC, indicating well-separated distributions for sequences A–B and B–A, thus supporting direction sensitivity as shown in figures 14(g) and (h) [63].

5. Challenges, solutions and perspectives

As for most emerging memory technologies, the development of HfO_2-based RRAM is still facing several challenges, mostly originating from the non-ideal reliability and performance of the device. The most relevant limitations of HfO_2-based RRAM, which are shared with all other filamentary RRAMs, are the rather high variability, the random fluctuation of resistance, the relatively narrow resistance window and the limited endurance. These issues affect all IMC applications, particularly those where the device is supposed to operate in a multilevel mode to maximize the density and performance. Two types of variability effects are present, namely device-to-device and cycle-to-cycle variation of the programmed states [77]. Stochastic variations are inherently arising from the filamentary nature of the conduction path in the RRAM device, where the variation of defect number and of the filament shape can result in a relatively large change of resistance. Both variation phenomena can be addressed by accurate program/verify techniques to finely tune the resistance to the desired multilevel state [26]. However, post-programming rediffusion of defects and RTN can affect the resistance even after the program/verify operation [73]. These fluctuation phenomena result in the broadening of the distribution, thus affecting the bit precision of RRAM conductance levels.

In addition to variations and fluctuations, HfO_x RRAM shows an intrinsic limit of the resistance window, particularly for the HRS which generally shows a finite, non-zero value of resistance. The resulting leakage current can affect the IMC accuracy if not properly compensated. In general, a differential synapse, including two RRAM devices with opposite currents to represent the positive and negative components of the synaptic weight, is recommended to achieve a highly precise zero weight. This is shown in figure 15(a), illustrating a differential synapse where the two opposite currents are obtained by biasing the two RRAM devices with opposite voltage, so that the overall weight is given by $W = G^{+} - G^{-}$, where G^{+} and G^{-} are the conductance values of the two devices in the differential pair [127]. Given the distribution of programmed conductance in figure 15(b), obtained by the IGVVA10 algorithm, one can properly combine the LRS levels L_1–L_9 to realize the differential weight distribution in figure 15(c). Although the HRS distribution in figure 15(b) displays a non-negligible conductance of about 8 μS, the weight W_0 in figure 15(c) displays an average zero value thanks to the subtraction of two L_9 levels in the differential pairs.

Note that a significant drawback of the differential approach in figure 15 is the presence of relatively large currents to achieve relatively low weight values, e.g., two conductances of about 225 μS are needed to achieve a zero-valued weight with relatively low standard deviation. More generally, HfO_2-based RRAMs display rather large conductance, in the range of several tens of μS, as shown in figure 15(b). While relatively high conductance values are beneficial thanks to a relatively small variation and a higher robustness against high-T annealing [77], large currents can cause significant voltage drop across the bitlines (usually referred to as IR drop), due to the summation of synaptic currents in the MVM operation and to the parasitic wire resistance of the metallic interconnections. The IR drop can be mitigated by several techniques including both hardware training techniques [128] and replication circuits for compensation [129]. From this standpoint, the adoption of low-current emerging memory technologies, such as electro-chemical random access memories (ECRAMs) [130, 131], is the most efficient solution.

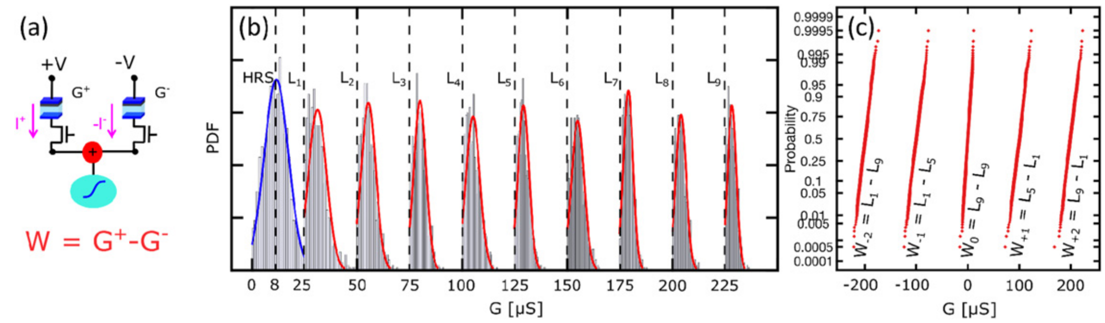

Figure 15. (a) Circuital scheme of a synapse composed of two RRAM devices; (b) conductance distribution of multilevel states of one RRAM device; (c) cumulative distributions of 5 weight levels of the overall synaptic unit. © 2022 IEEE. Reprinted, with permission, from [127].

Neuromorphic computing may require extensive programming of devices, e.g., for continuous learning and adaptation of synaptic weights or integration-and-fire neuron applications. From this point of view, a potential concern is the limited endurance upon repeated set/reset cycles [132]. The endurance of HfO_2 has been shown to be typically in the range of one million cycles, although the maximum number of cycles exponentially drops for increasing reset voltage [133]. For incremental set/reset operations, which might be needed for neuromorphic plasticity and integration, a larger endurance might be expected, although a comprehensive study of endurance for neuromorphic applications is not available yet.

Regarding the programming operation, another key concern is the forming operation, which is needed to initialize the device from the initial, pristine state of high resistance. Forming generally requires a relatively high voltage, which is a burden from the circuit point of view as it may requires charge pumps or high-reliability select transistor that can sustain the applied voltage. To minimize this burden, the HfO_2 layer is engineered to minimize the forming voltage by introducing a suitable concentration of defects [42]. This can be achieved by an oxygen exchange layer generated by redox exchange at the interface between the switching HfO_2 layer and a scavenging layer of moderately reactive metal, such as Ti, Hf or Ta [19], as also discussed in section 2.1.

Hardware accelerators of network training require in-memory execution of the outer product, namely an element-wise vector–vector product for updating the weight matrix in the RRAM array [134]. For this type of application, the device should display a high linearity of conductance vs number of pulses at fixed voltage, to enable potentiation or depression which is proportional to the pulse width at a given amplitude [102]. From this standpoint, the HfO_2-based RRAM is not optimized to achieve high linearity of conductance update, due to saturation effects (figure 5). RRAM with bilayer oxide stacks, such as HfO_2/TaO_2, have been recently reported to improve the linearity of conductance update [135]. Alternative memory technologies have shown a better linearity, usually combined with a lower operating current [136, 137].

Table 1 shows a summary of the properties of RRAM compared to other nonvolatile memory technologies [138], including commercial flash [139] and emerging memories such as phase change memory (PCM) [140], spin-transfer torque magnetic random access memory (STT-MRAM) [141], spin–orbit torque magnetic random access memory (SOT-MRAM) [142], ferroelectric random access memory (FeRAM) [143], ferroelectric field-effect transistor (FeFET) [144] and Li-ion based ECRAM [136]. In general, RRAM displays good compatibility for integration in CMOS circuits, simple fabrication process in the back end and small cell size. However, challenges still exist in terms of current operation, programming speed and reliability, including variability, endurance and fluctuations. Further progress is possible by a suitable combination of material engineering, programming/read/training algorithm and circuit/architecture design.

For what concerns the employment of volatile device for the implementation of dynamical components of neuromorphic networks, the investigation is still at the beginning and problems and challenges have not been identified precisely yet. However, one can consider that the use of the relaxation dynamics of volatile devices in real system is beneficial only in case the decay times are longer compared to the time constants that can be alternatively achieved with reasonably small capacitors and charging/discharging currents. In CMOS neuromorphic chips time constants of hundreds of milliseconds have been reported thanks to the use of very low subthreshold transistor currents and relatively large capacitors used to emulate the dynamics of a row of many synaptic devices exploiting the superposition principle [118]. Relaxation times of seconds are closed to the maximum values reported for HfO_2-based devices. Therefore, it is not clear whether the replacement of capacitors with volatile device is a real advancement. In turn, in case of the emulation of the dynamics of individual network units, like single neuron or synapse with its own dynamics, the superimposition principle does not hold any more and use of volatile RRAMs for each network units provides a scaling advantage compared to

Table 1. Comparison of figures of merit of different technologies useful for the nonvolatile storage of synaptic weights. Adjusted from [102] under the terms of Creative Commons Attribution 4.0 License, Copyright 2019, Institute of Physics.

Technology	CMOS mainstream memories		Memristive emerging memories						
	NOR flash	NAND flash	RRAM	PCM	STT-MRAM	FeRAM	FeFET	SOT-MRAM	Li-ion
On/off ratio	10^4	10^4	10 to 10^2	10^2 to 10^4	1.5–2	10^2 to 10^3	5–50	1.5–2	40 to 10^3
Multilevel operation	2 bit	4 bit	2 bit	2 bit	1 bit	1 bit	5 bit	1 bit	10 bit
Write voltage	$<$10 V	$>$10 V	$<$3 V	$<$3 V	$<$1.5 V	$<$3 V	$<$5 V	$<$1.5 V	$<$1 V
Write time	1–10 μs	0.1–1 ms	$<$10 ns	$\sim$50 ns	$<$10 ns	$\sim$30 ns	$\sim$10 ns	$<$10 ns	$<$10 ns
Read time	$\sim$50 ns	$\sim$10 μs	$<$10 ns	$<$10 ns	$<$10 ns	$<$10 ns	$\sim$10 ns	$<$10 ns	$<$10 ns
Stand-by power	Low	Low	Low	Low	Low	Low	Low	Low	Low
Write energy (J bit^{-1})	$\sim$100 pJ	$\sim$10 fJ	0.1–1 pJ	10 pJ	$\sim$100 fJ	$\sim$100 fJ	$<$1 fJ	$<$100 fJ	$\sim$100 fJ
Linearity	Low	Low	Low	Low	None	None	Low	None	High
Drift	No	No	Weak	Yes	No	No	No	No	No
Integration density	High	Very high	High	High	High	Low	High	High	Low
Retention	Long	Long	Medium	Long	Medium	Long	Long	Medium	—
Endurance	10^5	10^4	10^5 to 10^8	10^6 to 10^9	10^{15}	10^{10}	$>10^5$	$>10^{15}$	$>10^5$
Suitability for DNN training	No	No	No	No	No	No	Moderate	No	Yes
Suitability for DNN inference	Yes	Yes	Moderate	Yes	No	No	Yes	No	Yes
Suitability for SNN applications	Yes	No	Yes	Yes	Moderate	Yes	Yes	Moderate	Moderate

the use of capacitors. In particular, volatile devices can be used to implement biological properties of individual synapses, like paired pulse facilitation and depression [57, 62, 145], metaplasticity [146] or short-to-long term memory [145] transition in case the same RRAM device shows both volatile to nonvolatile retention [92, 93, 147]. These functions, which cannot be efficiently implemented in conventional CMOS technology, have not been demonstrated neither with large statistics nor at array level, yet. In general, however, a criticality that can already be identified for volatile device is their variability whose impact may depend a lot on their specific use in a neuromorphic system. In some case the parallel of many volatile device instead of only one is used in order to mitigate the effect of variability [63].

6. Conclusions

This review article presents the status of HfO_2-based RRAM devices for neuromorphic computing. The key device properties are illustrated for RRAMs devices, highlighting the role of the top electrode material in controlling the forming voltage and the volatile/nonvolatile memory behavior. The programming algorithms for nonvolatile RRAM are described, covering both high-precision multilevel cell programming for off-line training and analog weight update for on-line training. An overview on computing schemes is provided, covering ANNs with binary, multilevel and stochastic weights, as well as SNNs for unsupervised learning and spatiotemporal recognition. Finally, we discuss about open challenges, solutions and perspective of HfO_2-based RRAMs for neuromorphic applications in comparison with other existing and emerging technologies. From this report, HfO_2 appears as one of the most important RRAM material for demonstrating and prototyping neuromorphic circuits.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

All authors contributed equally to the writing of the manuscript. This work was partially funded by the European Union's Horizon 2020 Research and Innovation Program under the Grant No. 899559 and under the Grant No. 871371 (MeM-Scales project).

Data availability statement

The experimental data of figure 6(d) are available from the authors upon reasonable request. All the other data are from figures reprinted from published works as indicated in the caption.

ORCID iDs

S Brivio https://orcid.org/0000-0003-2386-7953
S Spiga https://orcid.org/0000-0001-7293-7503
D Ielmini https://orcid.org/0000-0002-1853-1614

References

[1] Ielmini D, Wang Z and Liu Y 2021 Brain-inspired computing via memory device physics *APL Mater.* **9** 050702

[2] Spiga S, Sebastian A, Querlioz D and Rajendran B (ed) 2020 *Memristive Devices for Brain-Inspired Computing Memristive Devices for Brain-Inspired Computing* (Duxford: Woodhead Publishing)

[3] Christensen D V *et al* 2022 2022 roadmap on neuromorphic computing and engineering *Neuromorphic Comput. Eng.* **2** 022501

[4] Wong H-S, Lee H-Y, Yu S, Chen Y-S, Wu Y, Chen P-S, Lee B, Chen F T and Tsai M-J 2012 Metal-oxide RRAM *Proc. IEEE* **100** 1951–70

[5] Yang J, Strukov D B and Stewart D R 2013 Memristive devices for computing *Nat. Nanotechnol.* **8** 13–24

[6] Ielmini D 2016 Resistive switching memories based on metal oxides: mechanisms, reliability and scaling *Semicond. Sci. Technol.* **31** 063002

[7] Chen H-Y *et al* 2017 Resistive random access memory (RRAM) technology: from material, device, selector, 3D integration to bottom-up fabrication *J. Electroceram.* **39** 21–38

[8] Yu S, Wu Y, Jeyasingh R, Kuzum D and Wong H S P 2011 An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation *IEEE Trans. Electron Devices* **58** 2729–37

[9] Garbin D, Vianello E, Bichler O, Rafhay Q, Gamrat C, Ghibaudo G, DeSalvo B and Perniola L 2015 HfO_2-based OxRAM devices as synapses for convolutional neural networks *IEEE Trans. Electron Devices* **62** 2494–501

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

[11] Yao P *et al* 2017 Face classification using electronic synapses *Nat. Commun.* **8** 15199

[12] Chicca E, Stefanini F, Bartolozzi C and Indiveri G 2014 Neuromorphic electronic circuits for building autonomous cognitive systems *Proc. IEEE* **102** 1367–88

[13] Govoreanu B *et al* 2011 10 × 10 nm^2 Hf/HfO_x crossbar resistive RAM with excellent performance, reliability and low-energy operation *Electron Devices Meeting (IEDM), 2011 IEEE Int.* p 31

[14] Seok J *et al* 2014 A review of three-dimensional resistive switching cross-bar array memories from the integration and materials property points of view *Adv. Funct. Mater.* **24** 5316–39

[15] Ielmini D and Wong H-S 2018 In-memory computing with resistive switching devices *Nat. Electron.* **1** 333–43

[16] Indiveri G and Liu S-C 2015 Memory and information processing in neuromorphic systems *Proc. IEEE* **103** 1379–97

[17] Merolla P A *et al* 2014 A million spiking-neuron integrated circuit with a scalable communication network and interface *Science* **345** 668–73

[18] Wang R, Yang J-Q, Mao J-Y, Wang Z-P, Wu S, Zhou M, Chen T, Zhou Y and Han S-T 2020 Recent advances of volatile memristors: devices, mechanisms, and applications *Adv. Intell. Syst.* **2** 2000055

[19] Lee H Y *et al* 2008 Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO_2 based RRAM *2008 IEEE Int. Electron Devices Meeting* pp 1–4

[20] Robertson J 2004 High dielectric constant oxides *Eur. Phys. J. Appl. Phys.* **28** 265–91

[21] Spiga S, Driussi F, Congedo G, Wiemer C, Lamperti A and Cianci C 2018 Sub-1 nm equivalent oxide thickness Al–HfO_2 trapping layer with excellent thermal stability and retention for nonvolatile memory *ACS Appl. Nano Mater.* **1** 4633–41

[22] Mikolajick T, Schroeder U and Slesazeck S 2020 The past, the present, and the future of ferroelectric memories *IEEE Trans. Electron Devices* **67** 1434–43

[23] Yu S, Hur J, Luo Y-C, Shim W, Choe G and Wang P 2021 Ferroelectric HfO_2-based synaptic devices: recent trends and prospects *Semicond. Sci. Technol.* **36** 104001

[24] Covi E, Mulaosmanovic H, Max B, Slesazeck S and Mikolajick T 2022 Ferroelectric-based synapses and neurons for neuromorphic computing *Neuromorph. Comput. Eng.* **2** 012002

[25] Ielmini D, Nardi F and Balatti S 2012 Evidence for voltage-driven set/reset processes in bipolar switching RRAM *IEEE Trans. Electron Devices* **59** 2049–56

[26] Milo V *et al* 2021 Accurate program/verify schemes of resistive switching memory (RRAM) for in-memory neural network circuits *IEEE Trans. Electron Devices* **68** 3832–7

[27] Sawa A 2008 Resistive switching in transition metal oxides *Mater. Today* **11** 28–36

[28] Brivio S and Menzel S 2020 Resistive switching memories *Memristive Devices for Brain-Inspired Computing Memristive Devices for Brain-Inspired Computing* ed S Spiga, A Sebastian, D Querlioz and B Rajendran (Duxford: Woodhead Publishing) pp 17–61

[29] Waser R, Dittmann R, Staikov G and Szot K 2009 Redox-based resistive switching memories—nanoionic mechanisms, prospects, and challenges *Adv. Mater.* **21** 2632–63

[30] Brivio S, Tallarida G, Cianci E and Spiga S 2014 Formation and disruption of conductive filaments in a HfO_2/TiN structure *Nanotechnology* **25** 385705

[31] Bersuker G *et al* 2011 Metal oxide resistive memory switching mechanism based on conductive filament properties *J. Appl. Phys.* **110** 124518

[32] Ielmini D 2011 Modeling the universal set/reset characteristics of bipolar RRAM by field- and temperature-driven filament growth *IEEE Trans. Electron Devices* **58** 4309–17

[33] Chen Y Y *et al* 2013 Endurance/retention trade-off on HfO_2/metal cap 1T1R bipolar RRAM *IEEE Trans. Electron Devices* **60** 1114–21

[34] Frascaroli J, Volpe F, Brivio S and Spiga S 2015 Effect of Al doping on the retention behavior of HfO_2 resistive switching memories *Microelectron. Eng.* **147** 104–7

[35] Azzaz M *et al* 2016 Improvement of performances HfO_2-based RRAM from elementary cell to 16 kb demonstrator by introduction of thin layer of Al_2O_3 *Solid-State Electron.* **125** 182–8 extended papers selected from ESSDERC 2015

[36] Pérez E, Javier Pérez-Avila A, Romero-Zaliz R, Mahadevaiah M K, Pérez-Bosch Quesada E, Bautista Roldan J, Jimenez-Molinos F and Wenger C 2021 Optimization of multi-level operation in RRAM arrays for in-memory computing *Electronics* **10** 1084

[37] Chou C-C, Lin Z-J, Tseng P-L, Li C-F, Chang C-Y, Chen W-C, Chih Y-D and Chang T-Y J 2018 An N40 256k × 44 embedded RRAM macro with SL-precharge SA and low-voltage current limiter to improve read and write performance *2018 IEEE Int. Solid-State Circuits Conference (ISSCC)* pp 478–80

[38] Grenouillet L *et al* 2021 16 kbit 1T1R OxRAM arrays embedded in 28 nm FDSOI technology demonstrating low BER, high endurance, and compatibility with core logic transistors *2021 IEEE Int. Memory Workshop (IMW)* pp 1–4

[39] Yu S, Chen H-Y, Gao B, Kang J and Wong H-S P 2013 HfO_x-based vertical resistive switching random access memory suitable for bit-cost-effective three-dimensional cross-point architecture *ACS Nano* **7** 2320–5

[40] Calka P *et al* 2014 Engineering of the chemical reactivity of the Ti/HfO_2 interface for RRAM: experiment and theory *ACS Appl. Mater. Interfaces* **6** 5056–60

[41] Clima S *et al* 2014 RRAMs based on anionic and cationic switching: a short overview *Phys. Status Solidi RRL* **8** 501–11

[42] Young-Fisher K G, Bersuker G, Butcher B, Padovani A, Larcher L, Veksler D and Gilmer D C 2013 Leakage current-forming voltage relation and oxygen gettering in HfO_x RRAM devices *IEEE Electron Device Lett.* **34** 750–2

[43] Padovani A, Larcher L, Padovani P, Cagli C and De Salvo B 2012 Understanding the role of the Ti metal electrode on the forming of HfO_2-based RRAMs *2012 4th IEEE Int. Memory Workshop* (Milan, Italy) (IEEE) pp 1–4

[44] Chen Y S *et al* 2009 Highly scalable hafnium oxide memory with improvements of resistive distribution and read disturb immunity *Electron Devices Meeting (IEDM), 2009 IEEE Int.* pp 1–4

[45] Magyari-Köpe B, Duncan D, Zhao L and Nishi Y 2016 Doping technology for RRAM—opportunities and challenges *2016 Int. Symp. on VLSI Technology, Systems and Application (VLSI-TSA)* pp 1–2

[46] Zhang H *et al* 2009 Effects of ionic doping on the behaviors of oxygen vacancies in HfO_2 and ZrO_2: a first principles study *2009 Int. Conf. on Simulation of Semiconductor Processes and Devices* pp 1–4

[47] Brivio S, Frascaroli J and Spiga S 2017 Role of Al doping in the filament disruption in HfO_2 resistance switches *Nanotechnology* **28** 395202

[48] Peng C-S, Chang W-Y, Lee Y-H, Lin M-H, Chen F and Tsai M-J 2012 Improvement of resistive switching stability of HfO_2 films with Al doping by atomic layer deposition *Electrochem. Solid-State Lett.* **15** H88–90

[49] Roy S *et al* 2020 Toward a reliable synaptic simulation using Al-doped HfO_2 RRAM *ACS Appl. Mater. Interfaces* **12** 10648–56

[50] Gao B, Liu L and Kang J 2015 Investigation of the synaptic device based on the resistive switching behavior in hafnium oxide *Prog. Nat. Sci.: Mater. Int.* **25** 47–50

[51] Chandrasekaran S, Simanjuntak F M, Saminathan R, Panda D and Tseng T-Y 2019 Improving linearity by introducing Al in HfO_2 as a memristor synapse device *Nanotechnology* **30** 445205

[52] Yu S, Gao B, Fang Z, Yu H, Kang J and Wong H-S 2013 A low energy oxide-based electronic synaptic device for neuromorphic visual systems with tolerance to device variation *Adv. Mater.* **25** 1774–9

[53] Woo J, Moon K, Song J, Lee S, Kwak M, Park J and Hwang H 2016 Improved synaptic behavior under identical pulses using AlO_x/HfO_2 bilayer RRAM array for neuromorphic systems *IEEE Electron Device Lett.* **37** 994–7

[54] Yao P, Wu H, Gao B, Tang J, Zhang Q, Zhang W, Yang J and Qian Q 2020 Fully hardware-implemented memristor convolutional neural network *Nature* **577** 641–6

[55] Chekol S A, Menzel S, Ahmad R W, Waser R and Hoffmann-Eifert S 2021 Effect of the threshold kinetics on the filament relaxation behavior of Ag-based diffusive memristors *Adv. Funct. Mater.* **32** 2111242

[56] Covi E, Wang W, Lin Y-H, Farronato M, Ambrosi E and Ielmini D 2021 Switching dynamics of Ag-based filamentary volatile resistive switching devices: I. Experimental characterization *IEEE Trans. Electron Devices* **68** 4335–41

[57] Abbas H, Abbas Y, Hassan G, Sokolov A S, Jeon Y-R, Ku B, Kang C J and Choi C 2020 The coexistence of threshold and memory switching characteristics of ALD HfO_2 memristor synaptic arrays for energy-efficient neuromorphic computing *Nanoscale* **12** 14120–34

[58] Bricalli A, Ambrosi E, Laudato M, Maestro M, Rodriguez R and Ielmini D 2016 SiO_x-based resistive switching memory (RRAM) for crossbar storage/select elements with high on/off ratio *2016 IEEE Int. Electron Devices Meeting (IEDM)* pp 4.3.1–4

[59] Song J, Woo J, Prakash A, Lee D and Hwang H 2015 Threshold selector with high selectivity and steep slope for cross-point memory array *IEEE Electron Device Lett.* **36** 681–3

[60] Sun Y *et al* 2019 Performance-enhancing selector via symmetrical multilayer design *Adv. Funct. Mater.* **29** 1808376

[61] Midya R *et al* 2017 Anatomy of Ag/Hafnia-based selectors with 10^{10} nonlinearity *Adv. Mater.* **29** 1604457

[62] Wang Z *et al* 2017 Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing *Nat. Mater.* **16** 101–8

[63] Wang W, Covi E, Milozzi A, Farronato M, Ricci S, Sbandati C, Pedretti G and Ielmini D 2021 Neuromorphic motion detection and orientation selectivity by volatile resistive switching memories *Adv. Intell. Syst.* **3** 2000224

[64] Wang W, Covi E, Lin Y-H, Ambrosi E, Milozzi A, Sbandati C, Farronato M and Ielmini D 2021 Switching dynamics of Ag-based filamentary volatile resistive switching devices: II. Mechanism and modeling *IEEE Trans. Electron Devices* **68** 4342–9

[65] Wang W, Wang M, Ambrosi E, Bricalli A, Laudato M, Sun Z, Chen X and Ielmini D 2019 Surface diffusion-limited lifetime of silver and copper nanofilaments in resistive switching devices *Nat. Commun.* **10** 81

[66] Li Y *et al* 2020 High-uniformity threshold switching HfO_2-based selectors with patterned Ag nanodots *Adv. Sci.* **7** 2002251

[67] Milo V, Anzalone F, Zambelli C, Pérez E, Mahadevaiah M K, Ossorio Ó G, Olivo P, Wenger C and Ielmini D 2021 Optimized programming algorithms for multilevel RRAM in hardware neural networks *2021 IEEE Int. Reliability Physics Symposium (IRPS)* pp 1–6

[68] Ambrogio S, Balatti S, Cubeta A, Calderoni A, Ramaswamy N and Ielmini D 2014 Statistical fluctuations in HfO_x resistive-switching memory: I. Set/reset variability *IEEE Trans. Electron Devices* **61** 2912–9

[69] Covi E, George R, Frascaroli J, Brivio S, Mayr C, Mostafa H, Indiveri G and Spiga S 2018 Spike-driven threshold-based learning with memristive synapses and neuromorphic silicon neurons *J. Phys. D: Appl. Phys.* **51** 344003

[70] Ambrogio S, Balatti S, Cubeta A, Calderoni A, Ramaswamy N and Ielmini D 2014 Statistical fluctuations in HfO_x resistive-switching memory: II. Random telegraph noise *IEEE Trans. Electron Devices* **61** 2920–7

[71] Puglisi F M, Pavan P, Vandelli L, Padovani A, Bertocchi M and Larcher L 2015 A microscopic physical description of RTN current fluctuations in HfO_x RRAM *2015 IEEE Int. Reliability Physics Sympo.* pp 5B.5.1B.5.6–5

[72] Ambrogio S, Balatti S, McCaffrey V, Wang D C and Ielmini D 2015 Noise-induced resistance broadening in resistive switching memory: I. Intrinsic cell behavior *IEEE Trans. Electron Devices* **62** 3805–11

[73] Ambrogio S, Balatti S, McCaffrey V, Wang D C and Ielmini D 2015 Noise-induced resistance broadening in resistive switching memory: II. Array statistics *IEEE Trans. Electron Devices* **62** 3812–9

[74] Perez E, Grossi A, Zambelli C, Olivo P, Roelofs R and Wenger C 2017 Reduction of the cell-to-cell variability in $Hf_{1-x}Al_xO_y$ based RRAM arrays by using program algorithms *IEEE Electron Device Lett.* **38** 175–8

[75] Covi E, Brivio S, Fanciulli M and Spiga S 2015 Synaptic potentiation and depression in Al:HfO_2-based memristor *Microelectron. Eng.* **147** 41–4

[76] Hou Y *et al* 2013 Self-compliance multilevel resistive switching characteristics in tin/HfOx/Al/Pt RRAM devices *2013 IEEE Int. Conf. of Electron Devices and Solid-State Circuits* pp 1–2

[77] Milo V, Zambelli C, Olivo P, Pérez E, Mahadevaiah M K, Ossorio O G, Wenger C and Ielmini D 2019 Multilevel HfO_2-based RRAM devices for low-power neuromorphic networks *APL Mater.* **7** 081120

[78] Pedretti G, Mannocci P, Li C, Sun Z, Strachan J P and Ielmini D 2021 Redundancy and analog slicing for precise in-memory machine learning-Part I: programming techniques *IEEE Trans. Electron Devices* **68** 4373–8

[79] Dalgaty T, Payvand M, Moro F, Ly R, Pebay-Peyroula B, Casas F, Indiveri J, Vianello G and Vianello E 2019 Hybrid neuromorphic circuits exploiting non-conventional properties of RRAM for massively parallel local plasticity mechanisms *APL Mater.* **7** 081125

[80] Payvand M, Demirag Y, Dalgaty T, Vianello E and Indiveri G 2020 Analog weight updates with compliance current modulation of binary ReRAMS for on-chip learning *2020 IEEE Int. Symp. on Circuits and Systems (ISCAS)* pp p 1–5

[81] Frascaroli J, Brivio S, Covi E and Spiga S 2018 Evidence of soft bound behaviour in analogue memristive devices for neuromorphic computing *Sci. Rep.* **8** 7178

[82] Brivio S, Covi E, Serb A, Prodromakis T, Fanciulli M and Spiga S 2016 Experimental study of gradual/abrupt dynamics of HfO_2-based memristive devices *Appl. Phys. Lett.* **109** 133504

[83] Brivio S, Conti D, Nair M V, Frascaroli J, Covi E, Ricciardi C, Indiveri G and Spiga S 2019 Extended memory lifetime in spiking neural networks employing memristive synapses with nonlinear conductance dynamics *Nanotechnology* **30** 015102

[84] Brivio S, Ly R B, Vianello E and Spiga S 2021 Non-linear memristive synaptic dynamics for efficient unsupervised learning in spiking neural networks *Front. Neurosci.* **15** 580909

[85] Yu S, Gao B, Fang F, Yu H, Kang J and Wong H-S 2013 Stochastic learning in oxide binary synaptic device for neuromorphic computing *Front. Neurosci.* **7** 186

[86] Stathopoulos S, Serb A, Khiat K, Ogorzalek M and Prodromakis T 2019 A memristive switching uncertainty model *IEEE Trans. Electron Devices* **66** 2946–53

[87] Brivio S, Frascaroli J, Covi E and Spiga S 2019 Stimulated ionic telegraph noise in filamentary memristive devices *Sci. Rep.* **9** 6310

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

[89] Garbin D, Bichler O, Vianello E, Rafhay Q, Gamrat C, Perniola L, Ghibaudo G and DeSalvo B 2014 Variability-tolerant convolutional neural network for pattern recognition applications based on OxRAM synapses *Electron Devices Meeting (IEDM), 2014 IEEE Int.* pp 28.4.1–4

[90] Jiang H, Belkin D, Saval'ev S E, Lin S and Wang Z 2017 A novel true random number generator based on a stochastic diffusive memristor *Nat. Commun.* **8** 882

[91] Chekol S A, Cüppers F, Waser R and Hoffmann-Eifert S 2021 An Ag/HfO_2/Pt threshold switching device with an ultra-low leakage ($<$10 fA), high on/off ratio ($>10^{11}$), and low threshold voltage ($<$0.2 V) for energy-efficient neuromorphic computing *2021 IEEE Int. Memory Workshop (IMW)* pp 1–4

[92] Lee Y, Mahata C, Kang M and Kim S 2021 Short-term and long-term synaptic plasticity in Ag/HfO_2/SiO_2/Si stack by controlling conducting filament strength *Appl. Surf. Sci.* **565** 150563

[93] Wu Q *et al* 2018 Full imitation of synaptic metaplasticity based on memristor devices *Nanoscale* **10** 5875–81

[94] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44

[95] Li C *et al* 2018 Efficient and self-adaptive *in situ* learning in multilayer memristor neural networks *Nat. Commun.* **9** 2385

[96] Romero-Zaliz R, Pérez E, Jiménez-Molinos F, Wenger C and Roldàn J B 2021 Influence of variability on the performance of HfO_2 memristor-based convolutional neural networks *Solid-State Electron.* **185** 108064

[97] Hirtzlin T, Bocquet M, Ernoult M, Klein J O, Nowak E, Vianello E, Portal J M and Querlioz D 2019 Hybrid analog–digital learning with differential RRAM synapses *2019 IEEE Int. Electron Devices Meeting (IEDM)* pp 22.6.1–4

[98] Sun X, Yin S, Peng X, Liu R, Seo J-S and Yu S 2018 XNOR-RRAM: a scalable and parallel resistive synaptic architecture for binary neural networks *2018 Design, Automation Test in Europe Conf. Exhibition (DATE)* pp 1423–8

[99] Laborieux A, Bocquet M, Hirtzlin T, Klein J-O, Nowak E, Vianello E, Portal and Querlioz D 2021 Implementation of ternary weights with resistive RAM using a single sense operation per synapse *IEEE Trans. Circuits Syst.* I **68** 138–47

[100] Pedretti G, Graves C E, Serebryakov S, Mao R, Sheng S, Foltin M, Li C and Strachan J 2021 Tree-based machine learning performed in-memory with memristive analog CAM *Nat. Commun.* **12** 5806

[101] Pedretti G, Milo V, Ambrogio S, Carboni R, Bianchi S, Calderoni A, Ramaswamy N, Spinelli A S and Ielmini D 2017 Memristive neural network for on-line learning and tracking with brain-inspired spike timing dependent plasticity *Sci. Rep.* **7** 5788

[102] Ielmini D and Ambrogio S 2019 *Nanotechnology* **31** 092001

[103] Li C, Roth R M, Graves C, Sheng X and Paul Strachan J 2020 Analog error correcting codes for defect tolerant matrix multiplication in crossbars *2020 IEEE Int. Electron Devices Meeting (IEDM)* pp 36.6.1–4

[104] Bocquet M, Hirztlin T, Klein J-O, Nowak E, Vianello E, Portal J-M and Querlioz D 2018 In-memory and error-immune differential RRAM implementation of binarized deep neural networks *2018 IEEE Int. Electron Devices Meeting (IEDM)* pp 20.6.1–4

[105] Grossi A, Vianello E, Zambelli C, Royer P, Noel J-P, Giraud B, Perniola L, Olivo P and Nowak E 2018 Experimental investigation of 4 kb RRAM arrays programming conditions suitable for TCAM *IEEE Trans. VLSI Syst.* **26** 2599–607

[106] Moradi S, Qiao N, Stefanini F and Indiveri G 2018 A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs) *IEEE Trans. Biomed. Circuits Syst.* **12** 106–22

[107] Pedretti G, Ambrosi E and Ielmini D 2021 Conductance variations and their impact on the precision of in-memory computing with resistive switching memory (RRAM) *2021 IEEE Int. Reliability Physics Symposium (IRPS)* pp 1–8

[108] Dalgaty T, Esmanhotto E, Castellani N, Querlioz D and Vianello E 2021 *Ex situ* transfer of Bayesian neural networks to resistive memory-based inference hardware *Adv. Intell. Syst.* **3** 2000103

[109] Balatti S, Ambrogio S, Carboni R, Milo V, Wang Z, Calderoni A, Ramaswamy N and Ielmini D 2016 Physical unbiased generation of random numbers with coupled resistive switching devices *IEEE Trans. Electron Devices* **63** 2029–35

[110] Balatti S, Ambrogio S, Wang Z and Ielmini D 2015 True random number generation by variability of resistive switching in oxide-based devices *IEEE J. Emerg. Sel. Top. Circuits Syst.* **5** 214–21

[111] Dalgaty T, Castellani N, Turck C, Harabi K-E, Querlioz D and Vianello E 2021 *In situ* learning using intrinsic memristor variability via Markov chain Monte Carlo sampling *Nat. Electron.* **4** 151

[112] Covi E, Donati E, Liang X, Kappel D, Heidari H, Payvand M and Wang W 2021 Adaptive extreme edge computing for wearable devices *Front. Neurosci.* **15** 611300

[113] Ambrogio S, Balatti S, Milo V, Carboni R, Wang Z Q, Calderoni A, Ramaswamy N and Ielmini D 2016 Neuromorphic learning and recognition with one-transistor–one-resistor synapses and bistable metal oxide RRAM *IEEE Trans. Electron Devices* **63** 1508–15

[114] Bi G-Q and Poo M-M 1998 Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type *J. Neurosci.* **18** 10464–72

[115] Matveyev Y, Egorov K, Markeev A and Zenkevich A 2015 Resistive switching and synaptic properties of fully atomic layer deposition grown TiN/HfO_2/TiN devices *J. Appl. Phys.* **117** 044901

[116] Ambrogio S, Balatti S, Nardi F, Facchinetti S and Ielmini D 2013 Spike-timing dependent plasticity in a transistor-selected resistive switching memory *Nanotechnology* **24** 384012

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

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

[119] Nair M V and Dudek P 2015 Gradient-descent-based learning in memristive crossbar arrays *Proc. of IEEE Int. Joint Conf. on Neural Networks (IJCNN)* pp 1–7

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

[121] Bienenstock , Cooper L N and Munro P W 1982 Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex *J. Neurosci.* **2** 32–48

[122] Bengel C, Cüppers F, Payvand M, Dittmann R, Waser R, Hoffmann-Eifert S and Menzel S 2021 Utilizing the switching stochasticity of HfO_2/TiO_x-based ReRAM devices and the concept of multiple device synapses for the classification of overlapping and noisy patterns *Front. Neurosci.* **15** 621

[123] Valentian A, Rummens F, Vianello E, Mesquida T, de Boissac C L, Bichler O and Reita C 2019 Fully integrated spiking neural network with analog neurons and RRAM synapses *2019 IEEE Int. Electron Devices Meeting (IEDM)* pp 14.3.1–4

[124] Regev A, Bricalli A, Piccolboni G, Valentian A, Mesquida T, Molas G and Nodin J 2020 Fully-integrated spiking neural network using SiO_x-based RRAM as synaptic device *2020 2nd IEEE Int. Conf. on Artificial Intelligence Circuits and Systems (AICAS)* pp 145–8

[125] Zhao Y D, Kang J F and Ielmini D 2021 Materials challenges and opportunities for brain-inspired computing *MRS Bull.* **46** 978

[126] Wang W, Pedretti G, Milo V, Carboni R, Calderoni A, Ramaswamy N, Spinelli A S and Ielmini D 2018 Learning of spatiotemporal patterns in a spiking neural network with resistive switching synapses *Sci. Adv.* **4** eaat4752

[127] Glukhov A, Milo V, Baroni A, Lepri N, Zambelli C, Olivo P, Pérez E, Wenger C and Ielmini D 2022 Statistical model of program/verify algorithms in resistive-switching memories for in-memory neural network accelerators *2022 IEEE Int. Reliability Physics Symp. (IRPS)* pp 3C.33-7–1

[128] Chakraborty I, Roy D and Roy K 2018 Technology aware training in memristive neuromorphic systems for nonideal synaptic crossbars *IEEE Trans. Emerg. Top. Comput. Intell.* **2** 335–44

[129] Lepri N, Baldo M, Mannocci P, Glukhov A, Milo V and Ielmini D 2022 Modeling and compensation of IR drop in crosspoint accelerators of neural networks *IEEE Trans. Electron Devices* **69** 1575–81

[130] Kang H, Park J, Lee D, Kim H W, Jin S, Ahn M and Woo J 2021 Two- and three-terminal HfO_2-based multilevel resistive memories for neuromorphic analog synaptic elements *Neuromorph. Comput. Eng.* **1** 021001

[131] Kim S *et al* 2019 Metal-oxide based, CMOS-compatible ECRAM for deep learning accelerator *2019 IEEE Int. Electron Devices Meeting (IEDM)* pp 35.7.1–4

[132] Lanza M *et al* 2021 Standards for the characterization of endurance in resistive switching devices *ACS Nano* **15** 17214–31

[133] Balatti S, Ambrogio S, Wang Z, Sills S, Calderoni A, Ramaswamy N and Ielmini D 2015 Voltage-controlled cycling endurance of HfO_x-based resistive-switching memory *IEEE Trans. Electron Devices* **62** 3365–72

[134] Agarwal S, Plimpton S J, Hughart D R, Hsia A H, Richter I, Cox J A, James C D and Marinella M J 2016 Resistive memory device requirements for a neural algorithm accelerator *2016 Int. Joint Conf. on Neural Networks (IJCNN)* pp 929–38

[135] Stecconi T *et al* 2022 Filamentary TaO_x/HfO_2 ReRAM devices for neural networks training with analog in-memory computing *Adv. Electron. Mater.* **8** 2200448

[136] Tang J *et al* 2018 ECRAM as scalable synaptic cell for high-speed, low-power neuromorphic computing *2018 IEEE Int. Electron Devices Meeting (IEDM)* pp 13.1.1–4

[137] Ambrogio S *et al* 2018 Equivalent-accuracy accelerated neural-network training using analogue memory *Nature* **558** 60

[138] Milo V, Malavena G, Monzio Compagnoni C and Ielmini D 2020 Memristive and CMOS devices for neuromorphic computing *Materials* **13** 166

[139] Monzio Compagnoni C, Goda A, Spinelli A S, Feeley P, Lacaita A L and Visconti A 2017 Reviewing the evolution of the nand flash technology *Proc. IEEE* **105** 1609–33

[140] Raoux S, Wełnic W and Ielmini D 2010 Phase change materials and their application to nonvolatile memories *Chem. Rev.* **110** 240–67

[141] Apalkov D, Dieny D and Slaughter J 2016 Magnetoresistive random access memory *Proc. IEEE* **104** 1796–830

[142] Shao Q 2021 Roadmap of spin–orbit torques *IEEE Trans. Magn.* **57** 800439

[143] Lehninger D, Lederer M, Ali T, Kämpfe T, Mertens K and Seidel K 2021 Enabling ferroelectric memories in BEoL—towards advanced neuromorphic computing architectures *2021 IEEE Int. Interconnect Technology Conf. (IITC)* pp 1–4

[144] Mulaosmanovic H, Ocker J, Müller S, Noack M, Müller J, Polakowski P, Mikolajick T and Slesazeck S 2017 Novel ferroelectric FET based synapse for neuromorphic systems *2017 Symp. on VLSI Technology* pp T176–7

[145] Liu C, Wang L-G, Cao Y-Q, Wu M-Z, Xia Y-D, Wu D and Li A-D 2019 Synaptic functions and a memristive mechanism on $Pt/AlO_x/HfO_x/TiN$ bilayer-structure memristors *J. Phys. D: Appl. Phys.* **53** 035302

[146] Shin H S, Song H, Kim D H, Martinez A, Cheong W H and Kim K M 2022 Multimode synaptic operation of a $HfAlO_x$-based memristor as a metaplastic device for neuromorphic applications *ACS Appl. Electron. Mater.* **4** 3786

[147] Ryu J-H, Mahata C and Kim S 2021 Long-term and short-term plasticity of Ta_2O_5/HfO_2 memristor for hardware neuromorphic application *J. Alloys Compd.* **850** 156675