Emerging neuromorphic devices

Daniele Ielmini[1] and Stefano Ambrogio[2]
[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano and IU.NET, Piazza L. da Vinci 32 – 20133 Milano, Italy. Email: daniele.ielmini@polimi.it
[2]IBM Research-Almaden, 650 Harry Road, 95120 San Jose, CA, USA.
Email: stefano.ambrogio@ibm.com

Artificial intelligence (AI) has the ability of revolutionizing our lives and society in a radical way, by enabling machine learning in the industry, business, health, transportation, and many other fields. The ability to recognize objects, faces, and speech, requires however exceptional computational power and time, which is conflicting with the current difficulties in transistor scaling due to physical and architectural limitations. As a result, to accelerate the progress of AI, it is necessary to develop materials, devices, and systems that closely mimic the human brain.

In this work, we review the current status and challenges on the emerging neuromorphic devices for brain-inspired computing. First, we provide an overview of the memory device technologies which have been proposed for synapse and neuron circuits in neuromorphic systems. Then, we describe the implementation of synaptic learning in the two main types of neural networks, namely the deep neural network (DNN) and the spiking neural network (SNN). Bio-inspired learning, such as the spike-timing dependent plasticity (STDP) scheme, is shown to enable unsupervised learning processes which are typical of the human brain. Hardware implementations of SNNs for the recognition of spatial and spatiotemporal patterns are also shown to support the cognitive computation in silico. Finally, we explore the recent advances in reproducing bio-neural processes via the device physics, such as insulating-metal transitions, nanoionics drift/diffusion, and magnetization flipping in spintronic devices. By harnessing the device physics in emerging materials, neuromorphic engineering with advanced functionality, higher density and better energy efficiency can be developed.

1. Introduction

After more than 50 years from its start, the evolution of the microelectronic industry can no longer be adequately described by the Moore's law of scaling the transistor size [1]. For years, making the transistor smaller have meant improving the density, performance and power consumption of a digital circuit. More recently, transistor miniaturization has been replaced by more advanced approaches, such as the introduction of high-k materials for the gate dielectric [2], the adoption of enhanced transistor layouts such as trigate structures [3], and possibly in the future the use of alternative switch concepts [4]. Most importantly, new computing methodologies such as quantum computing [5], stochastic computing [6], and analogue computing [7] are currently under scrutiny to overcome the main limitations of the digital circuits.

Among the novel computing approaches under investigation, neuromorphic computing is probably the most promising. Neuromorphic engineering defines the development of systems that emulate the human brain to achieve high energy efficiency, parallelism, and ability in cognitive tasks, such as object recognition, association, adaptation, and learning. The concept of neuromorphic systems is not novel, being first introduced in the 1980s as a branch of analogue circuit engineering [8]. The original neuromorphic concept is based on analogue circuits with extensive use of subthreshold-biased transistors, as a means to minimize the energy consumption and exploit the similarity between carrier diffusion in the transistor and atomistic transport in the ionic channel of a biological synapse [9]. During the years, analogue circuits for neurons and synaptic functions have been proposed [10], and led to the development of general-purpose chip demonstrators [11]. Although the complementary metal-oxide semiconductor (CMOS) technology is essential to enable the integration of large-scale neuromorphic systems, it does not easily provide some of the inherent features of the neurobiological network, such as the long-term plasticity, the stochastic behavior, and the ability to update internal variable, such as synaptic weight and membrane potential, as a function of spike timing and frequency.

In the last 10 years, there has been a wide exploration of new devices as technology enablers of neuromorphic computing. The class of emerging memories is very promising for neuromorphic computation, thanks to the ability to store analogue values in nonvolatile way, combined with the extremely small device size [12,13]. Also, emerging nonvolatile memories feature a unique device physics that can provide a broad portfolio of functions, such as time-dependent dynamics, low-voltage operation and stochastic switching. Most importantly, emerging memories naturally enable the so-called in-memory computing paradigm, where data are processed directly within the memory, thus with no need for any energy- and time-consuming transfer to/from the memory circuit [14]. Note, in fact, that our brain is essentially built on the concept of in-memory computing, where neurons and synapses serve the function of both memory and computing elements [15,16]. Memory devices can also be organized with array architectures, such as the cross-point array [17,18] and the one-transistor/one-resistor (1T1R) array [19,20], which strictly resemble the structure of a neural network, where each memory conductance plays the role of a synaptic weight [14,21]. For all these reasons, emerging memories, also known as memristors, are considered a strong contender for the implementation of high-density, low energy neuromorphic circuits.

This work provides an overview on the emerging devices for neuromorphic computing with an emphasis on nonvolatile memories for synaptic and neuron applications. First, the emerging memory devices that are currently investigated for neuromorphic applications are reviewed in terms of the physical switching mechanisms and inherent performance in terms of speed, multilevel operation, and scaling. Then, the synaptic concepts based on emerging memories are described, referring to various types of neural networks and learning rules, aiming at either supervised or unsupervised training. Examples of neural networks implementing brain-inspired learning rules for pattern recognition are shown. Neuron circuits employing emerging devices are then reported, including various classes of oscillating, accumulating, and stochastic neurons. Finally, examples of neural networks combining emerging neuron and synaptic devices are presented. The open challenges and remaining gaps for the development of this field are finally summarized.

2. Neuromorphic networks

Neuromorphic engineering aims at developing circuits that compute as the human brain. An essential feature of any neuromorphic circuit is the neural network architecture, where data are sent by neuronal terminals through a highly-parallel net of synaptic paths. The concept of neural network can be traced back to the neuron model proposed by McCulloch and Pitts, who for the first time described the neuron as a mathematical function of the synaptic inputs [22]. The prototypical version of the neural network is the perceptron, which is capable of recognizing linearly separable classes of objects [23]. More advanced schemes of neural networks, generally referred under the term of deep learning, have been first proposed [24] and more recently led to an increased interest [25] for the outstanding performance in object and face recognition, even matching or surpassing the human capability [26]. More brain-inspired concepts have been developed, such as the concept of reinforcement learning which enables self-adaptation within a neural network as for the human brain. For instance, it has been shown that a neuromorphic platform can learn to play videogames [27] or the ancient game of Go [28] by iteratively playing games, and autonomously learning from successes and failures. Most of these achievements were however obtained by running software programs in digital computers, taking advantage of the outstanding performance of advanced central processing units (CPUs) and graphics processing units (GPUs) to expedite the supervised training for setting the synaptic weights within the network. The digital computation is extremely power hungry, while lacking any similarity with the brain architecture. To realize energy efficient, scalable neuromorphic hardware systems, it is necessary to mimic the brain from its very fundamental architecture, communication and computation schemes.

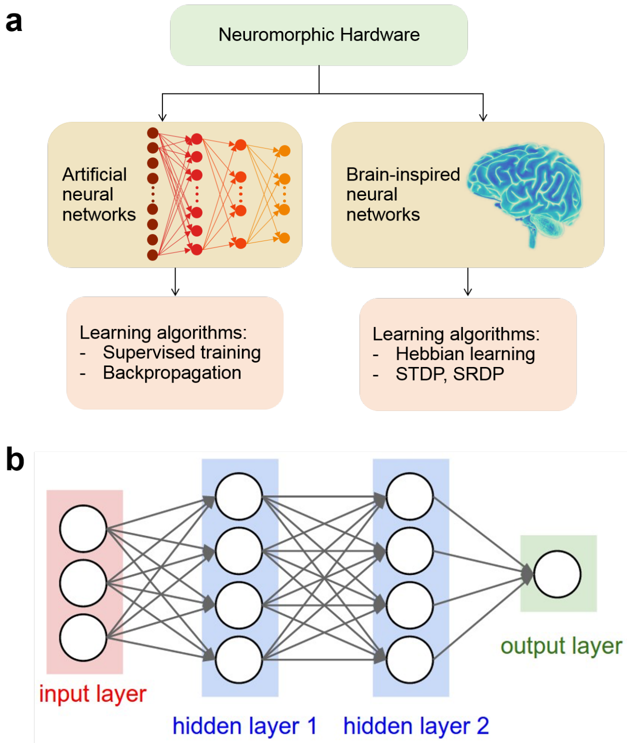

Fig. 1 Illustrative sketch of neuromorphic hardware options. (a) Neuromorphic hardware can be implemented via both artificial neural networks (ANNs) and brain-inspired neural networks. These two approaches rely on a different set of algorithms and learning rule, e.g., the backpropagation algorithm in ANNs, or the spike timing dependent plasticity (STDP) in brain-inspired concepts. (b) A generic neural network, organized as neuron layers where each neuron is connected by synapses to neurons in the previous layer, and the next layer (b). The number of layers defines the 'depth' of the neural network.

Fig. 1a schematically illustrates the two basic types of neuromorphic hardware, namely artificial neural networks (ANNs) and brain-inspired networks. These types of hardware differ mainly from the methodology for training the network synapses, while sharing the general neural network architecture. In ANN, the deep structure with many layers can only be trained by supervised learning algorithms such as the backpropagation scheme [25]. On the other hand, brain-inspired networks adopt learning rules which are derived from the neurobiological systems, such as Hebbian learning and the spike timing dependent plasticity (STDP) [29,30]. Note that, although ANNs are also inspired by the brain, the training algorithms such as the backpropagation technique are not, which justifies the nomenclature in Fig. 1a.

Both bio-inspired networks and ANNs in Fig. 1a rely on the fundamental neural network structure of Fig. 1b. This shows a fully-connected multilayer perceptron (MLP), which is the prototypical network for deep learning for object, face, and image recognition. The network consists of layers of neurons, including (i) input neurons, providing the input pattern of information, (ii) output neurons, providing the solution to the classification problem, and (iii) a number of hidden layers, where the intermediate solutions from input to output variables are found. Each neuron in the network can execute a summation or integration of the input signals, followed by a non-linear operation, such as logistic or sigmoidal function. The output of each neuron is then transmitted to neurons of the next layer via synaptic connections, which multiply the signal by a proper synaptic weight. The network can have a feed-forward structure, meaning that the information is sent from the input layer to the output layer, or a recurrent network, where a feedback connection is also present from a neuron layer back to another preceding layer. One example of recurrent network is the Hopfield network, where a layer of neurons sends information toward themselves through a single layer of synaptic connections [31]. Information among neurons is generally sent via synchronous or asynchronous spikes.

3. Emerging memory devices

To implement the neural network of Fig. 1b in hardware, it is necessary to identify the most suitable circuit to represent the neuron and synapse function. In this scenario, emerging memory devices play a major role, since they can provide added functionality to the conventional CMOS technology, such as the ability to implement analogue, nonvolatile memory within a nanoscale region on the chip. The emerging memory also enables the in-memory computing approach where data are processed in situ [14]. Emerging memory devices can be divided in two categories, namely 2-terminal devices and 3 terminal devices, as illustrated in the following.

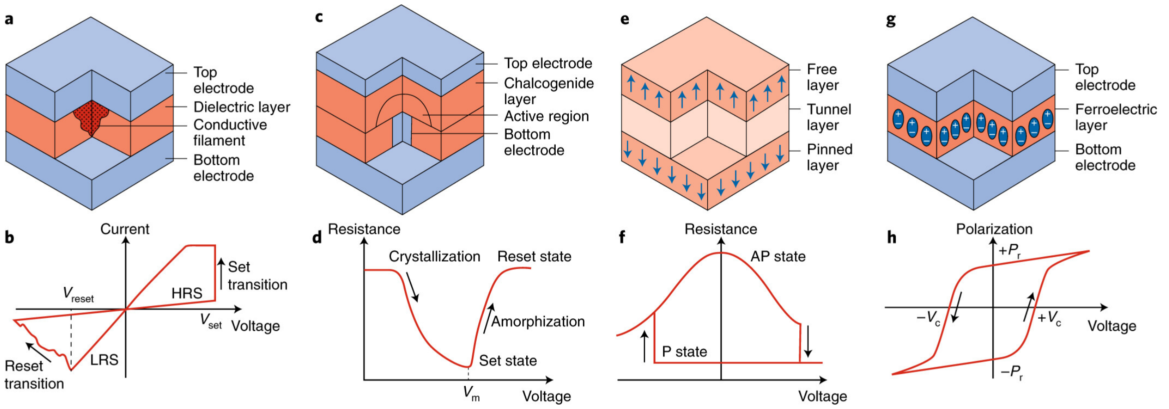

Fig. 2 Illustrative sketch of the memory devices that are considered for in-memory computing, including neuromorphic computing. (a) A resistive switching random access memory (RRAM), and (b) its representative current-voltage indicating bipolar switching between a low resistance state (LRS) and a high resistance state (HRS). The switching process originates from the ionic migration across a filamentary path across the metal-insulator-metal (MIM) stack of the RRAM. (c) A phase change memory (PCM), and (d) its resistance-voltage (R-V) characteristic, where the resistance drop at low voltage is due to the crystallization, and the resistance increase is due to melting and amorphization of the phase change material. (e) A spin-torque transfer magnetic random access memory (STT-MRAM), and (f) its R-V characteristic, where the transition between parallel (P) and antiparallel (AP) states dictates variations of the resistance. (g) Ferroelectric random access memory (FERAM), and (h) its polarization-voltage characteristic, indicating the typical hysteresis of ferroelectric insulating layer in the MIM stack. Reprinted with permission from [14]. Copyright 2018 Springer Nature Publishing.

3.1 2-terminal devices

Fig. 2 summarizes the 2-terminal devices that are currently studied for applications in neuromorphic computing circuits [14]. These include the resistive switching random access memory (RRAM), the phase change memory (PCM), the spin-transfer torque magnetic random access memory (STT-MRAM) and the ferroelectric random access memory (FERAM). These memory devices all share the same basic structure, with an insulator and/or active material sandwiched between 2 metal electrode layers. The application of external voltage pulses to the device will induce a change in a characteristic property of the memory device, which can be sensed as a variation in the resistance, or the electric/magnetic polarization. As a result, one can program, erase and read the memory by electrical operations on the memory device, which can retain the written state for long time, e.g., 10 years at elevated temperature. This is similar to the conventional nonvolatile Flash technology, which has been a consolidated memory device with extremely high density for the last 30 years [32,33]. However, Flash memory relies on the storage of charge within a floating gate of a MOS transistor, whereas all the emerging memory concepts in Fig. 2 are based on material properties which can be changed by electrical operations. Thanks to the charge-free material modification, the scalability of the emerging memory is generally superior to Flash memories.

The RRAM device in Fig. 2a consists of a metal-insulator-metal (MIM) structure, where the insulating layer can change its resistance from relatively large, in the high resistance state (HRS), to relatively low, in the low resistance state (LRS) [34-36]. The resistance change generally takes place at a localized region within the insulating layer, referred to as conductive filament (CF). To form the CF, RRAM is subjected to a preliminary operation, called forming, consisting of a soft dielectric breakdown to induce a local decrease of resistance. The set process allows to operate the transition from the HRS to the LRS, whereas the reset process is responsible for the transition from the LRS to the HRS. The set and reset processes are both induced by the application of voltage pulses, which can have the same bias polarity, in the case of unipolar RRAM [37,38], or, most typically, the opposite bias polarity, in the case of bipolar RRAM. Fig. 2b shows a typical current-voltage characteristic for a bipolar RRAM, indicating the set transition as a steep increase of the current at the positive set voltage V_{set} and the reset transition to the HRS starting at the negative voltage V_{reset}. The set and reset transitions are generally explained in terms of defect migration within the CF: for instance, the application of a reset voltage across the CF leads to drift and diffusion of the ionized defects, such as oxygen vacancies and metallic impurities, resulting in a retraction of the CF toward the negatively-biased electrode [39]. The CF retraction causes the formation of a depleted gap with low concentration of defects, hence high resistivity, which is responsible for the increase of resistance in the reset transition. Applying an opposite voltage leads to the migration of defects back into the depleted gap, which can decrease the resistance to the LRS. Various materials have been adopted for the insulating layer, most typically being a metal or semiconductor oxide such as HfO_2 [40], TaO_2 [41], or SiO_2 [42]. The RRAM resistance can be usually controlled with analogue precision between the HRS and the LRS, thus enabling multilevel cell (MLC) operation with storage of at least 3 bits [43,44]. RRAM also shows excellent downscaling to the 10 nm size [45] and the capability for 3D integration [46], thus serving as a promising technology for high density storage class memory.

Fig. 2c schematically illustrates a PCM structure, where the device resistance is changed upon a phase transformation of the active material [47-49]. The latter usually consists of a chalcogenide material, such as $Ge_2Sb_2Te_5$ (GST) [50]. The active material is usually in a crystalline phase, with a doped-semiconductor band structure and a relatively large conductivity [51,52]. The crystalline phase can be changed to amorphous by the application of an electrical pulse, called the reset pulse, which is large enough to locally induce melting in the chalcogenide material [53]. The amorphous phase has a high resistivity thanks to the pinning of the Fermi level at the midgap. The crystalline phase can then be obtained again by the application of a set pulse below the melting point, which causes the fast crystallization in the amorphous region thanks to the local Joule heating [53]. Fig. 2d shows the resistance-voltage (R-V) characteristic of a PCM, indicating the resistance R measured after the application of a pulse of voltage V to a device initially prepared in the amorphous phase. At relatively low voltage, the device shows a transition from the high-resistivity amorphous phase to the crystalline phase. For voltage above the melting point V_m, the resistance increases because of the increasing amorphization within the active layer. The PCM is generally operated by unipolar set/reset pulses, although bipolar operation of the PCM has also been reported [54]. Various chalcogenide materials have been proposed to date, most typically to increase the crystallization temperature with respect to conventional GST, thus enhancing the retention capability of the device. High-temperature materials include GeSb [55], InGeSbTe [56], and Ge-rich GST [57,58]. Similar to the RRAM, analogue control of resistance and MLC operation have been reported [59].

Fig. 2e shows the structure of a STT-MRAM device, consisting of a MIM stack with ferromagnetic (FM) metal electrodes, e.g., CoFeB, and a tunneling insulating layer, e.g., MgO. This structure is also known as a magnetic tunneling junction (MTJ), where the magnetic polarization in the two FM layers can be either parallel (P) or antiparallel (AP), resulting in a low or high resistance values, respectively. The different resistance is due to the coherent tunneling of spin polarized electrons, which has a high probability in the case where the FM polarization is parallel [60,61]. Of the two FM layers in the MTJ, one is the pinned layer, which is stabilized by the presence of adjacent magnetic layers, such as a synthetic antiferromagnetic (SAF) stack [62]. The other FM layer is instead free to change its polarization, which can be switched by the spin-transfer torque mechanism [63]. Fig. 2f shows the typical R-V curve of a STT-MRAM device, indicating an AP-to-P transition at positive voltage and a P-to-AP transition at negative voltage. Note the V-dependent resistance of the AP state, which is due to the non-linear transport across the tunneling layer. The abrupt transition at the switching points indicate a binary behavior of the STT-MRAM, which therefore is hardly compatible with MLC operation and analogue-state storage. On the other hand, the FM polarization switching in the STT-MRAM is purely electronic, which enables a fast switching [64] and an extremely high cycling endurance [65].

Fig. 2g shows a FERAM, which consists of a MIM stack where the insulating layer is made with a ferroelectric (FE) material [66]. The electrical dipoles within the FE material can be oriented by applying an external bias, as shown in the hysteretical polarization-voltage (P-V) characteristic in Fig. 2h. In particular, applying a positive voltage to the top electrode results in a residual positive polarization P_r in the FE layer, while a negative voltage will lead to negative residual polarization -Pr. The reversal of the dipole polarization occurs at voltages above the coercive voltage V_c. Note that the resistance is not sensitive to the FE polarization, thus the FERAM in Fig. 2g cannot be used as a resistive memory. On the other hand, the displacement currents induced by the polarization switching can be sensed externally to probe the FE state, which provides the basic read operation of the FERAM [66]. This read operation is, however, destructive of the pre-existing state, which makes the FERAM readout relatively expensive in terms of time and energy. Typical FERAM materials include $PbZrTiO_3$ (PZT) [67], $BiSrTiO_3$ (BST) [68] and doped HfO_2 [69]. The discovery of FE properties of HfO_2, which is a mainstream material in the front end of line of CMOS technology, has significantly revived the research on FERAM in the last 5 years.

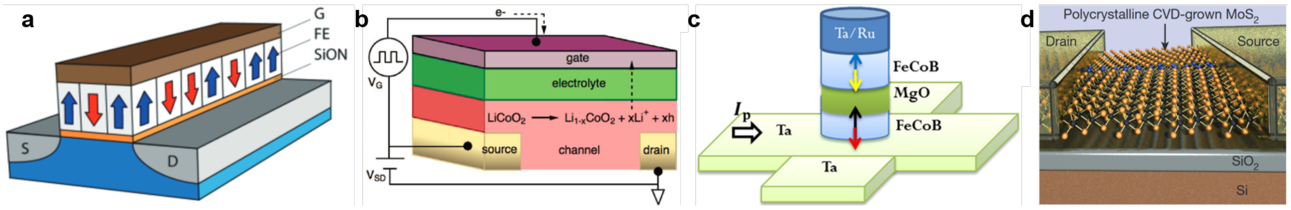

Fig. 3 Three-terminal synaptic memory devices. (a) A ferroelectric field effect transistor (FEFET), where the ferroelectric polarization of individual domains in the ferroelectric layer dictates the threshold voltage, thus serving as a nonvolatile memory and synaptic weight element. (b) Synaptic ionic transistor, where the application of a gate voltage signal induces migration of active ionic species (e.g., Li+) across the electrolyte, thus modifying the channel conductivity. (c) A spin-orbit-torque magnetic memory device (SOT-MRAM), where the current flowing across the heavy metal induces a change in the magnetic polarization in the adjacent ferroelectric layer, hence a change of resistance across the MTJ. (d) A memtransistor based on a 2D semiconductor, such as MoS_2, where the application of a source-drain voltage causes dislocation drift, hence a change of the source-drain conductivity. Reprinted with permission from [70,74,77,82]. Copyright 2017 IEEE, Copyright 2017 Wiley, Copyright 2014 AIP, Copyright 2018 Springer Nature Publishing.

3.2 3-terminal devices

Devices in Fig. 2 have a 2-terminal structure, which is practical in view of adopting a high density crosspoint architecture [14]. In many cases, however, memory devices can be associated with a transistor selector, which results in a 1T1R structure. The transistor gate thus introduces a third terminal, which may complicate the device structure and increase its footprint. However, the additional transistor improves the controllability of the memory state and prevents sneak paths in the array, which makes the 1T1R essential in most cases. Other device concepts or device structures show a 3-terminal architecture, as summarized in Fig. 3.

Fig. 3a shows the ferroelectric field-effect transistor (FEFET) which consists of a MOS transistor where the gate dielectric is a FE layer [70,71]. The polarization state of the FE layer can be controlled by the gate voltage and affects the threshold voltage V_T of the FEFET, which provides a straightforward, non-destructive read methodology. Second, the FEFET improves the compactness of the typical one-transistor/one-capacitor (1T1C) structure of the FERAM, and enables high density memory architectures such as the NAND arrays [68] and vertical 3D concepts [72]. HfO_2-based FEFET memory arrays have been recently demonstrated with 28 nm CMOS technology [73].

Fig. 3b shows a solid-electrolyte transistor, also known as electro-chemical random access memory (ECRAM), which consists of a transistor where the gate dielectric is made of a solid-state electrolyte for ion migration [74]. Typically, Li+ is used as migrating ion within a solid-state electrolyte, such as lithium phosphorous oxynitride (LiPON) [74]. The application of a positive gate voltage induces Li ion migration toward the $LiCoO_2$ channel, where Li reduction and intercalation cause the conductivity to decrease, thus resulting in a smaller drain current. A negative gate voltage instead induces Li de-intercalation from the channel and a consequent increase of conductivity [74]. Thanks to the decoupling of the write and read paths, the device shows enhanced linearity of conductance update, which makes this device concept extremely promising as a synaptic connection in supervised neural networks. Organic electro-chemical transistors based on proton migration were also shown for flexible circuits [75]. Device scaling and ns-operation were demonstrated with Li+ ECRAM with WO_3 channel, thus supporting this technology for fast, energy-efficient circuits [76].

Fig. 3c shows the spin-orbit torque magnetic random access memory (SOT-MRAM) [77,78]. Similar to the STT-MRAM, the core concept in the SOT-MRAM is an MTJ, where the P or AP states of the FM layers dictate the resistance. To switch the magnetization state, instead, a current is fed across the bottom electrode, consisting of a heavy metal (HM) such as Pt [79] or Ta [80]. The horizontal current can induce an accumulation of spin-polarized electrons at the HM/FM interface, thus inducing the magnetization switching in the MTJ [79]. The current-induced spin accumulation is generally explained by the spin Hall or the Rashba effects [81]. Fast switching time of about 0.4 ns was demonstrated by using a large current density of about 300 MAcm^{-2} in the HM electrode [79]. However, the large current is not fed through the sensitive MTJ, thus considerably extending the cycling endurance of the SOT-MRAM with respect to the STT-MRAM.

Fig. 3d shows the memristive transistor, or memtransistor, consisting of a polycrystalline 2D semiconductor, such as MoS_2, acting as a channel in a MOS transistor [82,83]. The application of a large source-drain voltage activates a resistance transition, which is explained by the migration of grain boundaries [83] or Li+ impurities [84] in the 2D semiconductor. As a result, the mem-transistor can be viewed as a transistor with a memory, depending on the previous history of pulses applied across the channel. Neuromorphic properties of spike accumulation and spike timing plasticity have been experimentally evidenced [82].

4. Artificial synapses in ANNs

Neuromorphic computing in the neural network of Fig. 1b requires both synapses and neurons. In the human brain, there is a ratio of about 10,000 between synapses and neurons, thus the synaptic element should be extremely small and energy-efficient to enable a cognitive computation with brain-like connectivity. To meet this goal, the emerging devices in Figs. 2 and 3 have been considered as potential artificial synapses in neural networks, in both ANN and SNN computing approach.

In general, an artificial synapse serves as an electrical connection between a pre-synaptic neuron, or PRE, and a post-synaptic neuron, or POST. The conductance of the artificial synapse provides the synaptic weight that multiplies the PRE signal before it is fed to the POST, according to the formula [22]:

$$y_j = \sum_i w_{ij} x_i \qquad (1)$$

where y_j is the input signal at the j-th POST, x_i is the signal of the i-th PRE, and w_{ij} is the synaptic weight connecting the i-th PRE with the j-th POST. Such scheme naturally enables the implementation of algorithms requiring the summation of many individual contributions. Typically, resistive devices are organized in crossbar arrays [21], providing high integration density and high computational parallelism, which represent essential elements to efficiently implement deep learning algorithms.

An example of a modern ANN, called deep neural network (DNN), is shown in Fig. 4. A DNN is a large neural network, composed of many neuron layers connected by synaptic weights in different connection schemes. For instance, Fig. 4 shows a typical fully-connected DNN, where all the neurons in one layer are connected to all the neurons in the subsequent layer, also referred to a multi-layer perceptron (MLP). Other implementations, instead, adopt the convolutional neural network (CNN) structure, where small sets of synaptic weights, organized in 2-D or 3-D kernels, are iteratively used to process and propagate information. The synaptic weights can be either positive or negative and must be trained with proper algorithms to perform a certain specific task, such as recognition of a dataset of objects or patterns.

DNNs typically adopt the backpropagation algorithm to train the synaptic weights for a wide variety of tasks, such as image recognition, speech processing and machine translation. The 4-layer MLP in Fig. 4, for instance, is trained on an image classification task, such as the recognition of handwritten digits from the MNIST dataset [85]. The images from a training dataset are forward propagated through the network, providing an x_i value for each neuron and a classification guess y_j for the images. Such guess is represented by the output of the last-layer neurons and is compared to the expected, or correct answer g_j, also known as the 'label'. By subtracting the two quantities, an error $\delta_j = y_j - g_j$ is obtained and backpropagated through the entire network, allowing the calculation of the error δ_j for the neurons of each layer. Finally, the synaptic weights w_{ij} are updated according to the formula:

$$\Delta w_{ij} = \eta \cdot x_i \cdot \delta_j, \qquad (2)$$

where η is the learning rate. This procedure is then repeated for every training image, and the entire training set is iteratively presented for many training cycles, called epochs. However, to evaluate the quality of such training, the network must be tested on previously unseen patterns, namely the test dataset [86].

Due to the large amount of weights typically involved in such networks, the training operation on CPUs and GPUs can become very expensive in terms of energy and time [87]. This is mainly because the multiple synaptic weights must be transferred between the memory and the processor, representing a major bottleneck (typically referred as "Von-Neumann bottleneck") and preventing a fast and energy efficient training process. To overcome this issue, several digital-custom implementations have been recently developed [88] to speed-up the training or forward inference of such networks [89]. However, the most promising approach in terms of density, speed and energy efficiency is the implementation of neural networks with crosspoint arrays, as shown in Fig. 5. The main advantage provided by crossbar arrays of non-volatile memories is the efficient calculation of Eq. (1), enabling a strong acceleration of training and forward inference tasks.

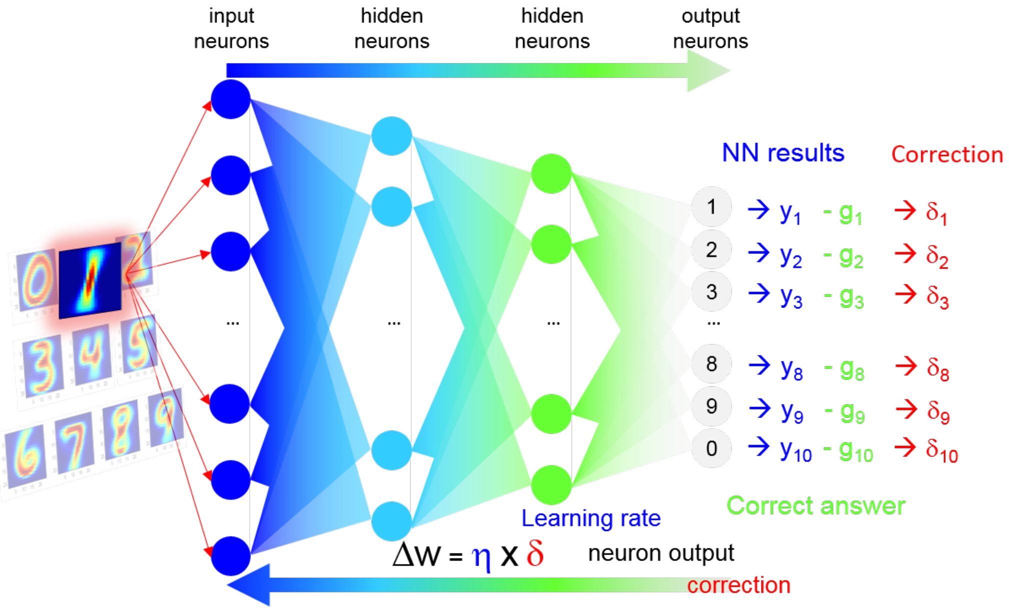
input neurons
hidden neurons
hidden neurons
output neurons
NN results
Correction
1 → y_1 - g_1 → δ_1
2 → y_2 - g_2 → δ_2
3 → y_3 - g_3 → δ_3
8 → y_8 - g_8 → δ_8
9 → y_9 - g_9 → δ_9
0 → y_{10} - g_{10} → δ_{10}
Correct answer
Learning rate
$\Delta W = \eta \times \delta$
neuron output
correction

Fig. 4 Backpropagation algorithm within a multilayer fully-connected network. The network is trained by the backpropagation algorithm to classify images of a specific dataset, such as the MNIST dataset of handwritten digits. Input patterns are forward propagated, then the output results y_j are compared with the correct answer g_j. The errors $\delta_j = y_j - g_j$ are back-propagated to previous layer and used to control the synaptic weight update in the network based on Eq. (2).

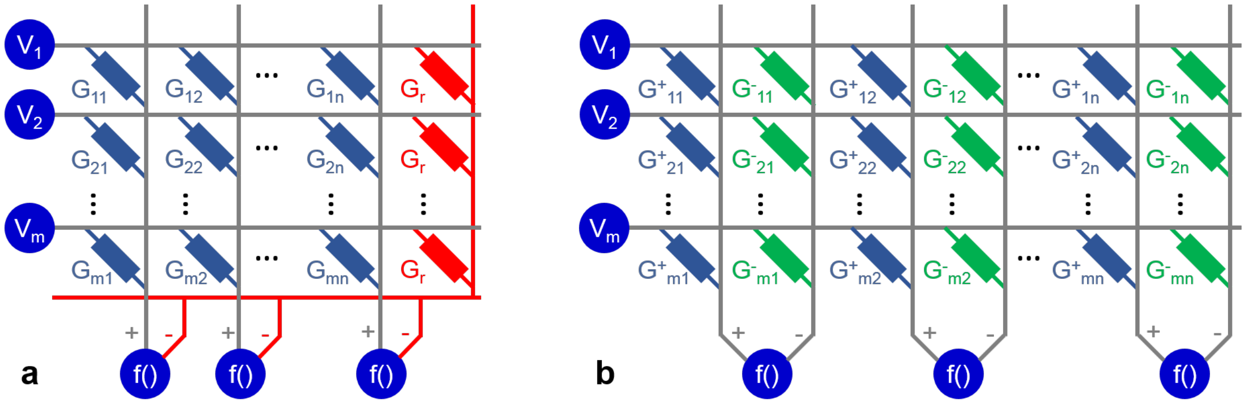

Fig. 5 Crosspoint array implementation of a neuromorphic circuit. (a) Synaptic weight is represented by a differential pair consisting of an array memory G_{ij} and a reference memory G_r., which is preferred when the memory device exhibits analogue behavior in both increase and decrease conductance directions. In this case, updates are performed on the array device, while the other device is shared among array cells in the same row as a reference device. (b) Synaptic weight is represented by a differential pair consisting of two memory devices G^+_{ij} and G^-_{ij}, which is preferred when the memory device exhibits analogue behavior in one direction only.

Fig. 5 shows a typical fully connected network implemented by crosspoint array of non-volatile memories. Here, voltage neuron signals V_i applied at the array rows induce currents I_j given by the Kirchhoff's and Ohm's laws, namely:

$$I_j = \sum_i w_{ij} V_i \tag{3}$$

which thus describes the neural network fundamental property of Eq. (1). Since the synaptic weight w_{ij} in Eq. (3) can have both positive and negative sign, two conductances are generally needed to implement a single weight in hardware. For instance, the conductance in Fig. 5a is obtained as the difference between a tunable conductance G_{ij} and a fixed reference conductance G_r, and the overall synaptic weight is thus given by $w_{ij} = G_{ij} – G_r$. Two tunable memory elements with conductances G_{ij}^+ and G_{ij}^- are instead used in Fig. 5b, thus yielding $w_{ij} = G_{ij}^+ - G_{ij}^-$ [21,90]. The scheme of Fig. 5a is adopted when a single resistive device can be tuned in analogue fashion for both conductance increase and decrease, whereas the scheme of Fig. 5b is instead preferred when the resistive memory device shows analog tuning capability in just one direction, e.g., the conductance increase for PCM [91] and the conductance decrease for filamentary RRAM [92]. Note that the architecture of Fig. 5a is more area efficient, since the reference device can be shared among an entire row. This structure is feasible for devices capable of analogue potentiation and depression, e.g., PCMO-based RRAM [93] and ECRAM memory devices [76]. In both circuits in Fig. 5, PRE neurons forward propagate information from row to column lines, and the POST neurons integrate the aggregate currents and apply a non-linear function *f*.

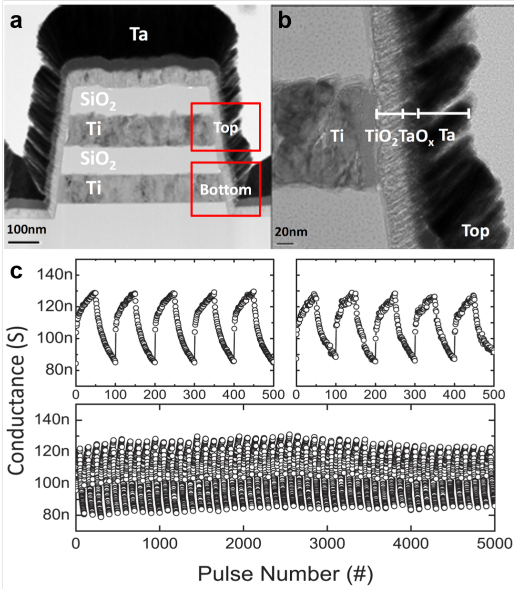

Fig. 6 Vertical RRAM synapse. (a) Structure of the vertical array, where each cell is interposed between stacked horizontal Ti electrodes and vertical Ta electrodes deposited on the sidewall of a trench. (b) RRAM structure, including a TaO_x/TiO_2 stack as switching layer. (c) Characteristics of pulsed potentiation and depression, indicating repeatable conductance window of about 35% with gradual switching. Reprinted with permission from [94,96]. Copyright 2013 IEEE, Copyright 2016 IOP.

An efficient and accurate training can be obtained by adjusting the weight correction Δw_{ij} to be linearly dependent on the product of x and δ, as expressed in Eq. (2). For this purpose, a high degree of linearity is requested to the weight updates of the memory devices. Such highly linear update of the memory conductance can in principle be obtained with any analogue resistive device by a closed-loop scheme, such as a program-verify algorithm to accurately update the device conductance. However, the degradation of speed performance becomes unacceptable, thus steering the programming technique towards an open-loop tuning. In the recent years, there has been a clear progress in improving the linearity of synaptic devices by means of pulse-width modulation or material stack engineering [94]. For instance, Fig. 6 shows an example of a vertical RRAM (a) where the switching layer consists of a stacking of TaO_x and TiO_2, interposed between a horizontal Ti electrode and a vertical Ta top electrode (b) [94]. This type of vertical RRAM is particularly promising in view of the superior scaling with respect to the horizontal stacking of horizontal crosspoint arrays [95]. Fig. 6c shows the corresponding weight update characteristics under open-loop application of pulses with a fixed voltage, indicating that a proper engineering of the RRAM stack enables good linearity and high repeatability of analogue update [96].

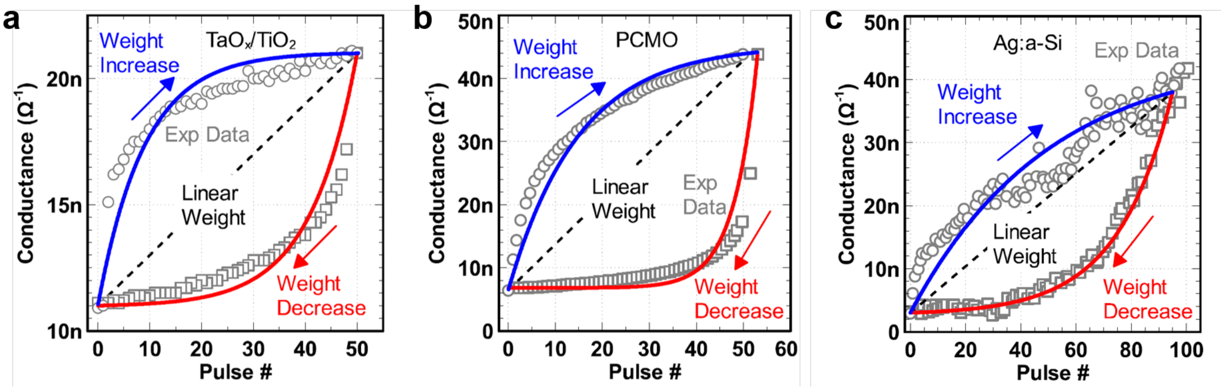

Fig. 7 Synaptic weight update characteristics for (a) TaO_x/TiO_2, (b) $Pr_{1-x}Ca_xMnO_3$, or PCMO, and (c) Ag:a-Si synapses, indicating various degree of linearity, symmetry and window of the conductance change. Pulses of fixed positive/negative voltage were applied during potentiation and depression, respectively. Reprinted with permission from [97]. Copyright 2015 IEEE.

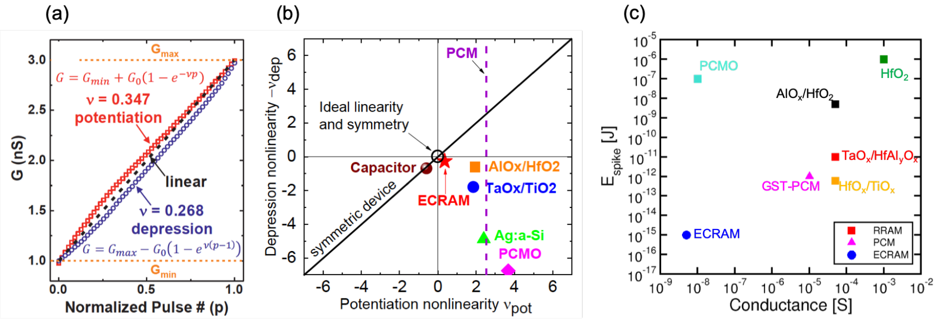

Fig. 8 Characteristics of the ECRAM synapse device. (a) The update characteristics indicate low non-linearity with low coefficient ν. (b) Summary of nonlinearity coefficients comparing ECRAM with various RRAM and PCM. (c) Summary of the energy per spike during potentiation/depression, as a function of the device conductance, the latter providing a metric for the energy consumption during spike transmission. Reprinted with permission from [76]. Copyright 2018 IEEE.

Fig. 7a shows the weight update characteristics for the TaO_x/TiO_2 bilayer stack of Fig. 6, where the depression characteristic is shown with reversed pulse count [94,97]. The weight update characteristics are compared with other synaptic RRAM devices, namely polycrystalline $Pr_{1-x}Ca_xMnO_3$ (PCMO) [93,98], and (c) amorphous Si with an Ag top electrode [99]. Pulses with equal voltage were applied during both potentiation and depression. In general, RRAM synapses show some degree of non-linearity in both potentiation and depression, which might result in inaccuracy for image and speech recognition [93].

The synaptic linearity can be improved by increasing the complexity of the synaptic structure. For instance, it has been shown that a one-transistor/2-resistor (1T2R) structure can improve the update linearity, despite the increased synapse area because of the additional transistor and resistance [100]. Alternatively, the update linearity is generally improved in synaptic transistors as in Fig. 3b. These 3-terminal structures such as Li-based memories [74,76] or organic based memories [75], can achieve a high linearity of potentiation/depression, despite a relatively small window of conductance. Fig. 8a shows the update characteristics for a Li-based ECRAM device, indicating an almost linear behavior for both potentiation and depression across a relative window by a factor 3 [76]. From Fig. 8a, a linearity factor can be defined as the coefficient ν in the potentiation formula $G = G_{min} + G_0(1+\exp(-\nu p))$, where G_{min} is the minimum conductance, G_0 is a reference conductance describing the synaptic window, and p is the normalized number of pulses [101]. Fig. 8b summarizes the linearity factor ν for various synaptic stacks, including TaO_x/TiO_2 RRAM [94], AlO_x/HfO_2 RRAM [102], PCMO [93,98], Ag/a-Si RRAM [99] and capacitor-based structures [103]. PCM synapses are also included in the comparison, although they can provide weight update in one direction only, namely potentiation by gradual crystallization of the phase change material [21]. Among all materials reported in Fig. 8b, ECRAM combines excellent linearity and symmetry of the potentiation/depression update. In addition, the update and read paths can be separated in the device, corresponding to the gate-channel and the source-drain terminals, respectively, thus enabling a good reliability and low energy operation. This is supported by Fig. 8c, showing the energy per spike for weight update as a function of the synapse conductance, which describes the energy during the feedforward operation of the neural network. The figure compares various synaptic technologies including ECRAM [76], $TaO_x/HfAl_yO_x$ RRAM [104], AlO_x/HfO_2 RRAM [102], PCMO [105], HfO_x/TiO_x [106], HfO_2 [107], and PCM [108]. ECRAM devices show a low conductance, enabling low current operation of the neural network in the inference mode, combined with a low energy per spike during weight update, thus offering a promising solution as a synaptic technology with high energy efficiency.

In addition to linearity, symmetry and energy efficiency, another key concerns for synaptic implementations is the granularity of conductance steps and the associated variability. To increase the available number of steps and reduce the stochastic variability, a novel weight structure combining multiple PCM devices in parallel was proposed [108]. The overall weight is obtained by the summation of all the parallel conductance. During training, the weight is updated by programming only one device each time. This implementation increases the overall dynamic range of the weight, since the maximum and minimum achievable weights are now N times larger, where N is the number of parallel memory devices in the synapse. The single conductance step is instead determined by a single device, thus reducing the impact of device variability [108].

Fig. 9a shows an advanced weight structure consisting of a differential pair of PCM devices, with conductances G^+ and G^-, and a third conductance g of a capacitance-controlled transistor. These synaptic conductances have different significance within the overall synaptic weight W, which is given by the formula [109,110]:

$$W = F\cdot(G^+ - G^-) + g - g^{shared}, \qquad (4)$$

where F is a gain factor for the multiplication of the PCM weight. As a result, the PCM devices are used as the most significant pair (MSP), which is amplified by the gain factor F = 3, which thus increases the maximum and minimum implementable weights. The least significant pair (LSP) is instead given by $g\text{-}g^{shared}$, where g^{shared} is a reference conductance shared by all the synaptic elements in the same row. The LSP is implemented by using 3-transistor/1-capacitor (3T1C) CMOS structures. The weight update during training is performed on the LSP by increasing or decreasing the amount of charge on the capacitor. The latter capacitor drives the gate of a MOS transistor, which results in a tunable source conductance. The capacitor is charged by a p-MOS and discharged by an n-MOS, thus the conductance updates have similar amplitudes in either potentiation or depression. This enables the sharing of the reference 3T1C structure with conductance g^{shared} among many weights, thus resulting in a high integration density [109]. Once every thousands of training epochs, the total weight W composed by the MSP and LSP is read and transferred into the PCM MSP, thus preserving the non-volatility of the overall weight. The LSP is then reset, enabling further training without the risk of capacitor saturation. Since the transfer operation is infrequent, the PCM devices can be programmed using a closed-loop procedure which ensures accurate weight programming [109]. The reduced number of PCM programming steps also allows to minimize any possible degradation due to potentiation/depression cycles.

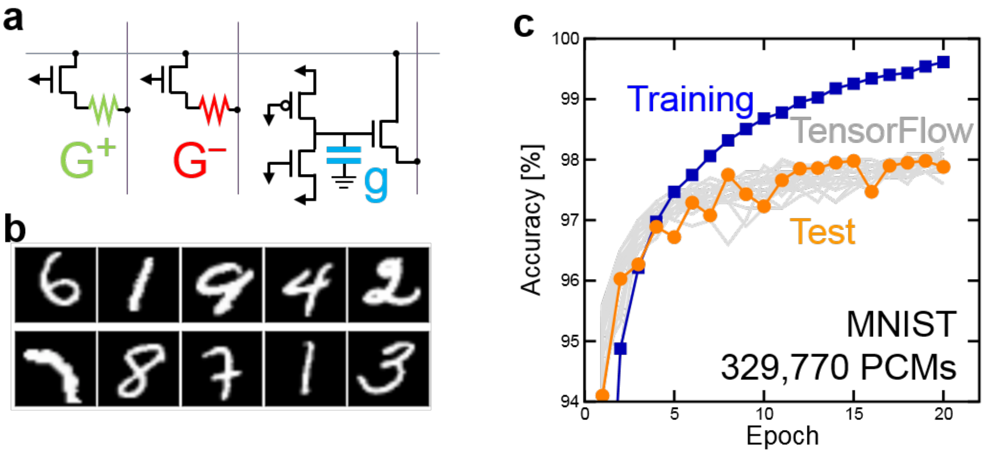

Fig. 9 Advanced synaptic structure to improve linearity. (a) The synaptic weight is implemented using two resistive devices, such as PCMs, and a 3-transistor/1-capacitor (3T1C) structure. The conductance g of the 3T1C structure represents the LSB, which is regularly updated, whereas the PCM represents the MSB, which is updated only once every 1000 cycles. (b) A sample of the MNIST dataset for training and testing the accuracy of the neural network. (c) Accuracy for learning the MNIST dataset, indicating that testing results are equivalent to TensorFlow results. Reprinted with permission from [109]. Copyright 2018 Springer Nature Publishing.

The weight structure in Fig. 9a allows to execute supervised training with conventional small and medium-size datasets such as MNIST (Fig. 9b), MNIST with noise, and CIFAR-10/100, and achieve software-equivalent training accuracy (Fig. 9c) [109]. A similar structure with multiple conductance of varying significance can be implemented with any pairs of non-volatile memories, thus allowing to replace the relatively bulky 3T1C circuits. The introduction of different pairs of conductance diversifies the requests on resistive devices in LSP or MSP. For example, while high linearity and endurance are required for LSP devices, weak retention is acceptable since the transfer operation preserves the overall weight value. On the other hand, devices in MSP should show a good retention, while linearity is not necessary, due to the closed-loop programming procedure [109, 111].

5. Artificial synapses in SNNs

Most practical applications of neural networks have a typical ANN structure as shown in Fig. 4 and adopt a supervised technique for training the network on specific dataset. However, this type of neural network is not similar to the brain, where learning does not take place with a supervised process such as the backpropagation algorithm. Brain-inspired learning typically occurs by unsupervised or semi-supervised processes, such as the spike-timing dependent plasticity (STDP) [29,30,112-116] or spike rate dependent plasticity (SRDP) [117-120]. In these processes, the simultaneous spiking activity at 2 neurons can lead to a potentiation of the synapse connecting them. This can be referred to as a generalized Hebb's learning rule, where 'neurons that fire together wire together', meaning that 2 neurons which are active in response to the same event should be linked by a relatively strong synaptic connection [121-122]. A similar concept is the associative memory, where the simultaneous spiking of two neurons within a recurrent network is awarded by synaptic potentiation to strengthen their respective synaptic link. Note that Hebbian and STDP rules are dictated by time, which is an essential variable describing the information within a SNN, as opposed to the synchronous timing in an ANN.

Fig. 10a illustrates the STDP learning rule, where the synaptic weight of a synapse is dictated by the relationship between the PRE spike time t_{PRE} and the POST spike time t_{POST}. In particular, if the POST fires after some time $\Delta t = t_{POST} - t_{PRE} > 0$ from the PRE spike, then the synaptic weight increases, while if Δt is negative, i.e., the PRE spike follows the POST spike, then the synaptic weight decreases. This is summarized in Fig. 10b, showing the change of synaptic conductance as a function of Δt: long-term potentiation (LTP) with an increase of conductance occurs for $\Delta t > 0$, whereas long-term depression (LTD) is seen for $\Delta t < 0$. For relatively long delays Δt, no synaptic change is observed. Fig. 10c shows experimental characteristics of STDP from the hippocampus, supporting the relevance of STDP as a learning rule in biological networks [29].

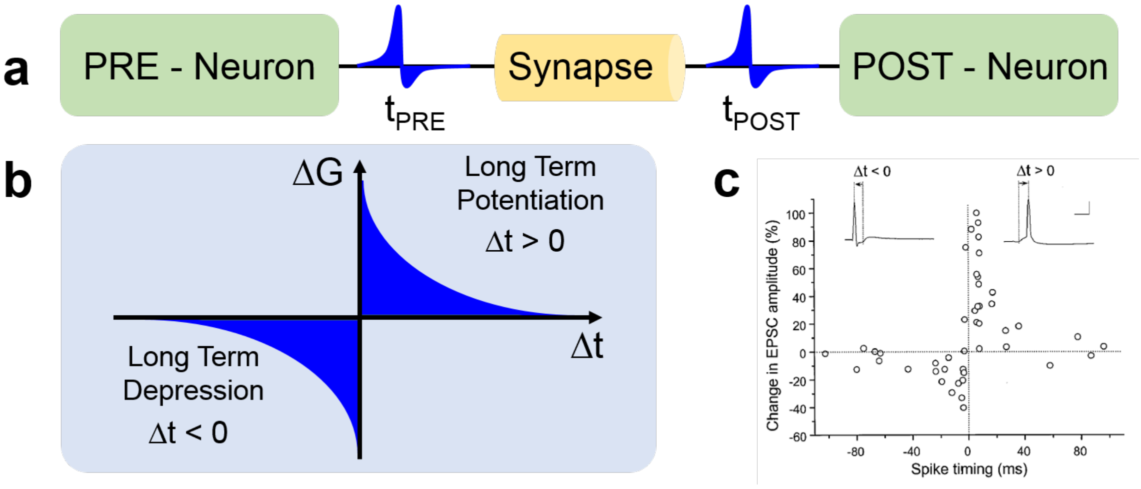

Fig. 10 Spike timing dependent plasticity (STDP) rule. (a) STDP controls the weight update within a synapse based on the timing between the spikes of the presynaptic neuron (PRE) and the postsynaptic neuron (POST). (b) Typical STDP characteristic, where the conductance variation ΔG is positive (long-term potentiation) for positive delay (POST spike following the PRE spike), and it is negative (long-term depression) for negative delay (PRE spike following the POST spike). (c) Data from in vivo biological samples from the hippocampus provide direct confirmation of the STDP. Reprinted with permission from [29]. Copyright 1998 Journal of Neuroscience.

5.1 Overlap STDP

There have been extensive efforts to implement STDP in hardware synapses adopting emerging memories. A synchronous protocol for STDP was proposed according to a time-division multiplexing (TDM) of neuronal signals passing through a synapse, the latter consisting of a bipolar switching memory such as a RRAM [123]. The proposed scheme allows to separate the two major roles of the synapse, namely (i) transmission of neuron spikes with proper weight, and (ii) learning or plasticity, where the weight is adjusted in response to the timing of PRE and POST spikes. The adjustment occurred via overlap of pulses for either LTP or LTD using pulse width modulation (PWM) of spiking signals. For instance, overlapping pulses at the two terminals of the RRAM device can result in a positive voltage drop across the memory device exceeding the threshold for set transition, thus leading to LTP. A similar concept was later introduced to implement STDP in real RRAM devices with Ag as top electrode and amorphous Si as switching layer [99]. In this type of devices, which are generally referred to as conductive bridge random access memory (CBRAM), the CF originated from migrating cations from the top electrode, such as Ag [124-128] or Cu [129-132].

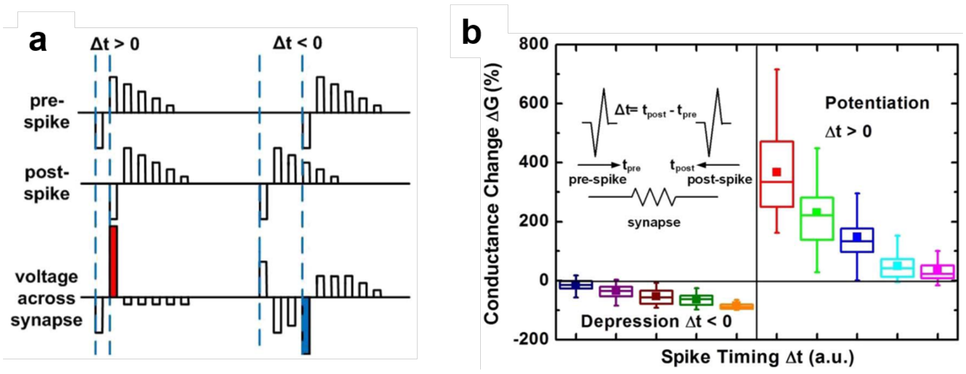

Fig. 11 Scheme for an overlap STDP with a RRAM synapse. (a) Trains of spikes generated by the PRE and the POST result in a potential difference across the synapse with a large positive pulse for Δt > 0, or a large negative pulse for Δt < 0. (b) Measured conductance change as a function of the spike delay Δt, indicating potentiation for Δt >0, and depression for Δt < 0. Reprinted with permission from [133]. Copyright 2011 IEEE.

Fig. 11 shows a more straightforward approach to STDP implemented in a RRAM device with HfO_2 switching layer [133]. Here, the shape of the neuron spikes is tailored to achieve a certain STDP response by overlap of pulses [13]. For instance, Fig. 11a shows that each spike consists of a train of 6 pulses, the first being negative, while the following 5 have decreasing positive amplitudes. Ideally, none of the pulses within each spike alone can result in a change of conductance, as the threshold voltage for set and reset transitions in the RRAM are larger than the spike amplitude. However, the application of 2 partially-overlapping spikes at the 2 terminals of the RRAM device results in a possible overlap exceeding the threshold for set and reset transitions. For instance, a large positive voltage arises from the spike overlap for Δt > 0, thus resulting in a set transition and consequent LTP. On the other hand, a large negative voltage exceeding the reset threshold occurs for Δt < 0, thus leading to a reset transition, or LTD. Fig. 11b shows the measured conductance change ΔG for a HfO_2 RRAM device, indicating LTP and LTD for positive and negative spike delay, respectively [133]. This type of overlap-based STDP has been applied to several device concepts, including metal-oxide RRAM [134-135], PCM [90,136,137], STT-MRAM [138], FeRAM [139], and also Flash memories with a floating gate [140,141] or a nanocrystal storage layer [142]. Both multiple pulse trains or single pulses were assumed for the spike in the STDP [116]. The pulse shape can be adapted to achieve the desired STDP function, where LTP and LTD are obtained for a certain range of Δt [13].

Despite the overlap STDP scheme of Fig. 11 can be applied to a wide variety of synaptic devices capturing various pulse shapes and STDP characteristics, the two-terminal device structure in the figure might have some drawbacks for RRAM synapses. For instance, typical filamentary switching requires a current limitation during set, which is typically achieved by adding a series transistor in a 1T1R structure [143]. The voltage driven set transition in Fig. 11 might thus cause uncontrolled growth of the CF with detrimental impact on synapse reliability. To overcome the limitations of the two-terminal synapse in Fig. 9, the 1T1R synaptic circuits were proposed for overlap STDP [91,92,144]. In a 1T1R structure, the larger number of terminals allows to independently control spike transmission and plasticity, while offering the capability to limit the RRAM current during set/potentiation for improved reliability. The 1T1R array also allows for a better control of the synaptic array, thanks to the select transistor in series with the memory element, either a RRAM [92,144] or a PCM [91].

Fig. 12a shows a RRAM synapse with 1T1R structure [145]. The RRAM synapse has 3 terminals including (i) the transistor gate, connected to the input node of the PRE, (ii) the transistor source, connected to the input node of the POST, and (iii) the top electrode of RRAM device. The latter is connected to the POST for the control of transmission and plasticity phases of the RRAM synapse. The PRE spike of amplitude V_G at the transistor gate induces a synaptic current, as the top electrode voltage V_{TE} is normally biased to a relatively small value below the threshold for set/reset processes [144]. The synaptic current is proportional to the synaptic weight, namely the RRAM conductance, as the gate voltage is large enough to prevent any significant voltage drop across the transistor. The current flows across the transistor and enters the POST input node, which behaves as a current-summing virtual ground. The POST can thus integrate the synaptic current from several synapses, all connected to the same input node, and fire in correspondence of the internal potential reaching a given threshold [145]. At fire, the POST applies a feedback spike to the top electrodes of all synapses activating STDP process. Fig. 12b shows the PRE spike (top) and the POST (feedback) spike (bottom) as a function of time, for the case of a small positive delay Δt between the spikes [145]. The feedback spike includes a positive pulse and a negative pulse, exceeding the threshold for set and reset transition of the RRAM device. For $\Delta t > 0$, there is an overlap between the PRE spike and the positive pulse of the POST spike, thus causing a set transition, or LTP. On the other hand, for $\Delta t < 0$, the overlap occurs at the negative top electrode voltage, which thus induces a reset transition, or LTD. Note that the transistor controls the current during set transition, thus setting a maximum limit to the RRAM conductance after LTP. On the other hand, LTD is self-limited by the highest resistance of the HRS, typically from one to two decades larger than LRS. Fig. 12c shows the STDP characteristics, namely the conductance change $\eta = \log(G/G_0)$, where G is measured before the STDP event and G_0 is measured before the STDP event, as a function of Δt and G_0. The data generally indicate that the synapse is potentiated for $\Delta t > 0$ and depressed for $\Delta t < 0$, although the amount of LTP/LTD depends on the initial conductance. For instance, LTD is inhibited when the initial state is HRS, since the conductance cannot be decreased below a certain minimum amount. Similarly, LTP is inhibited when the device is initiated in the LRS, thanks to the compliance current controlled by the transistor.

The 1T1R synapse can be generalized to other memory elements by changing the feedback spike of Fig. 12b. For instance, the feedback spike includes 2 positive pulses in the 1T1R PCM synapse, the first being a relatively small set pulse below the melting voltage, whereas the second is above melting voltage to induce a reset process [91]. A possible limitation of the STDP characteristics in Fig. 12c is the rectangular shape, instead of the exponential dependence on Δt [13]. For an improved control of the STDP characteristics, the two-transistor/1-resistor (2T1R) structure was proposed for RRAM [146] or PCM [147]. In the 2T1R synapse, the two gates activate the memory element for either transmission or plasticity, while the top electrode node is connected to the PRE. Applying a time-dependent gate voltage at the plasticity transistor from the POST, one can freely change the set transition current at the device, thus shaping the time dependence of the LTP. The time dependence of the LTD can be instead dictated by the top electrode voltage from the POST [146]. The 1T1R synapse of Fig. 12 was also adapted to SRDP, with the addition of few transistors resulting in a four-transistor/one-resistor (4T1R) structure [148,149].

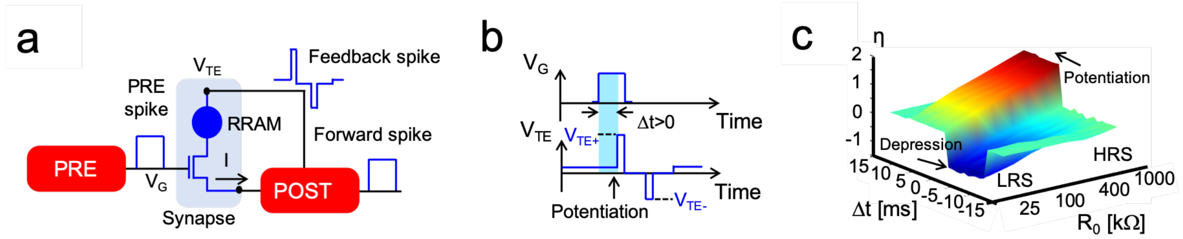

Fig. 12 Scheme for an overlap STDP with a 1T1R structure. (a) Circuit scheme for the 1T1R synapse, where the PRE spike is applied to the transistor gate, while the POST spike is applied to the top electrode. The spike current is activated by the PRE spike and fed into the POST through the bottom electrode. As the integrated PRE spikes lead to fire, the POST spike is applied to the TE, thus inducing potentiation or depression depending on the spike delay Δt. (b) PRE and POST spike for Δt > 0, where the resulting overlap pulse is positive, thus inducing set transition, or potentiation. (c) Measured conductance ratio $\eta = \log(G/G_0)$ as a function of initial resistance R_0 and spike delay Δt, indicating potentiation for Δt >0, and depression for Δt < 0. Reprinted from [145].

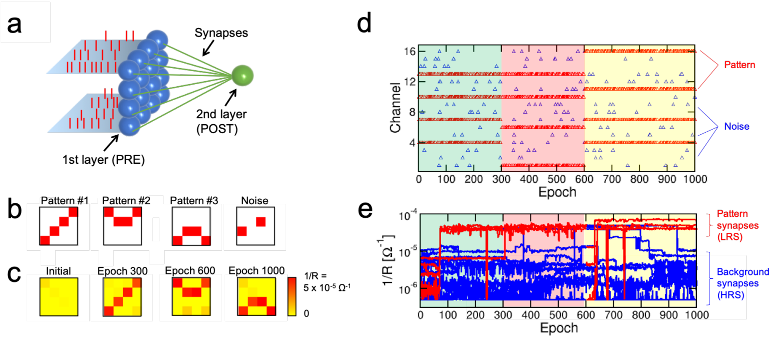

Fig. 13 Unsupervised learning in a one-layer perceptron with 1T1R synapses. (a) Illustration of the one-layer perceptron, where a pattern is presented at the input neuron layer (PRE) via spikes, while the single POST integrates the spiking currents across each synapse. (b) Examples of 4x4 visual patterns and typical noise presented at the PRE layer. (c) Initial and final synaptic weights after sequential presentation of three patterns. (d) Summary of the presented patterns during three sequential phases. (e) Time evolution of the synaptic weights for pattern synapses, i.e., those belonging to the pattern, and background synapses, i.e., those not belonging to the pattern. Pattern and background synapses display potentiation and depression thanks to unsupervised learning. Reprinted from [145].

STDP rule is an enabling algorithm for unsupervised learning patterns in space [91,144,145,148, 150-155] and space/time [156]. Fig. 13a shows a prototypical neural network, namely a one-layer perceptron with 16 PREs and one POST, to test the ability to learn via STDP [145]. This network was implemented in hardware by using RRAM synapses with 1T1R structure, as the one described in Fig. 12. PRE spiking signals were applied to the transistor gate electrodes, while all synaptic currents were summed in real time and integrated by a microcontroller Arduino Due to serve as the POST. At fire, POST feedback spikes were applied to all the synaptic array, thus resulting in LTP and LTD, depending on the relatively delay between PRE and POST spikes according to STDP. Fig. 13b shows the applied spikes, consisting of a space 4x4 pattern which was changed in time during three sequential phases of training of the network. Unsupervised learning is inherently stochastic, meaning that noise patterns must also be submitted randomly to the network to induce LTD as a result of uncorrelated PRE noise spikes following the POST fire. Fig. 13c shows the synaptic conductance before the training session, and after each of the three training phases, always indicating the correct learning of the input pattern with no remaining traces of the previous pattern. The ability to 'forget' the pattern from the previous phases can be attributed to the LTD at the background positions within submitted pattern, where only random noise is presented, thus inducing LTD [145]. Fig. 13d summarizes the submitted spikes, including both pattern and noise spikes, and Fig. 13e shows the time evolution of all the measured synaptic weights. The ability for unsupervised learning can be extended to larger networks, e.g. to enable learning of realistic input patterns such as the MNIST data set [91,157,158]. The STDP rule in 1T1R synaptic RRAM was also shown to support spatio-temporal learning [156] and associative memory behavior within a recurrent network [159].

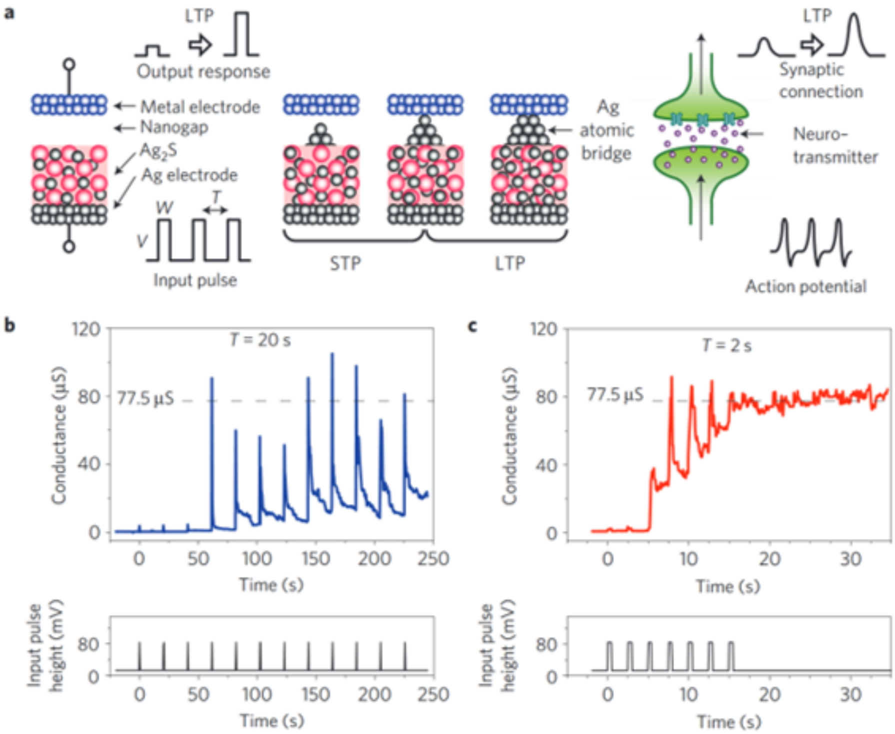

Fig. 14 Scheme for a non-overlap synapse with a RRAM synapse. (a) Schematic of the atom switch RRAM consisting of a conductive bridge RRAM where the applied pulse induces migration of Ag cation. The pulse frequency dictates the transition from STP to LTP, similar to the biological synapse. (b) Measured conductance (top) in response to a low-frequency stimulation (bottom) indicating STP with fast decay after each pulse. (c) Measured conductance (top) in response to a high-frequency stimulation (bottom) indicating LTP reaching permanent increase of conductance. Reprinted with permission from [161]. Copyright 2011 AAAS.

5.2 Non-overlap STDP

The STDP learning rule by overlapping spikes is generally robust and reliable, however it also has drawbacks in terms of energy consumption and pulse-width. In fact, the PRE spike pulse-width must be at least the same as the scale of delay time of the STDP characteristics in Fig. 12, because there is no memory effect taking place in the RRAM element after the end of the PRE spike. The long spike pulse-width results in a large energy consumption during the spike, as well as an excessive occupation of interconnect lines in address-event-representation (AER) architectures which are typical for multicore SNNs [160].

To overcome the limitations of the overlap learning scheme and enable the development of non-overlap neuromorphic devices, the inherent memory properties of emerging memories can be considered. Fig. 14a shows a schematic of an atom-switch RRAM, consisting of a stack of a metal top electrode, an Ag bottom electrode, and an Ag_2S film as solid-state electrolyte [161]. Differently from the CBRAM device, a vacuum nanogap is interposed between the top electrode and the solid electrolyte. The application of voltage pulses to the top electrode can lead to the formation of an Ag atomic bridge across the nanogap [162]. As shown in Fig. 12a, the atomic switch can act as an artificial synapse since repeated pulses can lead to the gradual formation of an atomic bridge, according to a sequence of a first phase of short-term potentiation (STP), followed by a second phase of LTP [161]. This dynamic behavior can also lead to SRDP, where STP or LTP behaviors are observed depending on the spiking frequency. In particular, the application of spikes with relatively low frequency results in STP, where the RRAM conductance increases temporarily, then decays within the period before the next spike is applied (Fig. 14b). On the other hand, the application of spikes at higher frequency induces LTP, as each spikes contributes an additional increase of conductance reaching a maximum value of about 80 µS after just 5 spikes (Fig. 14c).

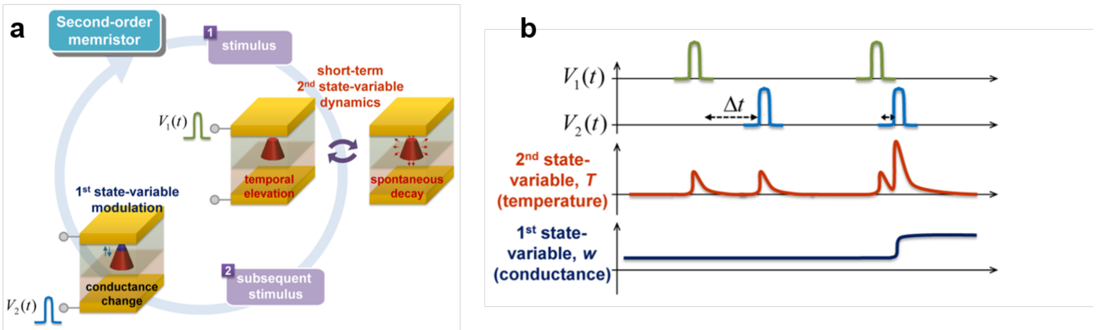

Fig. 15 Non-overlap synapse based on second-order memristor. (a) Schematic of the second-order memristor where the application of a first stimulus causes a temporary increase of conductance, followed by a spontaneous decay. The application of a subsequent stimulus can induce a permanent conductance change. (b) Illustration of the second order memristor dynamics, where the application of a sequence of two spikes can lead to potentiation only if the spike delay Δt is sufficiently short as a result of the second-order memristor dynamics. Reprinted with permission from [163]. Copyright 2015 ACS.

Similar dynamic effects have been reported in RRAM devices. Fig. 15a schematically illustrates the concept of second-order memristor, namely a RRAM synapse which can learn depending not only on the pulse amplitude, but also on the time delay between two successive spikes [163]. Namely, if the time delay Δt between two pulses with relatively low amplitude is long, no conductance change is observed, whereas the conductance increases if the second pulse is applied at short time delay from the first one. This second-order dynamic behavior has been explained by the role of the local temperature within the device: when the second pulse is applied at short delay after the first one, the device temperature is still relatively high, due to the limited thermal time constant which was estimated around 500 ns [163]. The local high temperature thus assists the ionic migration at the origin of the conductance change, which could lead to non-overlap STDP and SRDP process in a Pd/Ta2O_{5-x}/TaO_2/Pd RRAM synapse [163].

Fig. 16a shows a similar concept for a non-overlap synapse based on a one-selector/one-resistor (1S1R) structure, where the resistor device is a nonvolatile RRAM, while the selector device is a volatile RRAM [164]. The volatile RRAM, also referred to as diffusive memristor, consists of a stack of Ag top electrode and an oxide solid electrolyte, such as SiO_x [42], SiON [164] and HfO_x [165]. In these devices, similar to the atomic switch in Fig. 14, the CF formed by the application of a positive voltage pulse to the top electrode spontaneously disconnects after a characteristic retention time in a variable timescale from few ns [165] to the µs or ms range [166]. This volatile behavior has been explained in terms of the impact of the mechanical stress surrounding the CF [128] and the surface tension aiming at minimizing the surface-to-volume ratio of the metallic filament [164,167].

The combination of volatile and nonvolatile RRAM devices in the synapse of Fig. 16a enables SRDP as summarized by the weight change Δw as a function of time separation between spikes t_{zero} in Fig. 16b. The dependence on spiking frequency can be explained by the competition between synaptic potentiation induced by the electric field during the spike, and synaptic relaxation induced by surface diffusion during the waiting time t_{zero} [168]. Similarly, a non-overlap concept for STDP was demonstrated, as summarized in Fig. 16c and d [164]. These and other artificial synapses based on the rich physics of emerging memory devices might spur the development of novel neuromorphic systems which can parallel the energy efficiency and parallelism of the human brain.

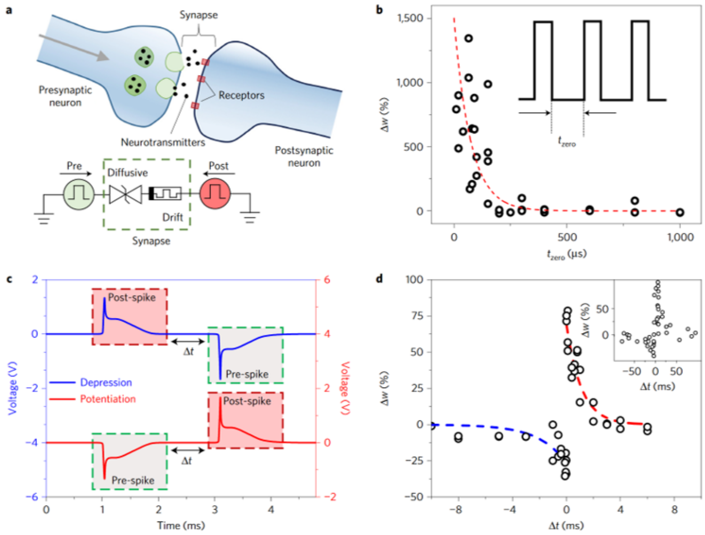

Fig. 16 Non-overlap synapse based on volatile RRAM, also called diffusive memristor. (a) Schematic of the 1S1R synapse, consisting of a volatile RRAM (1S) and a nonvolatile RRAM (1R), where PRE and POST spikes are applied, similar to a biological synapse with neurotransmitters and receptors. (b) Conductance change Δw as a function of the spike delay for a train of spikes. Potentiation decreases with increasing delay between spikes, thus mimicking the SRDP of biological synapses. (c) Illustration of the sequence of POST spike followed by PRE spike (top), or the sequence of PRE spike followed by POST spike (bottom), with a delay Δt between the spikes. (d) Conductance change as a function of the delay between PRE and POST spike, showing STDP dynamics similar to the biological synapse. Reprinted with permission from [164]. Copyright 2016 Springer Nature Publishing.

6. Neuron circuits

The synapse is generally assumed to serve as passive element to store the weight, while the neuron plays the role of active element executing current summation, integration and threshold firing. Due to the large area occupation and power consumption that CMOS circuits require for neuron implementation, it is desirable to follow a different approach where portions of neurons functionalities could be realized with compact and low-power resistive devices. As an example, the integration of spikes in a spike-based network, which is typically performed with large capacitors receiving current contributions, has been recently demonstrated via emerging memory devices [169]. A similar type of accumulating neuron has been proposed by using ReRAM devices based on PCMO [170] and Ag/SiON stack [171].

Fig. 17 shows a possible implementation of an integrating neuron using a PCM device [169]. While conventional CMOS integrators rely on capacitive elements to accumulate charge induced by voltage or current spikes, this type of memory-based neuron integrates the voltage spikes by inducing a partial crystallization within the amorphous volume, as shown in Fig. 17a. Here, each voltage pulse leads to a local Joule heating within the memory device, thus resulting in an increasing crystallization by nucleation and/or growth of the crystalline phase. Since the crystalline phase has a lower resistivity than the amorphous phase, the PCM conductance increases at each step, thus serving as an internal membrane potential of the biological neuron. When the conductance reaches a threshold, a fire event is triggered, corresponding to an output voltage spike generated by the neuron circuit. In addition, the PCM device is reset back to the amorphous phase, as shown in Fig. 17b, thus allowing to operate the spike integration in the next cycle. The modulation of the voltage pulse in terms of voltage amplitude and pulse-width leads then to different and controllable firing rates.

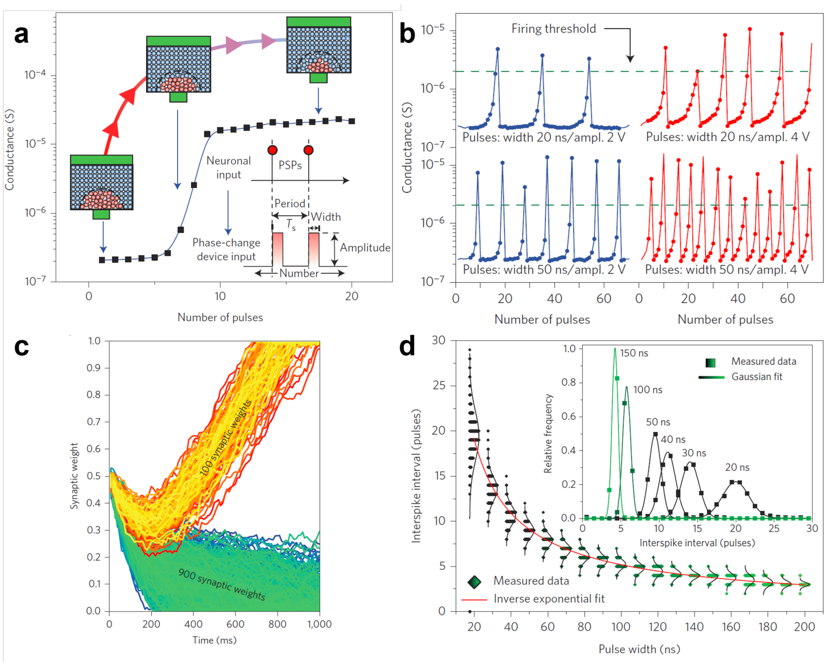

Fig. 17 Phase change neuron. (a) PCM conductance as a function of the number of applied spikes, indicating a gradual increase due to the growth of the crystalline phase within the active region of the phase change material. (b) Conductance change as a function of the number of spikes, or various pulse width (top: 20 ns, bottom: 50 ns) and amplitude (left: 2 V, right: 4 V). The conductance increases until it exceeds a certain threshold, then the low conductance is restored by a reset operation. (c) Synaptic weight for two distinct sets of PCMs, corresponding to active synapses where the applied spikes induce potentiation, or inactive spikes leading to depression. (d) Number of spikes to reach the threshold as a function of the pulse width. Reprinted with permission from [169]. Copyright 2016 Springer Nature Publishing.

Note that the PCM device is only suitable for integration, while an additional CMOS circuit is needed to provide the spike generation and the optional refractory period. Based on this concept, the authors demonstrated a single layer perceptron specialization with STDP algorithm and one PCM neuron. This is illustrated in Fig. 17c, showing the conductance of two distinct families of synapses, corresponding to synapses being active within the pattern, thus receiving potentiation, and synapses being non-active because not involved in the pattern channels, thus receiving depression [169]. Another crucial feature for brain-inspired artificial neurons is the intrinsic stochasticity of integration and fire. To this purpose, the PCM firing pulses were shown to feature a natural stochastic behavior, as every reset operation results in a statistically different atomic structure within the amorphous phase, thus causing variable conductance paths and crystallization times during integration, hence random neuron fire events. Fig. 17d reports the distribution of times between successive fire events as a function of the input voltage pulse-width, which supports the ability to control the amount of stochasticity by changing the programming conditions [169].

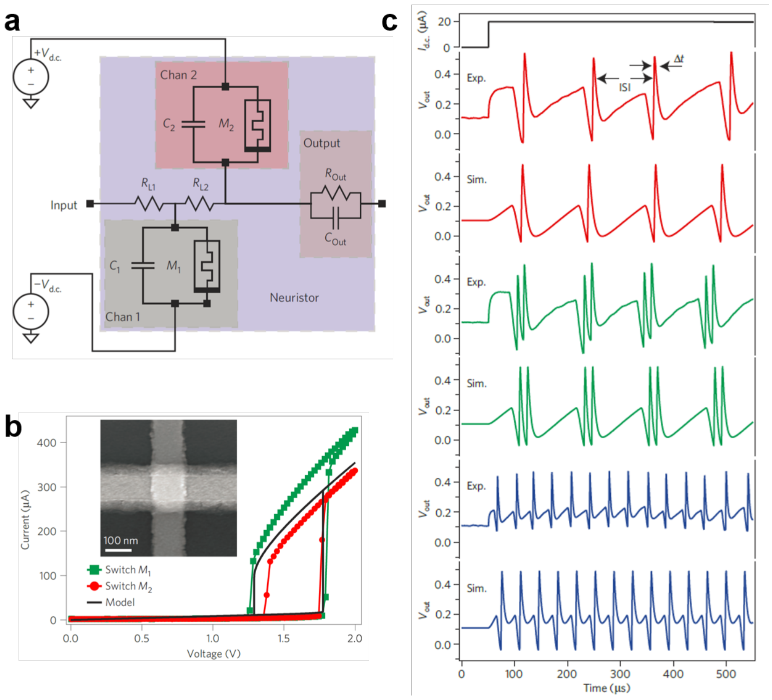

Fig. 18 Mott insulator neuron. (a) Schematic of the neuron circuit including two coupled relaxation oscillators based on NbOx RRAM. (b) I-V characteristics of the NbOx RRAM with crosspoint structure (inset). The curve indicates threshold switching from the off state to the on state, followed by a return back to the off state at the characteristic holding voltage of about 1.3 V. (c) Oscillating dynamics of the neuron. As the DC bias voltage increases, the fire rate increases as a result of the relaxation oscillator dynamics. Reprinted with permission from [173]. Copyright 2013 Springer Nature Publishing.

Many learning schemes in SNNs, such as STDP, require neurons with leaky integrate-and-fire functionality, to elaborate the incoming spike and provide signals to the next neuron layers. Other computing schemes instead rely on neurons emitting a train of spikes with controllable rate. According to the neuron model of Hodgkin and Huxley [172], the cell membrane releases an action potential with a specific shape and a refractory period. Volatile memory devices enable the implementation of such neuron by their switching and recovery processes [173]. Fig. 18a shows an example of neuron circuit implemented with a couple of RRAM devices based on NbO_2, described by the I-V curves in Fig. 18b [173]. NbO_2 -based RRAM devices are characterized by a typical threshold switching behavior, where the device shows a transition from the off-state to the on-state, followed by a transition back to the off-state [174-176]. Combining one or more threshold switches with parallel capacitor as in the circuit of Fig. 18a thus results in controlled relaxation oscillations which can be adopted to generate spike trains with various shape. The neuron circuit of Fig. 18a is in fact able to generate a time-oscillating dynamics thanks to a constant input current. Fig. 18c provides a comparison of experimental and simulated results, where a neuron biased at a constant current I = 20 μA delivers spike trains with tunable inter-spike intervals depending on the capacitance values C_1 and C_2 [173].

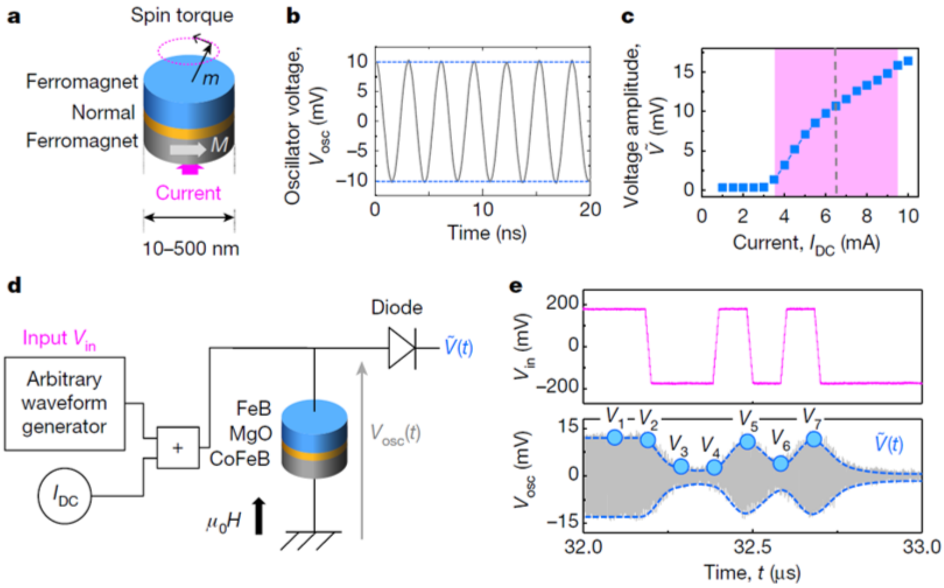

Fig. 19 Magneto-resistive neuron. (a) Schematic of a superparamagnetic junction, where a bias current I_{DC} induces an oscillating dynamics due to magnetic precession. (b) Oscillating output voltage as a result of the resistance change due to the MTJ effect. (c) Amplitude of the voltage oscillation as a function of I_{DC}. (d) Neuron circuit combining a DC bias and an external stimulation. (e) Typical input/output voltage, where a constant voltage input results in an oscillating output voltage with variable amplitude. Reprinted with permission from [177]. Copyright 2017 Springer Nature Publishing.

Although the NbO_x oscillating neuron shows excellent agreement with biological neurons, the need for integrated capacitors prevents easy scalability toward very high density. For this reason, achieving a neuron with oscillating functionalities inherently in a single device is desirable. Fig. 19a shows an example of a single nano-device oscillator [177]. The device structure consists of two ferromagnetic layers separated by a non-ferromagnetic material, such as a MgO tunneling barrier. The applied current flows across the bottom layer which shows a fixed in-plane magnetization. The bottom layer causes spin-polarization of the flowing electrons, which eventually cause the rotation of the magnetic polarization of the top ferromagnetic layer. The rotation of the magnetization leads to a variation in the device electrical resistance due to tunneling magneto-resistance (TMR) effect [60], with a consequent oscillation of the voltage drop across the device, as shown in Fig. 19b. Fig. 19c shows the amplitude of the voltage oscillations as a function of the bias current, indicating that the oscillation amplitude increases with the applied current [177]. This nanomagnet oscillator thus provides a remarkable scheme for controllable spiking neurons with extremely small area and high density.

Based on this principle, the input information can be sent to the oscillator by superposing a signal V_{IN} to a constant current I_{DC}. The input signal thus causes a modulation of the oscillation amplitude, as shown in Figs. 19d and e. Interestingly, the envelope curve $\tilde{V}(t)$ in Fig. 19e shows a relaxation time on the scale of tenths of nanoseconds providing a memory effect which can be used to perform computation [177]. In fact, the state of downstream neurons in a trained neural network depends on the state of neurons in the previous layer. Similarly, the same concept can be translated in time, where the state of a neuron in the future depends on the state of the same neuron in the past [177]. Based on this concept, a neural network adopting a single oscillating nanomagnet neuron allows to recognize audio waveforms of single spoken digits pronounced by five female speakers [177]. A similar concept of oscillating nanodevice was proposed by harnessing random telegraph signal within a single RRAM device [178]. These results support memory devices as a very promising approach for small-area neuron circuits, where computation takes place via the inherent device physics.

7. Conclusions and outlook

Neuromorphic circuits are gaining momentum as an alternative computing paradigm complementing the conventional digital circuits at the end of Moore's law. By taking inspiration from the architecture and processing scheme in the brain, the neuromorphic circuits offer the possibility to perform advanced tasks, such as recognizing objects in an image frame, or translating a speech to a different language. In carrying out such important tasks, however, it is essential that the minimum energy is consumed in the least computing time within the smallest possible circuit. In this scenario, memory devices show a very attractive wealth of physical processes, combined with an extreme scalability thanks to material switching, which makes this class of device the most promising technology for future neuromorphic circuits.

Despite the very strong opportunities of memory devices in neuromorphic computing, there are indeed several steps to be taken to fully support a memory-centric neuromorphic technology. First, it clearly appears that DNNs and SNNs rely on different computing concepts, thus the synapse/neuron devices should have correspondingly different properties. For instance, it appears that analogue or multilevel switching is essential for DNN synapses, whereas SNNs might rely on stochastic, binary synapses. Also, SNNs make large use of integrate-and-fire neurons, whereas a nonlinear amplifier is needed as neuronal element in DNNs. In addition, artificial synapses and neurons are usually implemented with different type of nanodevices, e.g., a nanomagnetic spiking neuron and a hybrid CMOS-RRAM synapse. Finally, 3D integration technology might be needed to achieve the necessary neuron/synapse density to reproduce a brain-like cognitive function. As a result, a full neuromorphic hardware might result in a complicated system combining various types of devices with 3D process integration.

The stochastic variation and lack of repeatability might still be an issue toward the development of DNNs, where the recognition accuracy is subject to a careful, extensive training of the synaptic weight. Reliability issues such as the resistance drift in PCM [179] or the stochastic fluctuation of resistive states in RRAM [180] might represent serious threats toward achieving a high recognition accuracy matching the software implementations of the neural network. A solution to these issues might require careful engineering of the device from the materials, geometry, and algorithm points of view, which are all subject to a better understanding of the switching mechanisms and the associated reliability issues. A deeper insight into the switching processes in the memory devices might also set the stage for improved numerical and compact models for the design of neuromorphic circuits and a better device-circuit codesign of the integrated system.

In addition to device-specific challenges, understanding the scope of neuromorphic hardware is still a general point of debate. DNNs find natural applications in image and speech recognition, where the development of neuromorphic devices enabling a higher density, a faster training and a lower energy consumption will further spur the transition of AI from cloud computing to edge computing, e.g., the ability to recognize and translate a speech on a smartphone. On the other hand, the potential applications for brain-inspired SNNs are not clear yet. While unsupervised learning and spiking operation are clearly attractive features for a neuromorphic system aiming at replicating the human brain, there has not been any compelling evidence that a SNN can even barely get close to the capability of a full cognitive system. A strong barrier toward achieving a brain-like neuromorphic computer on a chip is the lack of our understanding about how information is processed within the human brain. Solving this mystery would spur the development of SNNs and drive the demand for neuromorphic devices, where memory devices might offer unprecedented advantages in terms of scaling and functionality thanks to their unique physical properties.

8. Acknowledgments

We acknowledge Valerio Milo for critical reading of the manuscript. This article has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 648635).

9. References

[1] Moore G E 1965 *Electronics* **38** 114-7
[2] Robertson J 2004 *Eur. Phys. J. Appl. Phys.* **28** 265-91
[3] Kuhn K J 2012 *IEEE Trans. Electron Devices* **59** 1813-28
[4] Theis T N and Solomon P M 2010 *Proc. IEEE* **98** 2005-14
[5] Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N and Lloyd S 2017 *Nature* **549** 195-202
[6] Maass W 2014 *Proc. IEEE* **102** 860-80
[7] Guo N, Huang Y, Mai T, Patil S, Cao C, Seok M, Sethumadhavan S and Tsividis Y 2016 *IEEE J. Solid-State Circuits* **51** 1514-24
[8] Mead C 1990 *Proc. IEEE* **78** 1629-36
[9] Liu S-C, Kramer J, Indiveri G, Delbruck T and Douglas R 2002 *Analog VLSI: Circuits and Principles* (Cambridge, MA, USA: MIT Press)
[10] Chicca E, Stefanini F, Bartolozzi C and Indiveri G 2014 *Proc. IEEE* **102** 1367-88
[11] Qiao N, Mostafa H, Corradi F, Osswald M, Stefanini F, Sumislawska D and Indiveri G 2015 *Front. Neurosci.* **9** 141
[12] Indiveri G, Linares-Barranco B, Legenstein R, Deligeorgis G and Prodromakis T 2013 *Nanotechnology* **24** 384010
[13] Zamarreño-Ramos C, Camuñas-Mesa L A, Pérez-Carrasco J A, Masquelier T, Serrano-Gotarredona T and Linares-Barranco B 2011 *Front. Neurosci.* **5** 26
[14] Ielmini D and Wong H-S P 2018 *Nat. Electron.* **1** 333-43
[15] Merolla P A *et al* 2014 *Science* **345** 668-73
[16] Indiveri G and Liu S-C 2015 *Proc. IEEE* **103** 1379-97
[17] Kau D *et al* 2009 *IEEE IEDM Tech. Dig.* 617-20
[18] Liu T-Y *et al* 2013 *IEEE ISSCC Tech. Dig.* 210-1
[19] Arnaud F *et al* 2018 *IEEE IEDM Tech. Dig.* 424-7
[20] Fackenthal R *et al* 2014 *IEEE ISSCC Tech. Dig.* 338-9
[21] Burr G W *et al* 2014 *IEEE IEDM Tech. Dig.* 697-700
[22] McCulloch W S and Pitts W A 1943 *Bulletin of Mathematical Biophysics* **5** 115-33
[23] Rosenblatt F 1957 *Report 85-460-1*, Cornell Aeronautical Laboratory, Buffalo, New York.
[24] LeCun Y 1985 *Proceedings of Cognitiva* **85** 599-604
[25] LeCun Y, Bengio Y and Hinton G 2015 *Nature* **521** 436-44
[26] Taigman Y, Yang M, Ranzato M and Wolf L 2014 *IEEE Conference on Computer Vision and Pattern Recognition* 1701-8
[27] Mnih V *et al* 2015 *Nature* **518** 529-33
[28] Silver D *et al* 2016 *Nature* **529** 484-9
[29] Bi G-Q and Poo M-M 1998 *J. Neurosci.* **18** 10464-72
[30] Markram H, Lübke J, Frotscher M and Sakmann B 1997 *Science* **275** 213-5
[31] Chicca E, Badoni D, Dante V, D'Andreagiovanni M, Salina G, Carota L, Fusi S and Del Giudice P 2003 *IEEE Trans. Neural Netw.* **14** 1297-307
[32] Momodomi M, Itoh Y, Shirota R, Iwata Y, Nakayama R, Kirisawa R, Tanaka T, Aritome S, Endoh T, Ohuchi K and Masuoka F 1989 *IEEE J. Solid-State Circuits* **24** 1238-43
[33] Monzio Compagnoni C, Goda A, Spinelli A S, Feeley P, Lacaita A L and Visconti A 2017 *Proc. IEEE* **105** 1609-33
[34] Waser R and Aono M 2007 *Nat. Mater.* **6** 833-40

[35] Wong H-S P, Lee H-Y, Yu S, Chen Y-S, Wu Y, Chen P-S, Lee B, Chen F T and Tsai M-J 2012 *Proc. IEEE* **100** 1951-70
[36] Ielmini D 2016 *Semicond. Sci. Technol.* **31** 063002
[37] Russo U, Ielmini D, Cagli C and Lacaita A L 2009 *IEEE Trans. Electron Devices* **56** 186-92
[38] Ielmini D, Bruchhaus R and Waser R 2011 *Phase Transition* **84** 570-602
[39] Larentis S, Nardi F, Balatti S, Gilmer D C and Ielmini D 2012 *IEEE Trans. Electron Devices* **59** 2468-75
[40] Lee H Y, Chen P S, Wu T Y, Chen Y S, Wang C C, Tzeng P J, Lin C H, Chen F, Lien C H and Tsai M-J 2008 *IEEE IEDM Tech. Dig.* 297-300
[41] Lee M-J, Lee C B, Lee D, Lee S R, Chang M, Hur J H, Kim Y-B, Kim C-J, Seo D H, Seo S, Chung U-I, Yoo I-K and Kim K 2011 *Nat. Mater.* **10** 625-30
[42] Bricalli A, Ambrosi E, Laudato M, Maestro M, Rodriguez R and Ielmini D 2018 *IEEE Trans. Electron Devices* **65** 115-21
[43] Balatti S, Larentis S, Gilmer D and Ielmini D 2013 *Adv. Mater.* **25** 1474-8
[44] Prakash A, Park J, Song J, Woo J, Cha E-J and Hwang H 2015 *IEEE Electron Device Lett.* **36** 32-4
[45] Govoreanu B *et al* 2011 *IEEE IEDM Tech. Dig.* 729-32
[46] Baek I G *et al* 2011 *IEEE IEDM Tech. Dig.* 737-40
[47] Wuttig M and Yamada N 2007 *Nat. Mater.* **6** 824-32
[48] Raoux S, Welnic W and Ielmini D 2010 *Chem. Rev.* **110** 240-67
[49] Raoux S, Ielmini D, Wuttig M and Karpov I V 2012 *MRS Bull.* **37** 118-23
[50] Yamada N, Ohno E, Nishiuchi K, Akahira N and Takao M 1991 *J. Appl. Phys.* **69** 284
[51] Kato T and Tanaka K 2005 *Jpn. J. Appl. Phys.* **44** 7340-44
[52] Kim J-J, Kobayashi K, Ikenaga E, Kobata M, Ueda S, Matsunaga T, Kifune K, Kojima R and Yamada N 2007 *Phys. Rev. B* **76** 115124
[53] Ielmini D, Lacaita A L, Pirovano A, Pellizzer F and Bez R 2004 *IEEE Electron Device Lett.* **25** 507-9
[54] Ciocchini N, Laudato M, Boniardi M, Varesi E, Fantini P, Lacaita A L and Ielmini D 2016 *Sci. Rep.* **6** 29162
[55] Chen Y C, *et al* 2006 *IEEE IEDM Tech. Dig.* 1-4
[56] Morikawa T *et al* 2007 *IEEE IEDM Tech. Dig.* 307-10
[57] Cheng H Y *et al* 2012 *IEEE IEDM Tech. Dig.* 725-8
[58] Zuliani P *et al* 2013 *IEEE Trans. Electron Devices* **60** 4020-6
[59] Nirschl T *et al* 2007 *IEEE IEDM Tech. Dig.* 461-4
[60] Chappert C, Fert A and Van Dau F N 2007 *Nat. Mater.* **6** 813-23
[61] Dieny B *et al* 2010 *Int. J. Nanotechnol.* **7** 591-614
[62] Wang D, Nordman C, Daughton, J M, Qian Z and Fink J 2004 *IEEE Trans. Magnetics* **40** 2269-71
[63] Slonczewski J 1996 *J. Magn. Magn. Mater.* **159** L1-7
[64] Kent A D and Worledge D C 2015 *Nat. Nanotechnol.* **10** 187-91
[65] Carboni R, Ambrogio S, Chen W, Siddik M, Harms J, Lyle A, Kula W, Sandhu G and Ielmini D 2018 *IEEE Trans. Electron Devices* **65** 2470-8
[66] Mikolajick T, Dehm C, Hartner W, Kasko I, Kastner M J, Nagel N, Moert M and Mazure C 2001 *Microelectron. Reliab.* **41** 947-50
[67] Takashima D *et al* 2001 *IEEE J. Solid-State Circuits* **36** 1713-20
[68] Sakai S, Takahashi M, Takeuchi K, Li Q H, Horiuchi T, Wang S, Yun K Y, Takamiya M and Sakurai T 2008 *IEEE Non-Volatile Semiconductor Memory Workshop (NVSMW)* 103-4
[69] Böscke T S *et al* 2011 *IEEE IEDM Tech. Dig.* 547-50
[70] Mulaosmanovic H, Ocker J, Müller S, Noack M, Müller J, Polakowski P, Mikolajick T and Slesazeck S 2017 *Symp. VLSI Tech. Dig.* 176-7
[71] Takahashi M and Sakai S 2005 *Jpn. J. Appl. Phys.* **44** 25 L800-2

[72] Florent K *et al* 2018 *IEEE IEDM Tech. Dig.* 43-6
[73] Trentzsch M *et al* 2016 *IEEE IEDM Tech. Dig.* 294-7
[74] Fuller E J, El Gabaly F, Léonard F, Agarwal S, Plimpton S J, Jacobs-Gedrim R B, James C D, Marinella M J and Talin A A 2017 *Adv. Mater.* **29** 1604310
[75] van de Burgt Y, Lubberman E, Fuller E J, Keene S T, Faria G C, Agarwal S, Marinella M J, Talin A A and Salleo A 2017 *Nat. Mater.* **16** 414-8
[76] Tang J *et al* 2018 *IEEE IEDM Tech. Dig.* 292-5
[77] Cubukcu M *et al* 2014 *Appl. Phys. Lett.* **104** 042406
[78] Miron I M *et al* 2011 *Nature* **476** 189-93
[79] Garello K *et al* 2014 *Appl. Phys. Lett.* **105** 212402
[80] Lo Conte R, Hrabec A, Mihai A P, Schulz T, Noh S-J, Marrows C H, Moore T A and Kläui M 2014 *Appl. Phys. Lett.* **105** 122404
[81] Garello K *et al* 2013 *Nat. Nanotechnol.* **8** 587-93
[82] Sangwan V K, Lee H-S, Bergeron H, Balla I, Beck M E, Chen K-S and Hersam M C 2018 *Nature* **544** 500-4
[83] Sangwan V K *et al* 2015 *Nat. Nanotechnol.* **10** 403-6
[84] Zhu X, Li D, Liang X and Lu W. D 2019 *Nat. Mater.* **18** 141-8
[85] LeCun Y, Bottou L, Bengio Y and Haffner P 1998 *Proc. IEEE* **86** 2278-324
[86] Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* Ch. 8 (MIT Press)
[87] Tsai H, Ambrogio S, Narayanan P, Shelby R M and Burr G W 2018 *J. Phys. D: Appl. Phys.* **51** 28
[88] Sze V, Chen Y, Yang T and Emer J S 2017 *Proc. IEEE* **105** 2295-329
[89] Jouppi N P *et al* 2017 *44th International Symposium on Computer Architecture (ISCA)* 1-17.
[90] Suri M *et al* 2011 *IEEE IEDM Tech. Dig.* 79-82
[91] Ambrogio S, Ciocchini N, Laudato M, Milo V, Pirovano A, Fantini P and Ielmini D *Front. Neurosci.* **10** 56
[92] Ambrogio S, Balatti S, Nardi F, Facchinetti S and Ielmini D 2013 *Nanotechnology* **24** 384012
[93] Jang J-W, Park S, Burr G W, Hwang H and Jeong Y-H. 2015 *IEEE Electron Device Lett.* **36** 457-9
[94] Hsu C-W, Wan C-C, Wang I-T, Chen M-C, Lo C-L, Lee Y-J, Jang W-Y, Lin C-H and Hou T-H 2013 *IEEE IEDM Tech. Dig.* 264-7
[95] Park S-G *et al* 2012 *IEEE IEDM Tech. Dig.* 501-4
[96] Wang I-T, Chang C-C, Chiu L-W, Chou T and Hou T-H 2016 *Nanotechnology* **27** 365204
[97] Yu S, Chen P-Y, Cao Y, Xia L, Wang Y and Wu H 2015 *IEEE IEDM Tech. Dig.* 451-4
[98] Park S, Sheri A, Kim J, Noh J, Jang J, Jeon M, Lee B, Lee B R, Lee B H and Hwang H 2013 *IEEE IEDM Tech. Dig.* 625-8
[99] Jo S H, Chang T, Ebong I, Bhadviya B B, Mazumder P and Lu W 2010 *Nano Lett.* **10** 1297-301
[100] Moon K, Kwak M, Park J, Lee D and Hwang H 2017 *IEEE Electron Device Lett.* **38** 1023-6
[101] Chen P-Y, Lin B, Wang I-T, Hou T-H, Ye J, Vrudhula S, Seo J-S, Cao Y and Yu S 2015 *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* 194-9
[102] Woo J, Moon K, Song J, Lee S, Kwak M, Park J and Hwang H 2016 *IEEE Electron Device Lett.* **37** 994-7
[103] Li Y, Kim S, Sun X, Solomon P, Gokmen T, Tsai H, Koswatta S, Ren Z, Mo R, Yeh C C, Haensch W and Leobandung E 2018 *IEEE Symp. VLSI Tech. Dig.* 25-6
[104] Yao P *et al* 2017 *Nat. Commun.* **8** 15199
[105] Fumarola A, Narayanan P, Sanches L L, Sidler S, Jang J, Moon K, Shelby R M, Hwang H and Burr G W 2016 *IEEE International Conference on Rebooting Computing (ICRC)* 1-8
[106] Yu S, Gao B, Fang Z, Yu H, Kang J and Wong H-S P 2013 *Adv. Mater.* **25** 1774-9
[107] Li C *et al* 2018 *Nat. Commun.* **9** 2385
[108] Boybat I *et al* 2018 *Nat. Commun.* **9** 2514

[109] Ambrogio S *et al* 2018 *Nature* **558** 60-7
[110] Agarwal S *et al* 2017 *IEEE Symp. VLSI Tech.* T174-5
[111] Cristiano G, Giordano M, Ambrogio S, Romero L P, Cheng C, Narayanan P, Tsai H, Shelby R M and Burr G W 2018 *J. Appl. Phys.* **124** 151901
[112] Masquelier T and Thorpe S J 2007 *PLoS Comput. Biol.* **3** e31
[113] Masquelier T, Guyonneau R and Thorpe S J 2008 *PLoS ONE* **3** e1377
[114] Serrano-Gotarredona T, Masquelier T, Prodromakis T, Indiveri G and Linares-Barranco B 2013 *Front. Neurosci.* **7** 2
[115] Saïghi S *et al* 2015 *Front. Neurosci.* **9** 51
[116] Kuzum D, Yu S and Wong H-S P 2013 *Nanotechnology* **24** 382001
[117] Bienenstock E L, Cooper L N and Munro P W 1982 *J. Neurosci.* **2**, 32-48
[118] Bear M F 1996 *Proc. Natl. Acad. Sci. USA* **93** 13453-9
[119] Gjorgjieva J, Clopath C, Audet J and Pfister J P 2011 *Proc. Natl. Acad. Sci. USA* **108** 19383-8
[120] Rachmuth G, Shouval H Z, Beard M F and Poon C-S 2011 *Proc. Natl. Acad. Sci. USA* **108** E1266-74
[121] Hebb D O 1949 *The Organization of Behavior: A Neuropsychological Study* (New York: Wiley)
[122] Gerstner W, Ritz R and Van Hemmen J 1993 *Biological Cybernetics* **69** 503-15
[123] Snider G 2008 *IEEE/ACM Int. Symp. on Nanoscale Architectures (NANOARCH 2008)* 85-92
[124] Kozicki M N, Park M and Mitkova M 2005 *IEEE Trans. Nanotechnol.* **4** 331-8
[125] Gilbert N, Zhang Y, Dinh J, Calhoun B and Hollmer S 2013 *IEEE Symp. VLSI Tech. Dig.* C204-5
[126] Guo X, Schindler C, Menzel S and Waser R 2007 *Appl. Phys. Lett.* **91** 133513
[127] Yang Y, Gao P, Gaba S, Chang T, Pan X and Lu W 2012 *Nat. Commun.* **3** 732
[128] Ambrogio S, Balatti S, Choi S and Ielmini D 2014 *Adv. Mater.* **26** 3885-92
[129] Kozicki M N, Gopalan C, Balakrishnan M and Mitkova M 2006 *IEEE Trans. Nanotechnol.* **5** 535-44
[130] Schindler C, Thermadam S C P, Waser R and Kozicki M N 2007 *IEEE Trans. Electron Devices* 2007 **54** 2762-8
[131] Schindler C, Staikov G and Waser R 2009 *Appl. Phys. Lett.* **94** 072109
[132] Schindler C, Weides M, Kozicki M N and Waser R 2008 *Appl. Phys. Lett.* **92** 122910
[133] Yu S, Wu Y, Jeyasingh R, Kuzum D and Wong H-S P 2011 *IEEE Trans. Electron Devices* **58** 2729-37
[134] Prezioso M, Merrikh Bayat F, Hoskins B, Likharev K and Strukov D 2016 *Sci. Rep.* **6** 21331
[135] Wang Z Q, Xu H Y, Li X H, Yu H, Liu Y C and Zhu X J 2012 *Adv. Funct. Mater.* **22** 2759-65
[136] Kuzum D, Jeyasingh R G D, Lee B and Wong H-S P 2012 *Nano Lett.***12** 2179-86
[137] Tuma T, Le Gallo M, Sebastian A and Eleftheriou E 2016 *IEEE Electron Device Lett.* **37** 1238-41
[138] Srinivasan G, Sengupta A and Roy K 2016 *Sci. Rep.* **6** 2954
[139] Boyn S *et al* 2017 *Nat. Commun.* **8** 14736
[140] Kim C-H, Lee S, Woo S Y, Kang W-M, Lim S, Bae J-H, Kim J and Lee J-H 2018 *IEEE Trans. Electron Devices* **65** 1774-8
[141] Malavena G, Spinelli A S and Monzio Compagnoni C 2018 *IEEE IEDM Tech. Dig.* 35-8
[142] Alibart F, Pleutin S, Bichler O, Gamrat C, Serrano-Gotarredona T, Linares-Barranco B and Vuillaume D 2012 *Adv. Funct. Mater.* **22** 609-16
[143] Ielmini D 2011 *IEEE Trans. Electron Devices* **58** 4309-17
[144] Ambrogio S, Balatti S, Milo V, Carboni R, Wang Z, Calderoni A, Ramaswamy N and Ielmini D 2016 *IEEE Trans. Electron Devices* **63** 1508-15
[145] Pedretti G, Milo V, Ambrogio S, Carboni R, Bianchi S, Calderoni A, Ramaswamy N, Spinelli A S and Ielmini D 2017 *Sci. Rep.* **7** 5288

[146] Wang Z-Q, Ambrogio S, Balatti S and Ielmini D 2015 *Front. Neurosci.* **8** 438
[147] Kim S *et al* 2015 *IEEE IEDM Tech.* Dig. 443-6
[148] Milo V, Pedretti G, Carboni R, Calderoni A, Ramaswamy N, Ambrogio S and Ielmini D 2016 *IEEE IEDM Tech. Dig.* 440-3
[149] Milo V, Pedretti G, Carboni R, Calderoni A, Ramaswamy N, Ambrogio S and Ielmini D 2018 *IEEE Trans. VLSI* **26** 2806-15
[150] Suri M *et al* 2013 *IEEE Trans. Electron Devices* **60** 2402-9
[151] Covi E, Brivio S, Serb A, Prodromakis T, Fanciulli M and Spiga S 2016 *Front. Neurosci.* **10** 482
[152] Serb A, Bill J, Khiat A, Berdan R, Legenstein R and Prodromakis T 2016 *Nat. Commun.* **7** 12611
[153] Hansen M, Zahari F, Kohlstedt H and Ziegler M 2018 *Sci. Rep.* **8** 8914
[154] Prezioso M, Mahmoodi M R, Merrikh Bayat F, Nili H, Kim H, Vincent A and Strukov D B 2018 *Nat. Commun.* **9** 5311
[155] Pedretti G, Milo V, Ambrogio S, Carboni R, Bianchi S, Calderoni A, Ramaswamy N, Spinelli A S and Ielmini D 2018 *IEEE J. Emerging Topics in Circuits and Systems (JETCAS)* **8** 77-85
[156] Wang W, Pedretti G, Milo V, Carboni R, Calderoni A, Ramaswamy N, Spinelli A S and Ielmini D 2018 *Sci. Adv.* **4** eaat475
[157] Diehl P U and Cook M 2015 *Front. Comput. Neurosci*. **9** 99
[158] Ambrogio S, Balatti S, Milo V, Carboni R, Wang Z, Calderoni A, Ramaswamy N and Ielmini D 2016 *IEEE Symp. VLSI Tech. Dig.* 196-7
[159] Milo V, Ielmini D and Chicca E 2017 *IEEE IEDM Tech. Dig.* 263-6
[160] Serrano-Gotarredona R *et al* 2009 *IEEE Trans. Neural Netw.* **20** 1417-38
[161] Ohno T, Hasegawa T, Tsuruoka T, Terabe K, Gimzewski J K and Aono M 2011 *Nat. Mater.* **10** 591-5
[162] Terabe K, Hasegawa T, Nakayama T and Aono M 2005 *Nature* **433** 47-50
[163] Kim S, Du C, Sheridan P, Ma W, Choi SH and Lu W D 2015 *Nano Lett.* **15** 2203-11
[164] Wang Z *et al* 2017 *Nat. Mater.* **16** 101-8
[165] Midya R *et al* 2017 *Adv. Mater.* **29** 1604457
[166] Wang M *et al* 2018 *Adv. Mater.* **30** 1802516
[167] Wang W, Wang M, Ambrosi E, Bricalli A, Laudato M, Sun Z, Chen X and Ielmini D 2019 *Nat. Commun.* **10** 81
[168] Wang W, Bricalli A, Laudato M, Ambrosi E, Covi E and Ielmini D 2018 *IEEE IEDM Tech. Dig.* 932-5
[169] Tuma T, Pantazi A, Le Gallo M, Sebastian A and Eleftheriou E 2016 *Nat. Nanotechnol.* **11** 693-9
[170] Lashkare S, Chouhan S, Chavan T, Bhat A, Kumbhare P and Ganguly U 2018 *IEEE Electron Device Lett.* **39** 484-7
[171] Wang Z *et al* 2018 *Nat. Electron.* **1** 137-45
[172] Hodgkin A L and Huxley A F 1952 *J. Physiol.* **117** 500-44
[173] Pickett M D, Medeiros-Ribeiro G and Williams R S 2013 *Nat. Mater.* **12** 114-7
[174] Pickett M D and Williams R S 2012 *Nanotechnology* **23** 215202
[175] Kim S *et al* 2012 *Proc. IEEE Symp. VLSI Technol. (VLSIT)* 155-6
[176] Nandi S K, Liu X, Venkatachalam D K and Elliman R G 2015 *J. Phys. D: Appl. Phys*. **48** 195105
[177] Torrejon J *et al* 2017 *Nature* **547** 428-31
[178] Mehonic A and Kenyon A J 2016 *Front. Neurosci.* **10** 57
[179] Ciocchini N, Palumbo E, Borghi M, Zuliani P, Annunziata R and Ielmini D 2014 *IEEE Trans. Electron Devices* **61** 2136-44
[180] Ambrogio S, Balatti S, McCaffrey V, Wang D and Ielmini D 2015 *IEEE Trans. Electron Devices* **62** 3812-9