Applied Physics Reviews

Pathways to efficient neuromorphic computing with non-volatile memory technologies

Cite as: Appl. Phys. Rev. **7**, 021308 (2020); https://doi.org/10.1063/1.5113536
Submitted: 05 June 2019 . Accepted: 01 May 2020 . Published Online: 03 June 2020

I. Chakraborty, A. Jaiswal, A. K. Saha, S. K. Gupta, and K. Roy

COLLECTIONS

Paper published as part of the special topic on Brain Inspired Electronics
Note: This paper is part of the special collection on Brain Inspired Electronics.

F This paper was selected as Featured

View Online

Export Citation

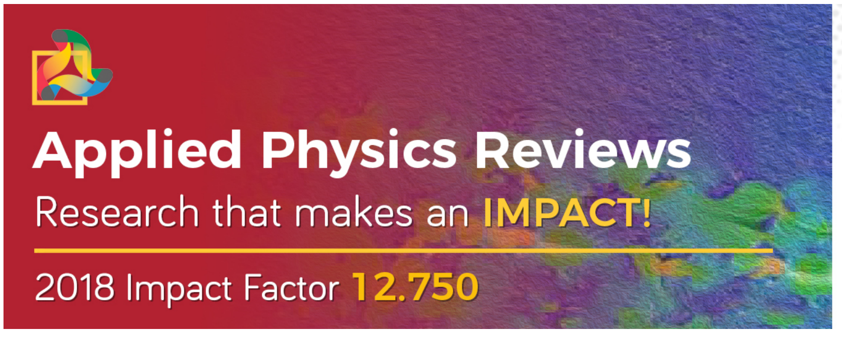
CrossMark

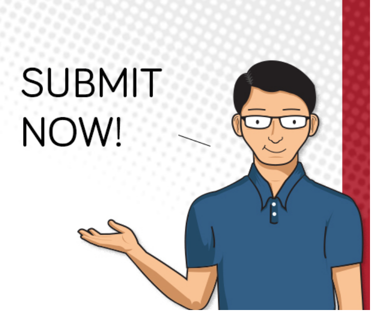
ARTICLES YOU MAY BE INTERESTED IN

Machine-learning-assisted metasurface design for high-efficiency thermal emitter optimization
Applied Physics Reviews **7**, 021407 (2020); https://doi.org/10.1063/1.5134792

A comprehensive review on emerging artificial neuromorphic devices
Applied Physics Reviews **7**, 011312 (2020); https://doi.org/10.1063/1.5118217

Nonlinear topological photonics
Applied Physics Reviews **7**, 021306 (2020); https://doi.org/10.1063/1.5142397





© 2020 Author(s).

Applied Physics Reviews REVIEW scitation.org/journal/are

Pathways to efficient neuromorphic computing with non-volatile memory technologies

Cite as: Appl. Phys. Rev. **7**, 021308 (2020); doi: 10.1063/1.5113536
Submitted: 5 June 2019 · Accepted: 1 May 2020 ·
Published Online: 3 June 2020

I. Chakraborty,[a] A. Jaiswal, A. K. Saha, S. K. Gupta, and K. Roy

AFFILIATIONS

School of Electrical and Computer Engineering, Purdue University, 465 Northwestern Ave., West Lafayette, Indiana 47906, USA

Note: This paper is part of the special collection on Brain Inspired Electronics.
[a]**Author to whom correspondence should be addressed:** ichakra@purdue.edu

ABSTRACT

Historically, memory technologies have been evaluated based on their storage density, cost, and latencies. Beyond these metrics, the need to enable smarter and intelligent computing platforms at a low area and energy cost has brought forth interesting avenues for exploiting non-volatile memory (NVM) technologies. In this paper, we focus on non-volatile memory technologies and their applications to bio-inspired neuromorphic computing, enabling spike-based machine intelligence. Spiking neural networks (SNNs) based on discrete neuronal "action potentials" are not only bio-fidel but also an attractive candidate to achieve energy-efficiency, as compared to state-of-the-art continuous-valued neural networks. NVMs offer promise for implementing both area- and energy-efficient SNN compute fabrics at almost all levels of hierarchy including devices, circuits, architecture, and algorithms. The intrinsic device physics of NVMs can be leveraged to emulate dynamics of individual neurons and synapses. These devices can be connected in a dense crossbar-like circuit, enabling in-memory, highly parallel dot-product computations required for neural networks. Architecturally, such crossbars can be connected in a distributed manner, bringing in additional system-level parallelism, a radical departure from the conventional von-Neumann architecture. Finally, cross-layer optimization across underlying NVM based hardware and learning algorithms can be exploited for resilience in learning and mitigating hardware inaccuracies. The manuscript starts by introducing both neuromorphic computing requirements and non-volatile memory technologies. Subsequently, we not only provide a review of key works but also carefully scrutinize the challenges and opportunities with respect to various NVM technologies at different levels of abstraction from devices-to-circuit-to-architecture and co-design of hardware and algorithm.

Published under license by AIP Publishing. https://doi.org/10.1063/1.5113536

TABLE OF CONTENTS

I. INTRODUCTION . . . 2
II. GENERIC NEURO-SYNAPTIC BEHAVIORAL AND LEARNING REQUIREMENTS . . . 4
A. Neurons . . . 4
B. Synapses . . . 5
1. Unsupervised learning . . . 5
2. Supervised learning . . . 6
III. NON-VOLATILE TECHNOLOGIES FOR NEUROMORPHIC HARDWARE . . . 6
A. Phase change devices . . . 7
1. PCM as neurons . . . 7
2. PCM as synapses . . . 8
3. PCM crossbars . . . 9
B. Metal-oxide RRAMs and CBRAMs . . . 10
1. Metal-oxide RRAMs and CBRAMs as neurons . . . 10
2. Metal-oxide RRAMs and CBRAMs as synapses . . . 11
3. Metal-oxide RRAM and CBRAM crossbars . . . 13
C. Spintronic devices . . . 13
1. Spin devices as neurons . . . 13
2. Spin devices as synapses . . . 15
3. Spintronic crossbars . . . 16
D. Ferroelectric FETs . . . 17
1. FEFETs as neurons . . . 18
2. FEFETs as synapses . . . 18
3. FEFET crossbars . . . 19
E. Floating gate devices . . . 19
1. Floating gate devices as neurons . . . 19
2. Floating gate devices as synapses . . . 20
3. Floating gate crossbars . . . 20
F. NVM architecture . . . 20
IV. PROSPECTS . . . 22

Published under license by AIP Publishing

A. Stochasticity—Opportunities and challenges 22
B. Challenges of NVM crossbars 22
C. Mitigating crossbar non-idealities 24
D. Multi-memristive synapses 24
E. Beyond neuro-synaptic devices and STDP 25
F. NVM for digital in-memory computing 25
G. Physical integrability of NVM technology with CMOS 25
V. CONCLUSION 25
AUTHORS' CONTRIBUTION 26

I. INTRODUCTION

The human brain remains a vast mystery and continues to baffle researchers from various fields alike. It has intrigued neuroscientists by its underlying neural circuits and topology of brain networks that result in vastly diverse cognitive and decision-making functionalities as a whole. Equivalently, computer engineers have been fascinated by the energy-efficiency of the biological brain in comparison to the state-of-the-art silicon computing solutions. For example, the Bluegene supercomputer[1] consumed mega-watts of power[2] for simulating the activity of cat's brain.[3] This is in contrast to ~20 W of power accounting for much more complex tasks including cognition, control, movement, and decision making, being rendered simultaneously by the brain. The massive connectivity of the brain fueling its cognitive abilities and the unprecedented energy-efficiency makes it by far the most remarkable known intelligent system. It is, therefore, not surprising that in the quest to achieve "brain-like cognitive abilities with brain-like energy-efficiency," researchers have tried building *Neuromorphic Systems* closely inspired by the biological brain (refer Fig. 1). Worth noting is the fact that neuromorphic computing not only aims at attaining the energy-efficiency of the brain but also encompasses attempts to mimic its rich functional principles such as cognition, efficient spike-based information passing, robustness, and adaptability. Interestingly, both the brain's cognitive ability and its energy-efficiency stem from basic computation and storage primitives called neurons and synapses, respectively.

Networks comprising artificial neurons and synapses have, therefore, been historically explored for solving various *intelligent* problems. Over the years, neural networks have evolved significantly and are usually categorized based on the characteristic neural transfer function as first, second, and third generation networks.[4] As shown in Fig. 2, the first generation neurons, called as *perceptrons*,[4] had a step function response to the neuronal inputs. The step perceptrons, however, were not scalable to deeper layers and were extended to Multi-Layer Perceptrons (MLPs) using non-linear functional units.[5] This is alluded to as the second generation neurons based on a continuous neuronal output with non-linear characteristic functions such as *sigmoid*[5] and *ReLU (Rectified Linear Unit)*.[6] Deep Learning Networks (DLNs) as we know it today are based on such second generation neural networks. The present revolution in artificial intelligence is being currently fueled by such DLNs using global learning algorithms based on the gradient descent rule.[7] Deep learning has been used for myriad of applications including classification, recognition, prediction, cognition, and decision making with unprecedented success.[8] However, a major requirement to achieve the vision of *intelligence everywhere* is to enable energy-efficient computing much beyond the existing Deep learning solutions. Toward that end, it is expected that networks of spiking neurons hold promise for building an energy-efficient alternative to traditional DLNs. Spiking neural networks (SNNs)—the third generation of neural networks—are based on the bio-plausible neural behavior and communicate through discrete spikes as opposed to the continuous valued signal of DLNs. Note that for this paper, we refer the second generation networks as DLNs and the third generation spiking networks as SNNs.

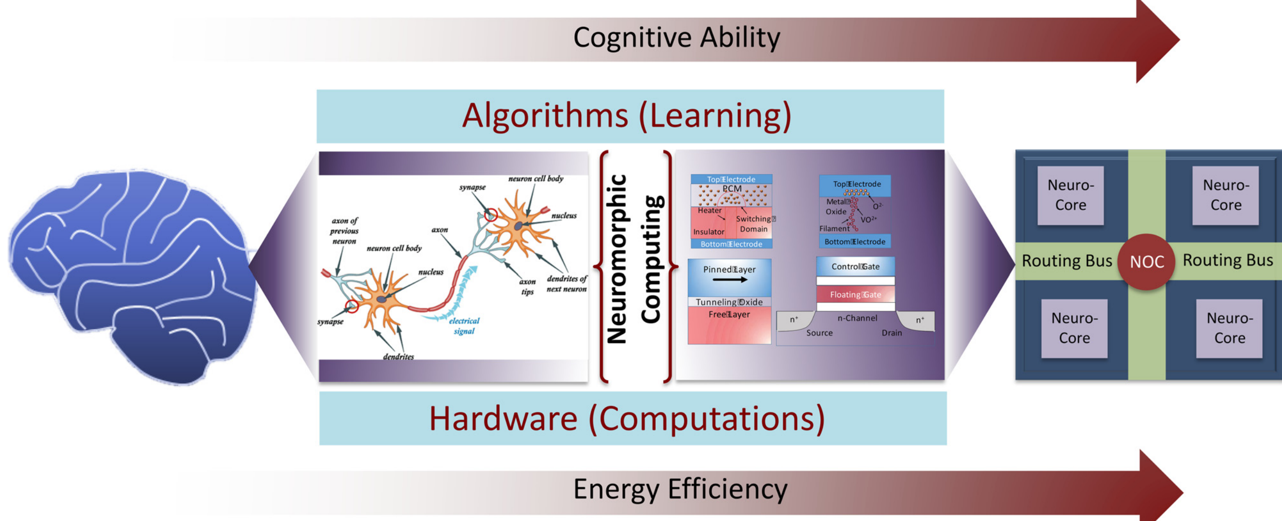


FIG. 1. Neuromorphic computing as a brain-inspired paradigm to achieve cognitive ability and energy-efficiency of the biological brain. "Hardware" and "Algorithms" form the two key aspects for neuromorphic systems. As shown in the right hand side, a generic neuromorphic chip consists of several "Neuro-Cores" interconnected through the address event representation (AER) based network-on-chip (NOC). Neuro-Cores consist of arrays of synapses and neurons at the periphery. Non-volatile technologies including PCM, RRAM, MRAM, and FG devices have been used to mimic neurons and synapses at various levels of bio-fidelity.

Published under license by AIP Publishing

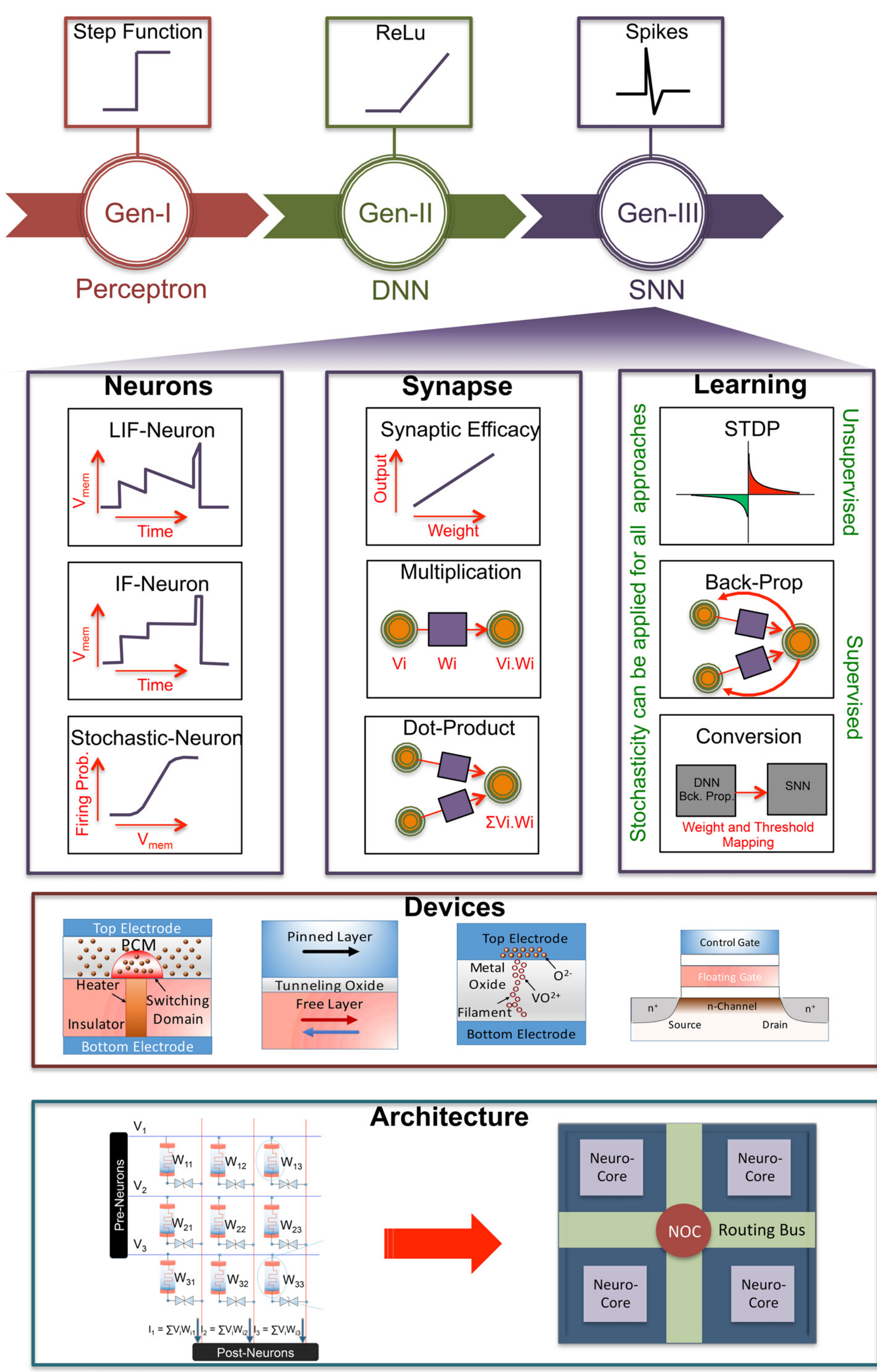


FIG. 2. Three generations of neural networks. First generation (Gen-I) of networks used step transfer functions and were not scalable, and second generation (Gen-II) uses transfer functions such as Rectified Linear Unit (ReLU) that has fueled today's deep learning networks. The third generation (Gen-III) uses spiking neurons resembling the neural activity of their biological counterparts. The three components of an SNN are (1) neurons, (2) synapses, and (3) learning. (1) Neurons: three broad classes of spiking neurons that researchers attempt to mimic using NVMs are Leaky-Integrate-Fire (LIF), Integrate-Fire (IF), and Stochastic Neurons. (2) Synapses: the key attributes needed for a particular device to function as a synapse are its ability to map synaptic efficacy (wherein a synaptic weight modulates the strength of the neuronal signal) and that they can perform multiplication and dot-product operations. (3) Learning: as shown in the figure, learning can be achieved either through supervised or unsupervised algorithms. From an NVM perspective, various NVM technologies are being used to mimic neuronal and synaptic functionalities with appropriate learning capabilities. At an architectural level, arrays of such NVMs are connected through the network-on-chip to enable seamless integration of a large neural network.

Published under license by AIP Publishing

From the energy-efficiency perspective, SNNs have two key advantages. First, the fact that neurons exchange information through discrete *spikes* is explicitly utilized in hardware systems to enable energy-efficient event-driven computations. By event-driveness, it is implied that only those units in the hardware system are active, which have received a spike, and all other units remain idle reducing the energy expenditure. Second, such an event-driven scheme also enables Address Event Representation (AER).[9] AER is an asynchronous communication scheme, wherein the *sender* transmits its address on the system bus and the *receiver* regenerates the spikes based on the addresses it receives through the system bus. Thereby, instead of transmitting and receiving the actual data, *event addresses* are exchanged between the sender and the receiver, leading to energy-efficient transfer of information.

In addition to emulation of neuro-synaptic dynamics and use of event-driven hardware, two notable developments, namely, (1) the emergence of various non-volatile technologies and (2) the focus on learning algorithms for networks of spiking neurons, have accelerated the efforts in driving neural network hardware closer toward achieving both energy-efficiency and improved cognitive abilities. Non-volatile technologies have facilitated area- and energy-efficient implementations of neuromorphic systems. As we will see in Sec. III of the manuscript, these devices are of particular interest since they are governed by intrinsic physics that can be mapped directly to certain aspects of biological neurons and synapses. This implies that instead of using multiple transistors to imitate neuronal and synaptic behavior, in many cases, a single non-volatile device can be used as a neuron or a synapse with various degrees of bio-fidelity. In addition, a major benefactor for non-volatile memory (NVM) technologies is that they can be arranged in dense crossbars of synaptic arrays with neurons at the periphery. This is of immense importance since the co-locations of compute (neuronal primitives) and storage (synaptic primitives) are inherent characteristics that make the biological brain so effective. Note that this closely intertwined fabric of compute and storage is conspicuously different from state-of-the-art computing systems that rely on the well-known von-Neumann model with segregated compute and storage units. Additionally, learning algorithms for networks of spiking neurons has recently attracted considerable research focus. For this paper, we would define neuromorphic computing as *SNN based neural networks, associated learning algorithms, and their hardware implementations*.

In this paper, we focus on non-volatile technologies and their applications to neuromorphic computing. With reference to Fig. 2, we start in Sec. II by first describing the generic neural and synaptic behavioral characteristics that are in general emulated through non-volatile devices. Subsequently, in Sec. III, we describe learning strategies for SNNs and associated topologies. With the knowledge of basic neuro-synaptic behavior and learning methodologies, Sec. IV presents non-volatile memories as the building block for neuromorphic systems. Finally, before concluding, we highlight on future prospects and key areas of research that can further the cause of neuromorphic hardware by exploiting non-volatile technologies.

II. GENERIC NEURO-SYNAPTIC BEHAVIORAL AND LEARNING REQUIREMENTS

One of the key advantages of non-volatile technologies is that their intrinsic device characteristics can be leveraged to map certain aspects of biological neurons and synapses. Let us highlight few *representative behaviors* for both neurons and synapses that form the basic set of neuro-synaptic dynamics usually replicated through non-volatile devices.

A. Neurons

Neural interactions are time varying electro-chemical dynamics that gives rise to brain's diverse functionalities. These dynamical behaviors in turn are governed by voltage dependent opening and closing of various charge pumps that are selective to specific ions such as Na^+ and K^+.[10,11] In general, a neuron maintains a *resting potential*, across its cell membrane by maintaining a constant charge gradient. Incoming spikes to a neuron lead to an increase in its membrane potential in a leaky-integrate manner until the potential crosses a certain threshold after which the neuron emits a spike and remains non-responsive for a certain period of time called as the *refractory period*. A typical spike (or action potential) is shown in Fig. 3 highlighting the specific movements of charged ions through the cell membrane. Additionally, it has been known that the firing activity of neurons is stochastic in nature.[12,13]

Having known the generic qualitative nature of neural functionality, it is obvious that a resulting model, describing the intricacies of a biological neuron, would consist of complex dynamical equations. In fact, detailed mathematical models such as Hodgkin–Huxley model[14] and spike response model have been developed, which closely match the behavior of biological neurons. However, implementing such models in hardware turns out to be a complex task. As such, hardware implementations mostly focus on simplified neuronal models, such as Leaky-Integrate-Fire (LIF) model[15–17] shown in Fig. 3. Consequently, the diverse works on mimicking neurons using non-volatile technologies can be categorized into three genres—(1) the Leaky-Integrate-Fire (LIF) neurons, (2) the Integrate-Fire (IF) neurons, and (3) Stochastic-Firing (s-F) neurons. Figure 2 graphically represents the typical neural behavior for each type of neuron, while Fig. 3(c) presents a Venn-diagram highlighting various works based on non-volatile technologies and the corresponding neural behavior that they are based on.

- Leaky-Integrate-Fire (LIF) neurons: The membrane potential of an LIF neuron is incremented at every instance when the neuron receives an input spike. In the interval between two spikes, the neuron potential slowly leaks, resulting in the typical leaky-integrate behavior shown in Fig. 2. If the neuron receives sufficient input spikes, its membrane potential crosses a certain threshold, eventually allowing the neuron to emit an output spike.
- Integrate-Fire (IF) neurons: The IF neuron is a simplified version of the LIF neuron without the leaky behavior. Essentially, an IF neuron increments its membrane potential at every spike maintaining its potential at a constant value between two spikes, as shown in Fig. 2. IF neurons fire when the accumulated membrane potential crosses a pre-defined threshold.
- Stochastic-Firing neurons: In contrast to deterministic neurons that fire whenever the neuron crosses its threshold, a stochastic firing neuron fires with a probability, which is proportional to its membrane potential. In other words, for a stochastic neuron, an output spike is emitted with a certain probability, which is a function of the instantaneous membrane potential. In its simplest

Published under license by AIP Publishing

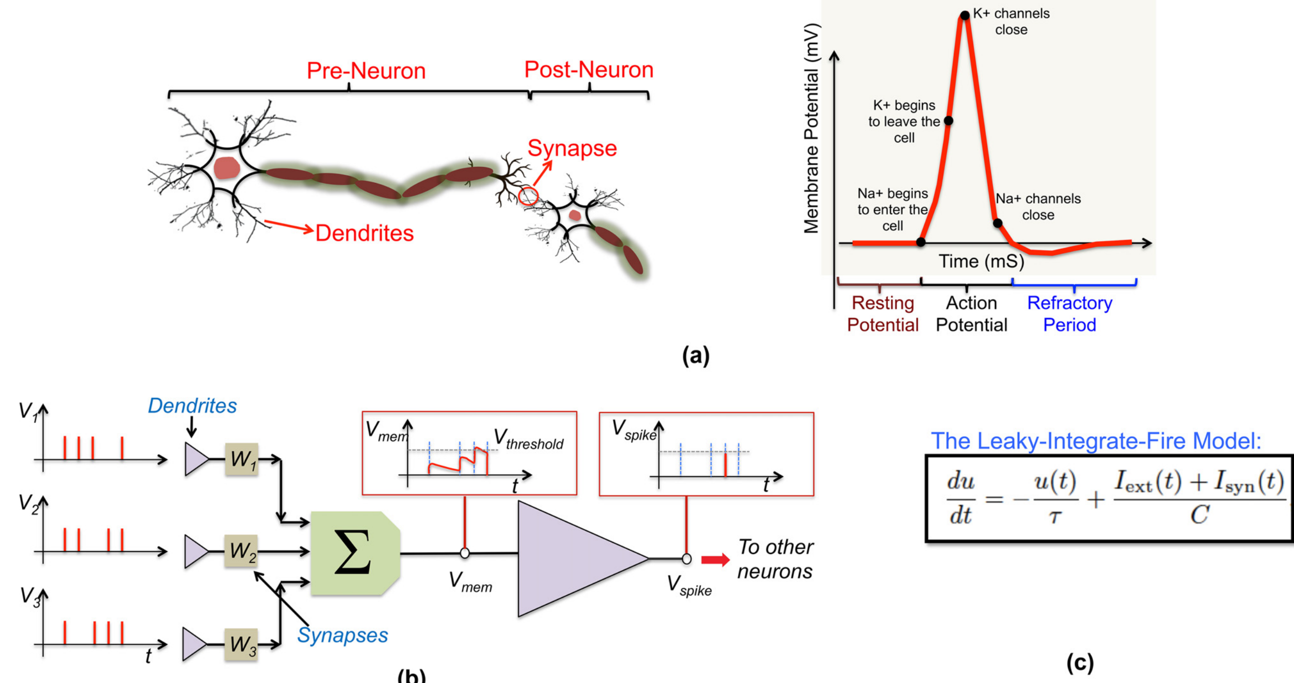


FIG. 3. (a) The biological neuron and a typical spiking event. Various ions and the role they play in producing the spiking event are shown. (b) A simplified neural computing model highlighting the flow of information from the input of neurons to the output. Spikes from various pre-neurons are multiplied by the corresponding weights and added together before being applied as an input to the neuron. The neuron shows a typical leaky-integrate behavior unless its membrane potential crosses a certain threshold, leading to emission of a spike. (c) The LIF differential equation.

form, a stochastic firing behavior can be modeled by a firing probability, which increases with the input stimulus. However, stochasticity can also be combined with LIF and IF neurons, such that once the neuron crosses the threshold, it only emits a spike based on a probabilistic function.

LIF neurons are most widely used in the domain of SNNs. The leaky nature of LIF neurons renders a regularizing effect on their firing rates. This can help particularly for frequency based adaptation mechanisms that we will discuss in the next section.[18] IF neurons are typically used in supervised learning algorithms. In these algorithms, the learning mechanism does not have temporal significance, and hence, temporal regularization is not required. Stochastic neurons, on the other hand, have a different computing principle. Due to the probabilistic nature of firing, it can also act as a regularizer and also lead to better generalization behavior in neural networks. All the aforementioned neurons can leverage the inherent device physics in NVM devices for efficient hardware implementation.

B. Synapses

Information in biological systems is governed by transmission of electrical pulses between adjacent neurons through connecting bridges, commonly known as synapses. *Synaptic efficacy*, representing the strength of connection through an internal variable, is the basic criterion for any device to work as an artificial synapse. Neuro-chemical changes can induce plasticity in synapses by permanently manipulating the release of neurotransmitters and controlling the responsiveness of the cells to them. Such plasticity is believed to be the fundamental basis of learning and memory in the biological brain. From the neuromorphic perspective, synaptic learning strategies can be broadly classified into two major classes: (1) unsupervised learning and (2) supervised learning.

1. Unsupervised learning

Unsupervised learning is a class of learning algorithms associated with self-organization of weights without the access to labeled data. In the context of hardware implementations, unsupervised learning relates to biologically inspired localized learning rules where the weight updates in the synapses depend solely on the activities of the neurons on its either ends. Unsupervised learning in spike-based systems can be broadly classified into (i) Spike Timing Dependent Plasticity (STDP) and (ii) frequency dependent plasticity.

Spike timing dependent plasticity (STDP), shown in Fig. 4, is a learning rule, which strengthens or weakens the synaptic weight based on the relative timing between the activities of the connected neurons. This kind of learning was first experimentally observed in rat's hippocampal glutamatergic synapses.[19] It involves both long-term potentiation (LTP),[20] which signifies the increase in the synaptic weight^{2+}, and long-term depression (LTD), which signifies a reduction in the synaptic weight. LTP is realized through STDP when the post-synaptic neuron fires after the pre-synaptic activity, whereas LTD results from an

Published under license by AIP Publishing

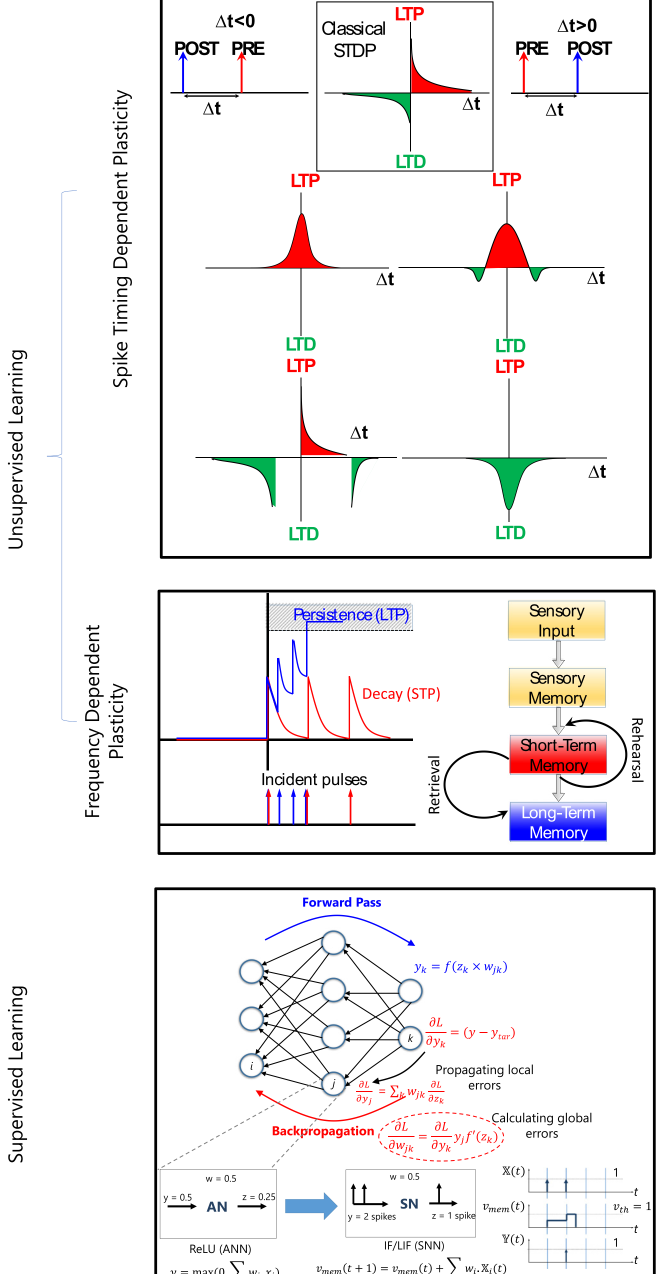


FIG. 4. Different kinds of learning strategies can be broadly classified into (i) spiking timing dependent plasticity (STDP), (ii) frequency dependent plasticity, and (iii) gradient-based learning. STDP induces both potentiation and depression of synaptic weights in a non-volatile fashion based on the difference in spike timing of pre-neurons and post-neurons, Δt. Classical STDP assumes an exponential relationship with Δt, as demonstrated by Bi and Poo.[19] Other variants of STDP have also been observed in mammalian brains. Frequency dependent plasticity manifests itself in the form of short-term plasticity (STP) and long-term potentiation (LTP). The change in the synaptic weight, in this case, depends on how frequently the synapse receives stimulus. STP and LTP form the basis of short-term and long-term memory in biological systems. Finally, gradient-based learning is a supervised learning scheme where the change in the synaptic weight depends on gradients calculated from error between the predicted and the ideal output.

acausal spiking between the pre-synaptic and post-synaptic neurons, wherein the post-synaptic neuron fires before the pre-synaptic neuron.

Mathematically, the relative change in synaptic strength is dependent on the timing difference of the post-synaptic and pre-synaptic spikes as

$$\delta w(\Delta t) = A^{+} \exp(-\Delta t/\tau^{+}) \text{ if } \Delta t > 0, \quad (1)$$

$$= A^{-} \exp(\Delta t/\tau^{-}) \text{ if } \Delta t < 0. \quad (2)$$

Here, A^{+}, A^{-}, τ^{+}, τ^{-} are the amplification coefficients and time-constants, respectively, and Δt is defined as the difference between the pre-synaptic and post-synaptic firing instants. STDP has been widely adopted in not only computational neuroscience but also neuromorphic systems as the *de facto* unsupervised learning rule for pattern detection and recognition.

In conjunction to long-term modification of synaptic weights, the physiology of synapses induces yet another type of learning, i.e., frequency dependent plasticity, dependent on the activity of the pre-synaptic potential.[21,22] Activity-dependent learning can induce two types of changes in the synaptic strength. The change occurring over a short timescale (hundreds of milliseconds in biological systems) is known as Short-Term Plasticity (STP), while the long-term effects are a form of LTP that can last between hours to years. In general, at a given instance, a pre-synaptic activity induces STP; however, when the pre-synaptic activity reduces, the synaptic efficacy is reverted back to its original state. Repeated stimuli eventually result in LTP in the synapses. As STP corresponds to the recent history of activity and LTP relates to long-term synaptic changes resulting from activity over a period of time, they are often correlated with short-term memory (STM) and long-term memory (LTM), respectively, in mammals.[23]

2. Supervised learning

Although unsupervised learning is believed to form the dominant part of learning in biological synapses, the scope of its applicability is still in its nascent stages in comparison to conventional deep learning. An alternative *ex situ* learning methodology to enable spike-based processing in deep SNNs is restricting the training to the analog domain, i.e., using the greedy gradient descent algorithm as in conventional DLNs and converting such an analog valued neural network to the spiking domain for inferencing. Various conversion algorithms[24–26] have been proposed to perform nearly lossless transformation from the DLN to the SNN. These algorithms address several concerns pertaining to the conversion process, primarily emerging due to differences in neuron functionalities in the two domains. Such conversion approaches have been demonstrated to scale to state-of-art neural network architectures such as ResNet and VGG performing classification tasks on complex image datasets as in ImageNet.[27] More recently, there has been a considerable effort in realizing gradient-based learning in the spiking domain itself[28] to eliminate conversion losses.

III. NON-VOLATILE TECHNOLOGIES FOR NEUROMORPHIC HARDWARE

As elaborated in Sec. II, SNNs not only are biologically inspired neural networks but also potentially offer energy-efficient hardware solutions due to their inherent sparsity and asynchronous signal processing. Advantageously, non-volatile technologies provide two

Published under license by AIP Publishing

additional benefits with respect to neuromorphic computing. First, the inherent physics of such devices can be exploited to capture the functionalities of biological neurons and synapses. Second, these devices can be connected in a crossbar fashion allowing analog-mixed signal in-memory computations, resulting in highly energy-efficient hardware implementations.

In this section, we first delve into the possibilities and challenges of such non-volatile devices, based on various technologies, used to emulate the characteristics of synapses and neurons. Subsequently, we describe how crossbar structures of such non-volatile devices can be used for in-memory computing and the associated challenges.

A. Phase change devices

Phase change materials (PCMs) such as chalcogenides are the front-runners among emerging non-volatile devices—with speculation about possible commercial offerings—for high density, large-scale storage solutions.[31] These materials can encode multiple intermediate states, rendering them the capability of storing multiple bits in a single cell. More recently, PCM devices have also emerged as a promising candidate for neuromorphic computing due to their multi-level storage capabilities. In this section, we discuss various neuromorphic applications of PCM devices.

1. PCM as neurons

PCM devices show reversible switching between amorphous and crystalline states, which have highly contrasting electrical and optical properties. In fact, this switching dynamics can directly lead to integrate and firing behaviors in PCM-based neurons. The device structure of such a neuron comprises a phase change material sandwiched between two electrodes, as shown in Fig. 5(a). The mushroom

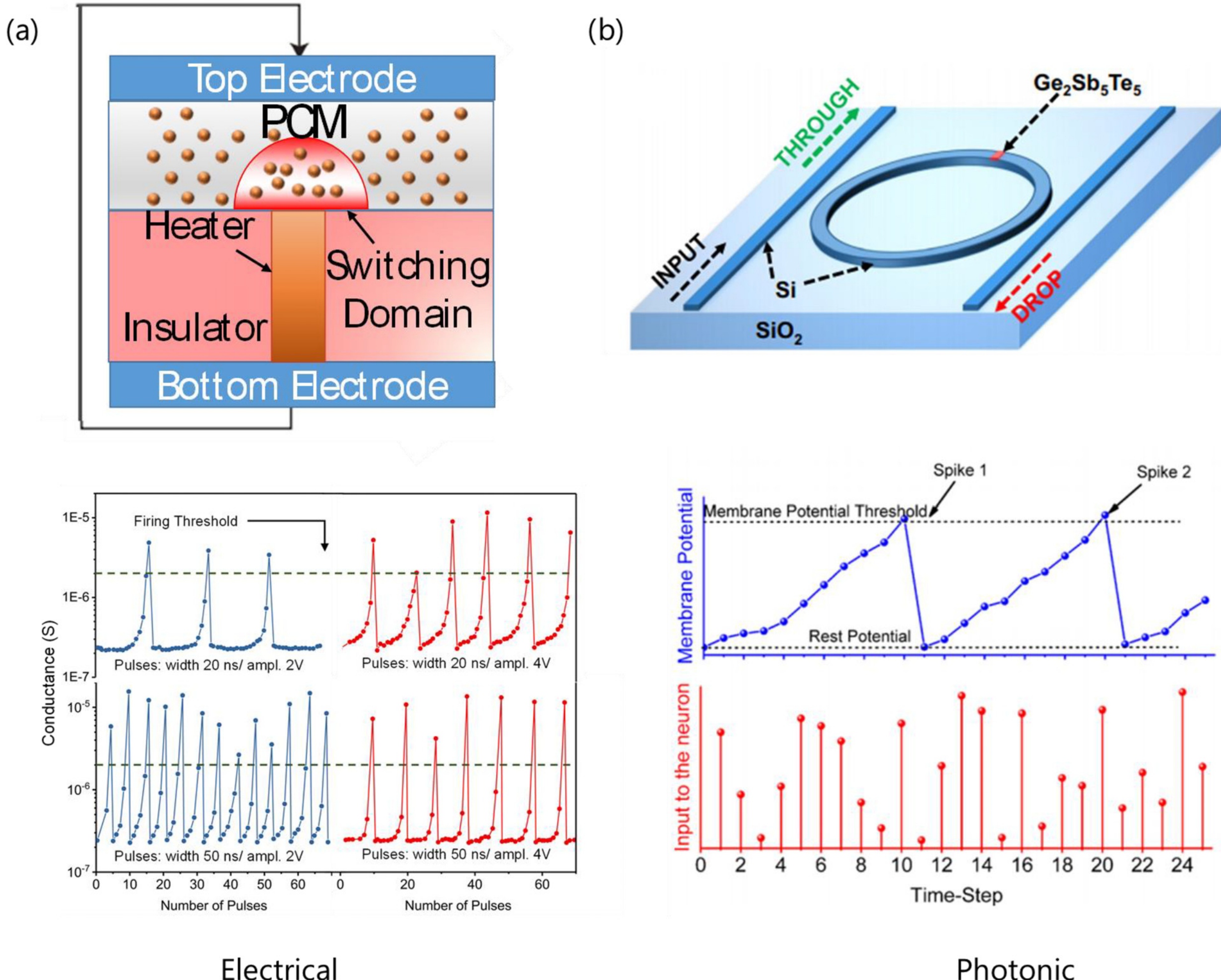


FIG. 5. (a) Device structure of a PCM-based IF neuron.[29] The thickness of the amorphous region (shown in red) represents the membrane potential of the neuron. The integrating and firing behaviors for different incident pulse amplitudes and frequencies are shown (bottom). (b) Device structure of a photonic IF neuron based on PCM (GST).[30] The input pulses coming through the INPUT port get coupled to the ring waveguide and eventually to the GST element, changing the amorphous thickness. The output at the "THROUGH" port represents the membrane potential, which depends on the state of the GST element.

Published under license by AIP Publishing

structure shows the shape of the switching volume just above the region known as the heater. The heater is usually made of resistive elements such as W, and high current densities at the contact interface between the phase change material and the heater cause locally confined Joule heating. When the PCM in the neuron is in its initial amorphous state, a voltage pulse that has an amplitude low enough so as to not melt the device but high enough to induce crystal growth can be applied. The resulting amorphous thickness, u_a, on application of such a pulse is given as[29]

$$\frac{du_a}{dt} = -v_g(R_{th}(u_a)P_p + T_{amb}), \quad u_a(0) = u_0 \tag{3}$$

where v_g is the crystal growth velocity dependent on the temperature determined by its argument $R_{th}(u_a)P_p + T_{amb}$. Here, R_{th} is the thermal resistance and T_{amb} is the interface temperature between amorphous and crystalline regions. The variable, u_a, in Eq. (1) can be interpreted as the neuron's membrane potential where P_p is the input variable controlling the dynamics. On successive application of crystallization pulses, the amorphous thickness, u_a, decreases, leading to lower conductance and temporal integration of the membrane potential. Beyond a certain threshold conductance level, the neuron fires, or in other words, the PCM changes to a crystalline state. A reset mechanism puts the neuron back in its original amorphous state. The aforementioned integrate-and-fire characteristics in PCM neurons are accompanied by inherent stochasticity. The stochasticity arises from different amorphous states created by repeated resets of the neuron. Different initial states lead to different growth velocities, which result in an approximate Gaussian distribution of inter-spike intervals, the interval between adjacent firing events. Populations of such stochastic IF neurons have also been used in detection of temporal correlation in parallel data streams.[32]

Thus far, we have talked about electronic devices mimicking neuronal behavior using PCM. Such behavior can also be achieved with Si-based photonic devices with PCM embedded on top of them.[30] Such a device is shown in Fig. 5(b), which consists of a Si microring resonator on the SiO_2 substrate with a phase change material, $Ge_2Sb_2Te_5$ (GST), deposited on top of the ring waveguide. The membrane potential of such a neuron or, in other words, the amorphous thickness of the PCM can be modulated by guiding laser pulses through Si waveguides. Light gets evanescently coupled to the PCM element and changes the thickness of the amorphous region, thereby allowing an optical IF neuron based on PCM elements, as shown in Fig. 5(b) (bottom).

2. PCM as synapses

We have discussed the ability of PCM to store multiple bits in a single cell. This multi-level behavior of PCM-based devices makes them a promising candidate to emulate synaptic characteristics. In addition, the large contrast in electrical properties allows for a significantly high ON/OFF resistance ratio in PCM devices. The same two-terminal structure described in Fig. 5(a) can be used as a synaptic device. The programming of such a synapse is performed through the phase transition mechanism between amorphous and crystalline states. Amorphization (or "RESET") is performed by an abrupt melt-quench process, where high and short voltage pulses are applied to heat the device followed by rapid cooling such that the material solidifies in the amorphous state. On the other hand, crystallization is performed when an exponential current above the threshold voltage leads to heating of the material above its crystallization temperature and switches it to the crystalline state, as depicted by the I–V characteristics in Fig. 6(a). The crystallization (or "SET") pulses are much longer as opposed to amorphization (or RESET) pulses, as shown in Fig. 6(b). Multiple states are achieved by progressively crystallizing the material, thus reducing the amorphous thickness.

These multi-level PCM synapses can be used to perform unsupervised on-chip learning using the STDP rule.[33] LTP and LTD using STDP involves a gradual increase and decrease in conductance of PCM devices, respectively. However, such a gradual increase or decrease in conductance needs to ensure precise control, which is difficult to achieve using identical current pulses. As a result, by configuring a series of programming pulses of increasing or decreasing amplitude [Fig. 7(a)], both LTP and LTD have been demonstrated using PCM devices.[34–36] In this particular scheme, the pre-spikes consist of a number of pulses of gradually decreasing or increasing pulses, whereas the post-spike consists of a single negative pulse. The difference between the magnitude of the pre-spike and post-spike due to overlap of the pulses varies with the time difference, resulting in the change in conductance of the synapse following the STDP learning rule. The scheme for potentiation is explained in Fig. 7(a). A simplified STDP learning rule with constant weight update can also be implemented using a single programming pulse by shaping the pulses appropriately[33] as shown in Fig. 7(b). However, such pulse shaping requires additional circuitry. These schemes rely on single PCM devices representing a synapse. Alternatively, using a "2-PCM" synapse, one can potentially implement LTP and LTD characteristics that can be independently programmed. Such a multi-device implementation becomes important for PCM technology as the amorphization is an abrupt process, and it is difficult to control the progression of different amorphization states, which poses a fundamental limitation toward realizing both LTP and LTD in a single device. Visual pattern recognition has been demonstrated using such 2-device synapses, which are able to learn directly from event-based sensors.[37] While these works focus on asymmetric STDP, which forms the basis of learning spatio-temporal features, PCM synapses can also exhibit symmetric STDP based learning enabling associative learning.[38] As we had discussed about IF neurons, the difference in optical responsivity of PCMs can also lead to emulation of synaptic behavior on Si-photonic devices. The change in optical transmission in photonic synaptic devices arises from the difference in the imaginary part of the refractive index of PCMs in their amorphous and crystalline states. The gradual increase

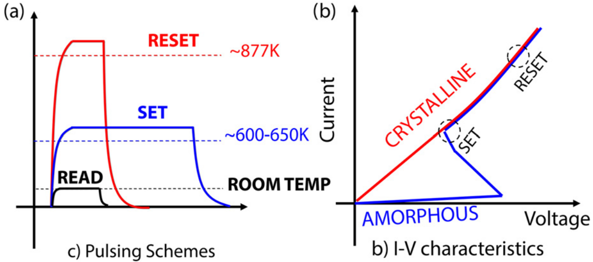


FIG. 6. (a) I–V characteristics of PCM devices showing SET and RESET points for two states. (b) Pulsing schemes for SET and RESET processes to occur, showing the temperatures reached due to the pulses.

Published under license by AIP Publishing

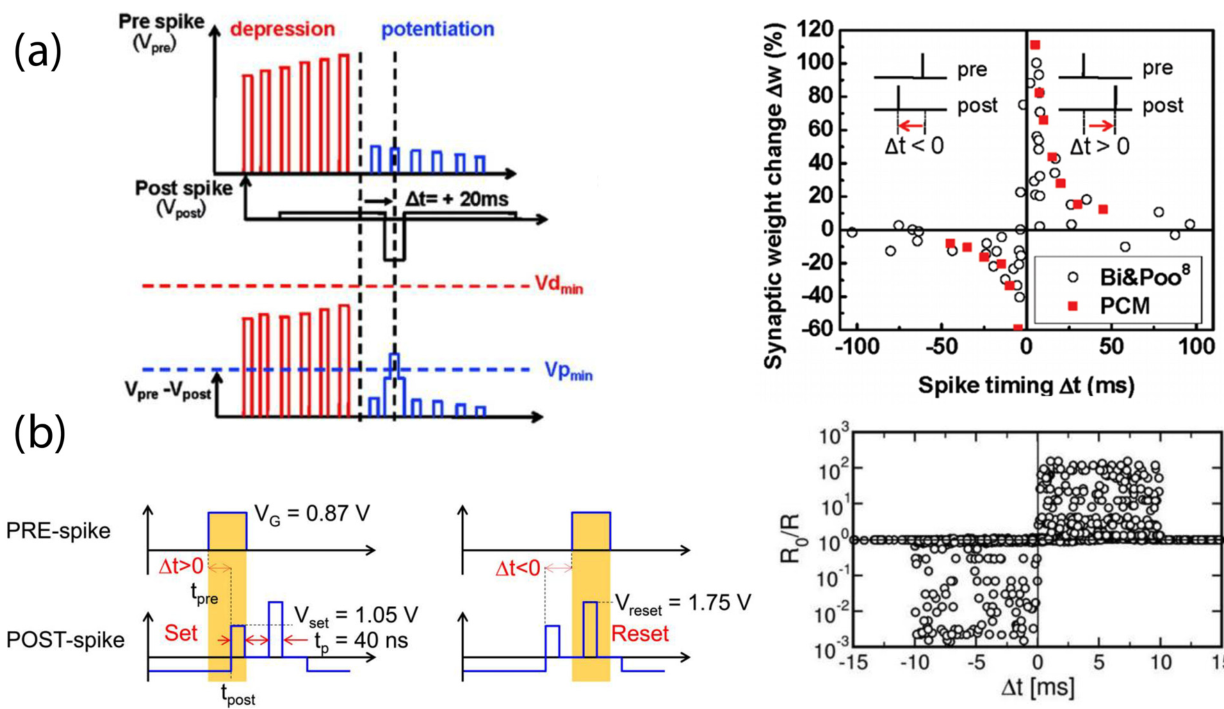


FIG. 7. (a) STDP learning in PCM synapses[34] by a series of pulses of increasing (decreasing) amplitude demonstrating LTP behavior (left) similar to neuroscientific experiments[19] (right). Reprinted with permission from Kuzum *et al.*, Nano Lett. **12**(5), 2179–2186 (2012). Copyright 2012 American Chemical Society. (b) STDP learning effected due to overlap of appropriately shaped pulses.[33] Reprinted with permission from Ambrogio *et al.*, Front. Neurosci. **10**, 56 (2016). Copyright 2016 Author(s), licensed under a Creative Commons Attribution (CC BY) license.

in the optical response of PCM elements by modulating the refractive index can be achieved through varying the number of programming pulses. This has been exploited to experimentally demonstrate unsupervised STDP learning in photonic synapses.[39] To scale beyond single devices, the rectangular waveguides used in this work can be replaced with microring resonators to perform unsupervised learning in an atemporal fashion.[40]

3. PCM crossbars

We have thus far talked about isolated PCM devices mimicking the neuronal and synaptic behaviors. Interestingly, these devices can be connected in an integrated arrangement to perform in-memory computations involving a series of multiply and-accumulate (MAC) operations. Such operations can be broadly represented as a multiplication operation between an input vector and the synaptic weight matrix, which is key to many neural computations. Vector–matrix multiplication (VMM) operations require multiple cycles in a standard von-Neumann computer. Interestingly, arranging PCM devices in a crossbar fashion (or in more general terms, arranging resistive memories in a crossbar fashion) can engender a new, massively parallel paradigm of computing. VMM operation, which is otherwise a fairly cumbersome operation, can be performed organically through the application of Kirchoff's laws as follows. This can be understood through Fig. 8, where each PCM device encodes the synaptic strength in the form of its conductance. The current through each device is proportional to the voltage applied and the conductance of the device. Currents from all the devices in a column get added in accordance with Kirchoff's current law to produce a column-current, which is a result of the dot-product of the voltages and conductance. Such a dot-product operation can be mathematically represented as

$$I_j = \sum_i V_i G_{ij}, \tag{4}$$

where V_i represents the voltage on the i-th row and G_{ij} represents the conductance of the element at the intersection of the i-th row and j-th columns. This ability of parallel computing within the memory array using single-element memory elements capable of packing multiple bits paves the way for faster, energy-efficient, and high-storage neuromorphic systems.

In addition to synaptic computations, PCM crossbars can also be used for on-chip learning that involves dynamic writing into individual devices. However, parallel writing to two-terminal devices in a crossbar is not feasible as the programming current might sneak to undesired cells, resulting in inaccurate conductance updates. To alleviate the concern of *sneak current paths*, two-terminal PCM devices are usually used in conjunction with a transistor or a selector. Such memory cell structures are termed as "1T-1R" or "1S-1R" (shown in Fig. 8) and are extensively used in NVM crossbar arrays. Such 1T-1R crossbar arrays can be seamlessly used for on-line learning schemes such as STDP. To that effect, PCM crossbars were used as one of the first of its kind to experimentally demonstrate on-chip STDP based learning,[41,42] and simple pattern recognition tasks were conducted using the arrays.

Published under license by AIP Publishing

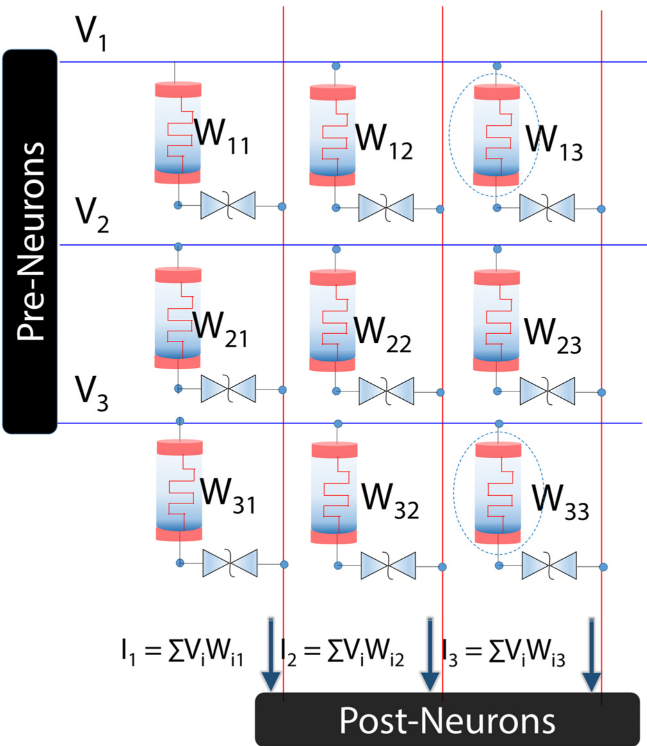


FIG. 8. Synaptic devices arranged in a crossbar fashion along with selector devices to perform dot-product operations. The input voltages are applied to the different rows of the crossbars, and the current from each column represents the dot-product, $I_j = \sum V_i W_{ij}$, between the input voltages and the conductance, W, of the devices.

Although these works focused on smaller scale crossbar arrays of size 10×10, slightly modified 2T-1R memory arrays have also been explored for *in situ* learning on a much bigger scale.[43] Using two transistors enables simultaneous LIF neurons and STDP learning characteristics in an integrated fashion.

We have discussed how unsupervised STDP learning can be implemented using PCM crossbars. However, on-line learning using STDP requires complex programming schemes and is difficult to scale to larger crossbars. On the other hand, networks trained with supervised learning can be mapped on to much larger PCM crossbar arrays for inferencing. These neural networks have been experimentally demonstrated to perform complex image recognition tasks[44,45] with reasonable accuracy. Note that for these works, the supervised learning schemes were implemented with software and the PCM crossbars were used for forward propagation both during training and inferencing.

We have discussed how PCM crossbars leverage Kirchoff's laws to perform neuro-synaptic computations in the electrical domain. In the optical domain, however, the dot-product operation can be implemented using wavelength-division-multiplexing (WDM).[40,46] The input is encoded in terms of different wavelengths, and each synaptic device modulates the input of a particular wavelength. The resulting sum is fed to an array of photo-detectors to realize the dot-product operation.

PCM technology shows remarkable scalability and high-storage density, making them amenable to efficient neuromorphic systems. However, further material and device research is necessary to truly realize the full potential of PCM-based neuromorphic accelerators. First, the most common PCM devices are based on the chalcogenide material group comprising elements Ge, Sb, and Te due to their high optical contrast, repeatability, and low reflectivity. In the GeSbTe system ranging from GeTe to Sb_2Te_3, $Ge_2Sb_2Te_5$ has been identified as the optimum material composition[47,48] based on the trade-offs between stability and switching speed. Despite this development, PCMs suffer from significantly high write power due to their inherent heat dependent switching and high latency. Second, PCM devices suffer from the phenomenon of resistance drift, which is more pronounced for high resistance states (HRSs). The resistance drift is the change in the programmed value of the resistance over time after programming is completed. This has been attributed to structural relaxations occurring shortly after programming.[49–51] The effect of drift on neural computing has been studied, and possible mitigation strategies have been proposed.[52] However, the inability to reliably operate PCM devices at high resistance states has an impact on large-scale crossbar operations. In light of these challenges, it is necessary to investigate newer materials that offer more stability and lower switching speeds for efficient and scalable neuromorphic systems based on PCM devices.

B. Metal-oxide RRAMs and CBRAMs

An alternative class of materials to PCMs for memristive systems are perovskite oxides such as $SrTiO_3$,[53] $SrZrO_3$,[54] $Pr_{0.7}Ca_{0.3}MnO_3$ (PCMO),[55] and binary metal oxides such as HfO_x,[56] TiO_x,[57] and TaO_x,[58] which exhibit resistive switching with lower programming voltages and durations. Such resistive switching is also observed when the oxide is replaced by a conductive element. Two-terminal devices based on these materials form the base of Resistive Random Access Memories (RRAMs). The devices with oxides in the middle are known as metal-oxide RRAMs, whereas the ones with conductive elements are known as the Conductive Bridge RAM (CBRAM). Although the internal physics of these two classes of resistive RAMs is slightly different, both kinds of devices have a similar behavior and hence applicability. In the initial years of research, RRAM was envisaged to be a non-volatile high-density memory system along with CMOS-compatible integration. With significant development over the years, various other applications leverage the non-volatility of RRAMs for power and area-efficient implementations. Among these, neuromorphic computing is a dominant candidate, which exploits the multi-level capability and the analog memory behavior of RRAMs to emulate neuro-synaptic functionalities. In this section, we will discuss how RRAMs can directly mimic neuronal and learning synaptic behaviors using single devices.

1. Metal-oxide RRAMs and CBRAMs as neurons

The dynamics of a voltage driven metal-oxide RRAM device was first investigated by HP labs in their iconic work on TiO_2, which identified the first device[61] showing the characteristics of a memristor, predicted by Chua in 1971.[62] The oxide material can be conceptually split into two regions, a conductive region and an insulating region. The conductance of such a device can be given by its state variable, *w*, which varies as

$$I = g_M(w/L)V(t), \frac{dw}{dt} = f(w(t), V(t)). \quad (5)$$

Interestingly, the RRAM device can be used in an integrator circuit as a resistor in parallel to an external capacitance, as shown in Fig. 9 (top), to emulate the LIF characteristics where the conductance of the device can be used as an internal variable.[59] When the memristor is in its OFF state, the current through the circuit is low, and hence, it does not output a spike. Once the memristor reaches its ON state, the current suddenly jumps, which can be converted to analog spike. The voltage across the memristor, in that case, obeys the dynamics of a LIF neuron, given by Eq. (1) in Sec. II A. A similar neuron circuit has also been explored for CBRAM devices based on $Cu/Ti/Al_2O_3$[60] [Fig. 9 (bottom)]. Unlike PCMs, to emulate the differential equations of the LIF neuron, an *R-C* circuit configuration is used. If the leaky behavior is not required, the internal state of the neuron or the membrane potential can be directly encoded in the oxygen concentration in the device. By manipulating the migration of oxygen vacancies using post-synaptic pulses, IF neurons can be realized by oxide-based devices.[63] To that effect, oxide-based devices have been used to design common neuronal models involving leaky behavior, such as the Hodgkin–Huxley model and leaky IF model.[64]

2. Metal-oxide RRAMs and CBRAMs as synapses

Much like PCM devices, RRAM devices can also be programmed to multiple intermediate states between the two extreme resistance states, which are known as the high resistance state (HRS) and the low resistance state (LRS). This capability of behaving as an analog memory makes RRAMs suitable for mimicking synaptic operations in neural networks. The physics behind emulating such synaptic behavior rests on soft di-electric breakdown in metal-oxide RRAM devices and dissolution of metal ions in CBRAM devices. The device structure for a metal-oxide RRAM is shown in Fig. 10(a). In the case of the metal-oxide RRAM, the switching mechanisms can be categorized as (a) filamentary and (b) non-filamentary. The filamentary switching results due to the formation and rupture of filamentary conductive paths due to thermal redox reactions between metal electrodes and the oxide material. The "forming" or SET process occurs at a high electric field due to the displacement and drift of oxygen atoms from the lattice. These oxygen vacancies form localized conductive filaments, which form the basis of filamentary conduction in RRAM devices. The forming voltage can be reduced by thinning down the oxide layer[65] and controlling annealing temperatures during deposition.[66] The RESET mechanism, on the other hand, is well debated, and ionic migration has been cited as the most probable phenomenon.[67,68] A unified model of RESET proposes that the oxygen ions that drifted to the negative electrode causes the insulator/anode interface to act as a "oxygen reservoir."[69] Oxygen ions diffuse back into the bulk due to a concentration gradient and possibly recombine with the vacancies that form the filament such that material moves back to the HRS. The *I–V* characteristics are shown in Fig. 10(b) where varying SET and RESET pulses lead to different resistance states. In order to emulate synaptic behavior through analog memory states in filamentary RRAMs, various programming techniques have been explored. For example, the SET current compliance can be used to modulate the device resistance by determining the number of conductive filaments. On the other hand, varying the external stimulus can control the degree of oxidation at the electrode and oxide interface, resulting in a gradual change in resistance.[70] These analog states in RRAM devices can be exploited to perform learning on devices using various pulsing techniques. To that effect, the time dependence of synaptic conductance change in STDP learning can be induced by manipulating the shapes of pre-synaptic and post-synaptic voltage waveforms,[71,72] shown in Fig. 11(a). Similar to programming PCM devices, a gradual increase or decrease in conductance can be achieved using a succession of identical pulses as well, as shown in the figure. Such a pulsing scheme, despite requiring a more number of pulses, provides a more granular control over the synaptic conductance,[73,74] shown in Fig. 11(b). Furthermore, adding more peripheral transistors to programming circuits can further enable precise control over STDP. For example, a 2T/1R synapse uses the overlapping window of two different pulses to generate programming current to induce time-dependent LTP and LTD.[75] In the case of filamentary RRAMs, variability in the forming process induces stochasticity in resistive switching, which can be leveraged to design stochastically learning synapses. The switching probability can be

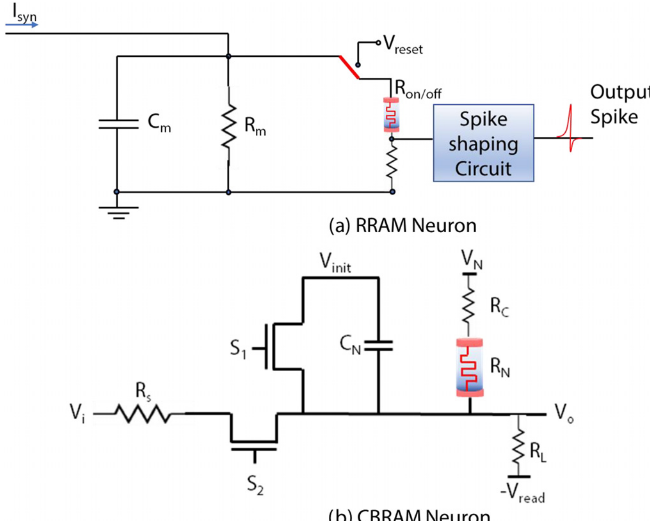


FIG. 9. (a) RRAM[59] and (b) CBRAM[60] neuron circuits showing the memristive device R_N (below) or $R_{ON/OFF}$ (top) in parallel to a capacitor to emulate LIF characteristics.

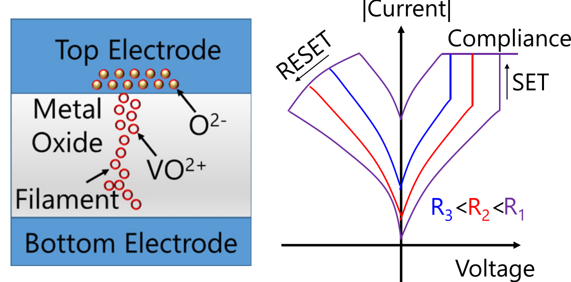


FIG. 10. (a) Basic device structure for RRAM devices consisting of a metal-oxide layer sandwiched between two electrodes. (b) *I–V* characteristics showing varying SET and RESET points, leading to different resistance states.

Published under license by AIP Publishing

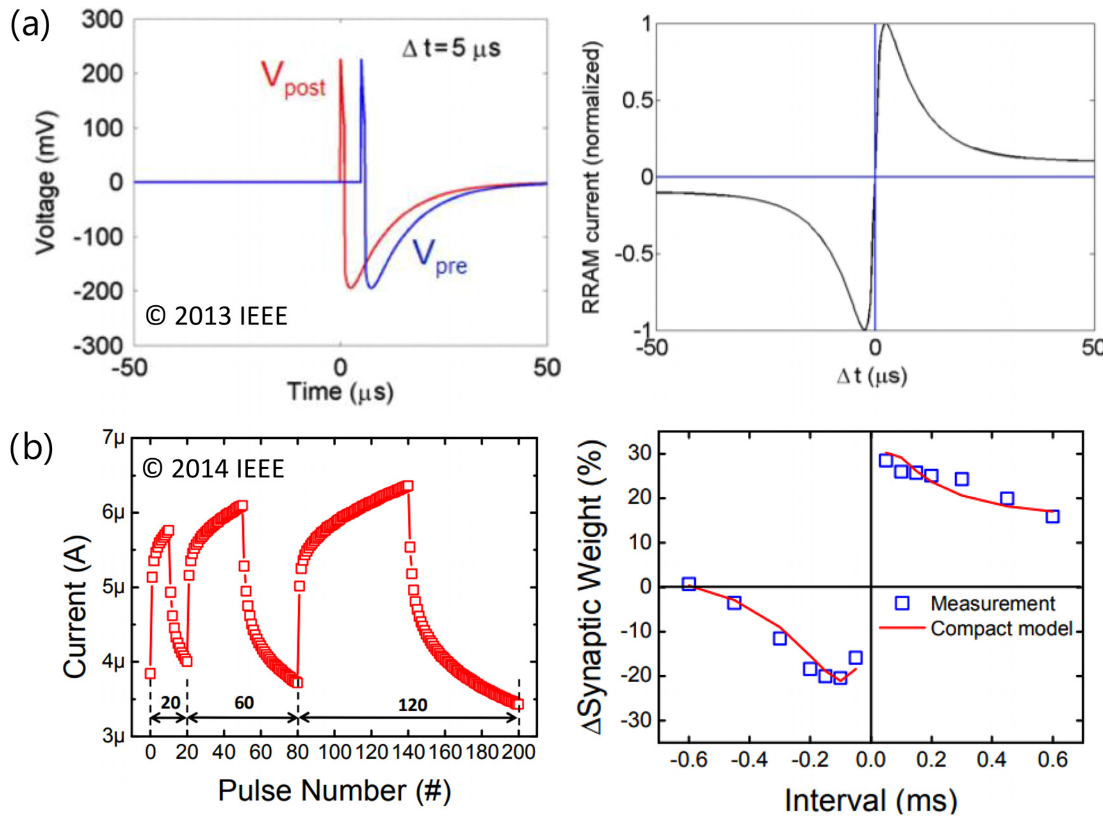


FIG. 11. (a) Appropriately shaped pulses representing the post-synaptic and pre-synaptic potential.[72] The overlap between the two pulses in time leads to STDP learning characteristics in the form of the writing current flowing through the device. Reprinted with permission from Rajendran *et al.*, IEEE Trans. Electron Devices **60**(1), 246–253 (2012). Copyright 2013 IEEE. (b) STDP characteristics can also be emulated by passing multiple pulses, repetitively.[74] Reprinted with permission from Wang *et al.*, in *2014 IEEE International Electron Devices Meeting* (IEEE, 2014), p. 28. Copyright 2014 IEEE.

controlled by using a higher pulse amplitude. Stochastic synapses have the ability to encode information in the form of probability, thus achieving significant compression over deterministic counterparts. Learning stochastically using binary synapses has been demonstrated to achieve pattern learning.[76] Unsupervised learning using multi-state memristors can also be performed probabilistically to yield robust learning against corrupted input data.[77]

Oxides of some transition metals, such as $Pr_{0.7}Ca_{0.3}MnO_3$ (PCMO), exhibit non-filamentary switching as well. This type of switching, on the other hand, results from several possible phenomena such as charge-trapping or defect migration at the interface of metal and oxide, which end up modulating the electrostatic or Schottky barrier. Although the switching physics in non-filamentary RRAM devices is different from that in filamentary RRAMs, the fundamental behavior of using these RRAM devices as synapses is quite similar. Non-filamentary RRAMs can also be programmed using different voltage pulses to exhibit multi-level synaptic behavior. Moreover, varying pulse widths can instantiate partial SET/RESET characteristics, which have been used to implement STDP characteristics in RRAM synapses.[78,79] By encoding the conductance change using the number of pulses coupled with appropriate waveform engineering can enable various kinds of STDP behaviors, explained in Sec. II B, of isolated RRAM devices showing non-filamentary switching.[80] In addition to long-term learning methods, RRAM devices with controllable volatility can also be used to mimic frequency dependent learning, thus enabling a transition from short-term to long-term memory.[81] By controlling the frequency and amplitude of the incoming pulses, STP-LTP characteristics have been achieved in WO_3 based RRAM synapses.[82] In general, higher amplitude pulses in quick succession are required to transition the device from decaying weights to a more stable persistent state. Such metastable switching dynamics of RRAM devices have been used to perform spatiotemporal computation on correlated patterns.[83]

Thus far, we have discussed how metal-oxide RRAM devices can emulate synaptic behavior. Next, we will discuss CBRAM devices, which also exhibit similar switching behavior by just replacing the oxide material with an electrolyte. The switching mechanism is analogous to filamentary RRAM except that the filament results in a metallic conductive path due to electro-chemical reactions. This technology has garnered interest due to its fast and low-power switching. Most CBRAM devices are based on Ag electrodes where resistive switching behavior is exhibited due to the contrast in conductivity in Ag-rich and Ag-poor regions. The effective conductance of such a device can be written as[88]

Published under license by AIP Publishing

$$G_{eff} = \frac{1}{R_{ON}w + R_{OFF}(1-w)}, \quad (6)$$

where w defines the normalized position of the end of the conducting region at the interface of Ag-rich and Ag-poor regions. The conductance of such a device can also be gradually manipulated to implement STDP using a succession of pulses.[88] Here, the exponential dependence on spike timing is implemented using time-division multiplexing where the timing information is encoded in the pulse width. CBRAM based STDP learning has been implemented on-chip using CMOS integrate-and-fire neurons.[89] As with filamentary RRAM devices, stochastic behavior in CBRAM devices can also enable low-power probabilistic learning. One such implementation uses the recency of spiking as a measure of manipulating the probability of the device for visual and auditory processing.[90] Some CBRAM devices also exhibit decay in conductance, which can be leveraged to implement short-term plasticity. Ag_2 S based synapses also show the properties of sensory memory, wherein conductance does not change for some time, before exhibiting STP.[91]

3. Metal-oxide RRAM and CBRAM crossbars

RRAMs are two-terminal devices, similar to PCMs. Hence, like PCMs, RRAM devices can also be arranged into large-scale resistive crossbars, shown in Fig. 8, for building neuromorphic systems. RRAM crossbar arrays can be integrated seamlessly with CMOS circuits for hybrid storage and neuromorphic systems. To that effect, a 40×40 array with CMOS peripheral circuits has been demonstrated to reliably store complex bitmap images.[92] Such an experimental demonstration is a testimony to the scalability of RRAM crossbars. Leveraging this scalability, studies have proposed RRAM crossbar arrays to perform *in situ* learning in single layer neural networks.[93,94] This scalability has been corroborated by the recent development in the process technology, which have led to the realization of large crossbars of sizes up to 128×128 to perform image processing tasks[95] and *in situ* learning for multi-layer networks.[45] The aforementioned works focus on using RRAM as an analog memory. To achieve more stability, RRAM crossbar arrays have also been used as binary weights in a scalable and parallel architecture[85] to emulate a large-scale XNOR network.[96] Both PCM and RRAM crossbars have been extensively explored at an array-level, and Table I provides a comparative study of different experimental demonstrations. It should be understood that large-scale RRAM crossbars have been primarily explored for non-spiking type networks; however, the compute primitives can be easily ported to realize spike-based computing. We will later discuss NVM architectures based on these RRAM crossbars, which show immense potential to achieve energy-efficiency and high density compared to standard CMOS-based computing.

Thus far, we have discussed metal-oxide RRAM crossbar arrays. From a scalability point of view, CBRAM crossbars exhibit similar trends. To that effect, high-density 32×32 crossbar arrays based on Ag–Si systems have been experimentally demonstrated, which can be potentially used to build neuromorphic circuits. Simulation studies based on such Ag–Si systems show significant potential of using large-scale crossbars for image classification tasks.[97]

Of the two classes of materials belonging to the RRAM family, metal-oxide RRAM devices have been more dominantly explored in the context of developing large-scale neuromorphic circuits. However, despite significant progress, RRAM-based devices suffer from significant variability, particularly in the filament formation process. On the other hand, non-filamentary RRAM devices, being barrier-dependent, may lead to trade-offs between stability and programming speed. Overall, further material research is crucial toward making RRAMs viable for large-scale neuromorphic systems.

C. Spintronic devices

Akin to other non-volatile technologies, spin based devices were conventionally investigated as a non-volatile replacement for the existing silicon memories. What makes spin devices particularly unique as compared to other non-volatile technologies is their almost unlimited endurance and fast switching speeds. It is therefore not surprising that among various non-volatile technologies, spin devices are the only ones that have been investigated and have shown promise as on-chip cache replacement.[98] With respect to neuromorphic computing, it is the rich device physics and spin dynamics that allow efficient mapping of various aspects of neurons and synapses into a single device. As we will discuss in this section, spintronics brings in an alternate paradigm in computing by using electron spin as the memory storage variable. The fact that spin dynamics can be controlled by multiple physics including current induced torques,[99] domain wall motion,[100] voltage based spin manipulation,[101] and elastic coupling adds to the rich device possibilities with spintronics and their applications to neuromorphic computing. In this section, we would describe key representative works with spin devices showing their applicability as IF-, LIF-, and stochastic neurons, and synaptic primitives.

1. Spin devices as neurons

As mentioned earlier, it is the rich spin dynamics that allows mapping of different aspects of biological neurons using a single device. In fact, the simplest and the most well-known spin device—the two-terminal Magnetic Tunnel Junction (MTJ)—can be seen as a

TABLE I. NVM Technologies.

Technology	PCM[45]	RRAM[84]	RRAM[85]	RRAM[86]	RRAM[87]
Crossbar size	512×512	108×54	128×128	128×16	512×512
ON/OFF ratio	10	5	N/A	10	N/A
Area per operation (μm^2)	22.12	24	0.05	31.15	N/A
Latency (ns)	80	10	13.7	0.6	9.8
Energy-efficiency (TOPS/W)	28	1.37	141	11	121.38

Published under license by AIP Publishing

stochastic-LIF neuron. MTJs are composed of two ferromagnetic (FM) nanomagnets sandwiching a spacer layer[105] as shown in Fig. 12(a). Nanomagnets encode information in the form of the direction of magnetization and can be engineered to stabilize in two opposite directions. The relative direction of the two FMs—parallel (P) vs antiparallel (AP)—results in two distinct resistive states—LOW vs HIGH resistance. Switching the MTJ from the P to the AP state or vice versa can be achieved by passing a current through the MTJ, resulting in transfer of torque from the incoming spins to the FMs. Interestingly, the dynamics of the spin under excitation from a current induced torque can be looked upon as a stochastic-LIF dynamics. Mathematically, the spin dynamics of an FM, shown in Fig. 12(b), can be expressed effectively using the stochastic-Landau–Lifshitz–Gilbert–Slonczewski (s-LLGS) equation,

$$\frac{\partial \hat{m}}{\partial t} = -|\gamma|(\hat{m} \times H_{EFF}) + \alpha\left(\hat{m} \times \frac{\partial \hat{m}}{\partial t}\right) + \frac{1}{qN_s}(\hat{m} \times I_s \times \hat{m})$$

$$\frac{1+\alpha^2}{\gamma}\left(\frac{\partial \hat{m}}{\partial t}\right) = -(\hat{m} \times H_{EFF}) + \alpha(\hat{m} \times \hat{m} \times H_{EFF}) + \frac{1}{qN_s}(\hat{m} \times I_s \times \hat{m}) \quad (7)$$

where $\hat{m}$ is the unit vector of free layer magnetization, γ is the gyromagnetic ratio for the electron, α is Gilbert's damping ratio, and H_{EFF}

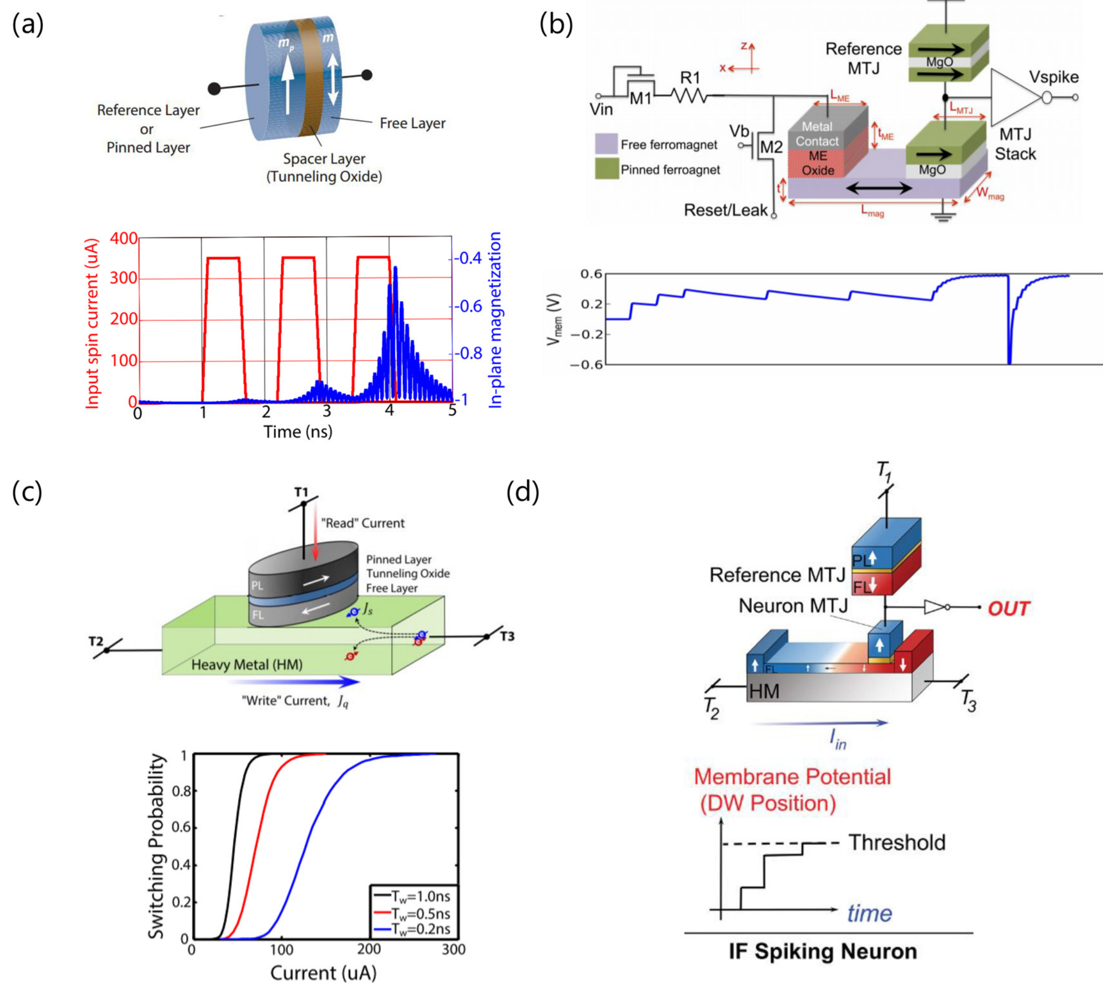


FIG. 12. (a) MTJ-based neuron[102] showing the device structure (top) and leaky-integrate characteristics (bottom). Sengupta *et al.*, Sci. Rep. **6**, 30039 (2016). Copyright 2016 Author(s), licensed under a Creative Commons Attribution (CC BY) license. The magnetization of the free layer of the MTJ integrates under the influence of incoming current pulses. (b) ME oxide-based LIF neuron[103] showing the device structure (top) and LIF characteristics (bottom). Reproduced with permission from Jaiswal *et al.*, IEEE Trans. Electron Devices **64**(4), 1818–1824 (2017). Copyright 2017 IEEE. (c) SHE-MTJ-based stochastic neuron[102] showing the device structure (top) and the stochastic switching characteristics (bottom). Reprinted with permission from Sengupta *et al.*, Sci. Rep., **6**, 30039 (2016); Copyright 2016 Author(s), licensed under a Creative Commons Attribution (CC BY) license. (d) DWM-based IF spiking neuron[104] showing the device structure (top) and integration and firing behavior (bottom) over time. For incident input spikes, the domain wall moves toward the MTJ at the end, thus decreasing the resistance of the device. When the domain wall is at the end, the resistance reaches its lowest, enough for the neuron fires. Reproduced with permission from Sengupta and Roy, Appl. Phys. Rev. **4**(4), 041105 (2017). Copyright 2017 AIP Publishing.

is the effective magnetic field including the shape anisotropy field, external field, and thermal field. This equation bears similarities with the leaky-integrate-and-fire behavior of a neuron. The last term represents the spin transfer torque (STT) phenomenon, which causes the magnetization to rotate by transferring the torque generated through the change in angular momentum of incoming electrons. Interestingly, the first two terms can be related to the "leak" dynamics in an LIF neuron, while the last term relates to the integrating behavior of the neuron as follows. When an input current pulse or "spike" is applied, the magnetization starts integrating or precessing toward the opposite stable magnetization state owing to the STT effect (last term). In the absence of such a spike, the magnetization leaks back toward the original magnetization state (Gilbert damping, second term). Furthermore, due to nano-scale size of the magnet, the switching dynamics is a strong function of a stochastic thermal field, leading to the stochastic behavior. This thermal field can be modeled using Brown's model.[106] In terms of Eq. (7), the thermal field can be incorporated into H_{EFF} as a magnetic field,

$$H_{thermal} = \zeta \sqrt{\frac{2\alpha kT}{|\gamma| M_s V}}, \tag{8}$$

where ζ is a zero mean, unit variance Gaussian random variable, V is the volume of the free layer, T is the temperature, and k is the Boltzmann constant. A typical, stochastic-LIF behavior using MTJ is shown in Fig. 12(a).[102] While the two-terminal MTJ does represent the stochastic-LIF dynamics, the very fact that the leaky and integrate behaviors are controlled by intricate device physics and intrinsic material parameters makes it difficult to control as needed for a large-scale circuit/system implementation. As a result, alternate physics such as the magneto-electric switching (ME) has been proposed as stochastic-LIF neurons, wherein the leaky and integrating behaviors can be easily controlled through device dimensions and associated circuit elements. In ME devices, the voltage induced electric field polarization induces a magnetic field at the interface of the FM and ME oxide, which induces switching of the FM layer. The ME oxide layer acts a capacitor, and a series resistance can enable LIF neuronal dynamics in such a device. The ME switching process is susceptible to noise like the conventional MTJ switching and hence inherently mimics the stochastic dynamics with the LIF behavior.[103] By controlling the ME oxide dimension in Fig. 12(b) and/or the leaky resistive path, the LIF dynamics can be easily tweaked as per requirement. In essence, we have seen that both current induced MTJ and voltage driven ME switching can act as stochastic-LIF neurons. However, on one hand, current based MTJ is difficult to control, while on the other hand, ME switching is still in its nascent stage of investigation and needs extensive material research for bringing the device to mainstream applications.

Alternatively, at the cost of reduced dynamics, three terminal Spin-Orbit-Torque MTJ (SOT-MTJ) has been used as a reliable stochastic spiking neuron while neglecting the leaky-integrate dynamics.[102] SOT-MTJ is reasonably mature, and also its three terminal nature brings in attractive circuit implications. First, SOT-MTJ is switched by passing a bi-directional current through a heavy-metal (HM) layer, as shown in Fig. 12(c). When a charge current enters the HM, electrons of opposite spins get scattered to the opposite sides of the layer, and a spin-polarized current is generated, which rotates the magnetization in the adjacent MTJ such that the switching probability increases as the magnitude of the input current is increased. This in turn implies that the incoming current passes through a much lower metal resistance and sees a constant metal resistance throughout the switching process as opposed to current based switching in conventional two-terminal MTJs. As we will see later, the existence of a low input resistance for the neuron allows easy interfacing with synaptic crossbar arrays. Second, the decoupled read-write path in SOT-MTJs allows for independent optimization of the read (inferencing) and write (learning) paths. A typical SOT-MTJ and its sigmoid-like stochastic switching behavior are shown in Fig. 12(c). While the aforementioned behaviors depicted in Fig. 12(c) correspond to an SOT-MTJ with a high energy-barrier (10–60 kT), telegraphic SOT-MTJ with an energy-barrier as low as 1 kT has also been explored as stochastic neurons.[107]

In addition to smaller magnets, wherein the entire magnet switches like a giant spin, longer magnets known as domain wall magnets (DWMs)[108] have been used as IF neurons. DWMs consist of two oppositely directed magnetic domains separated by a domain wall [see Fig. 12(d)]. Electrons flowing through the DWM continuously exchange angular momentum with the local magnetic moment. Current induced toque affects the misaligned neighboring moments around the domain wall region, thus displacing the domain wall along the direction of current flow. The instantaneous membrane potential is encoded in the position of the domain wall, which moves under the influence of post-synaptic input current. The direction of movement is determined by polarity of the incident current. The resulting magnetic polarity can be sensed by stacking a MTJ at an extremity of the DWM, and subsequent thresholding is performed when the domain wall reaches that extremity. The leak functionality in such a neuron can be implemented by passing a controlled current in the opposite direction. A constant current driven leak would result in increased energy consumption; as such, voltage driven DWMs based on elastic coupling can be used to reduce the energy consumption.[109] However, a concern with DWM-based neuromorphic devices is that the motion of domain walls might be pinned by the presence of defects.[110] To that effect, magnetic skyrmions promise enhanced stability and has been explored in the context of emulating neuromorphic behavior.[111] In summary, we have described multiple devices and their physics and extent of bio-fidelity, wherein spin is used as the basic state variable. Let us now consider the applicability of spin devices as synaptic elements (Fig. 13).

2. Spin devices as synapses

Recall that, for PCM and RRAM devices, the existence of multiple non-volatile resistance states between the two extreme HIGH and LOW resistances makes them suitable as synaptic elements. On similar lines, spin devices can be engineered to enable a continuous analog resistive stable state between its AP (HIGH) and P (LOW) resistances. This is achieved by stacking an MTJ over DWMs. The position of the domain wall determines the resistance state of the device. In extreme cases, the magnetization direction of the entire DWM aligns with that of the pinned layer, resulting in a LOW resistance state of the device, shown in Fig. 13. Conversely, the magnetization direction of the DWM in the opposite direction to that of the pinned layer leads to an

Published under license by AIP Publishing

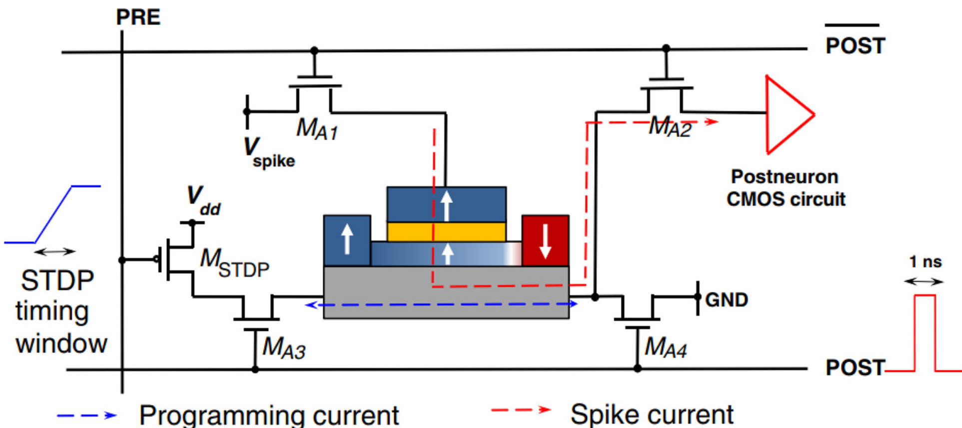


FIG. 13. STDP learning scheme in the DWM-based spin synapse[112] using peripheral transistors. The exponential characteristics of STDP are realized by operating M_{STDP} in the sub-threshold region and applying a linearly increasing voltage at its gate. M_{STDP} is activated when a pre-neuron spikes, and the programming current (shown in blue) through the transistor is injected into the HM layer (grey) when a post-neuron spikes. Reproduced with permission from Sengupta *et al.*, Phys. Rev. Appl. **6**(6), 064003 (2016). Copyright 2017 American Physical Society.

Anti-Parallel (AP) configuration, which defines the HIGH resistance state of the device. With respect to the position of the domain wall, x, the resistance of the device changes as

$$G_{eq} = G_P \frac{x}{L} + G_{AP} \frac{L-x}{L} + G_{DW}. \tag{9}$$

Here, G_P (G_{AP}) is the conductance of the MTJ when the domain wall is at the extreme right (left) of the DWM. G_{DW} is the conductance of the domain wall region, and L is the length of the DWM. Owing to low write currents, synaptic elements based on DWM devices[113] can achieve orders of magnitude lower energy consumption over corresponding realizations in other non-volatile technologies. Similar to spin neurons, inducing switching using the Spin-Hall effect (SHE) through a heavy-metal below the MTJ, the programming current can be further reduced. DWM-based devices have been explored to mimic the behavior of multi-level synapses in works such as Ref. 114. With a few extra transistors, STDP learning can be enabled by a relatively simple programming scheme as shown in Fig. 13.[112] This scheme leverages the exponential characteristics of transistors in the sub-threshold regime. A linearly increasing voltage is applied to the gate of the transistor, M_{STDP}, which is activated when the pre-neuron spikes. When the post-neuron fires, an appropriate programming current passes through the HM layer, which now depends exponentially to the timing difference due to the sub-threshold operation. It is worth noting that although the DWM provides a way to encode multiple stable states in spin devices, the key drawback of such devices is the extremely limited HIGH–LOW resistance range. The resistance range for spin devices is much lower than their PCM and RRAM counterparts. Encoding multiple states within the constrained resistance range raised functionality concerns considering variability.

Alternatively, non-domain wall devices such as two-terminal MTJs of three terminal SHE based MTJs can be used as synapses. In the absence of DWMs, MTJs can only encode binary information, i.e., two resistance states. In such a scenario, stochasticity can play an interesting role in realizing multi-level behavior by probabilistic switching. In spin devices, such thermally induced stochasticity can be effectively controlled by varying the amplitude or duration of the programming pulse as shown in Fig. 12(c). This benefit of controlled stochasticity leads to energy-efficient learning in binary synapses implemented using MTJs.[115,116] An advantage of on-chip stochastic learning is that the operating currents are lower than the critical current for switching, thus ensuring low-power operations. Such multiple stochastic MTJs can be represented as a single synapse to achieve an analog weight spectrum.[117] These proposals of stochastic synapses based on MTJs have shown applications of pattern recognition tasks on a handwritten digit dataset.

Finally, the precessional switching in the free FM layer in the MTJ inherently represents a dependence of switching on the frequency on programming inputs. On the incidence of a pulse, the magnetization of the free FM layer moves toward the opposite stable state. However, if the pulse is removed before the switching is completed, it reverts back to its original stable state. These characteristics can be used to represent volatile synaptic learning in the form of STP-LTP dynamics.[118]

3. Spintronic crossbars

Synapses based on 2-terminal MTJs can be arranged in a crossbar fashion, similar to other memristive technologies. The currents flowing through the MTJs of each column get added in the crossbar and represent the weighted sum of the inputs. Unlike the two-terminal devices, SHE based MTJs, being 3-terminal devices, have decoupled read and write paths. As a result, they require a modified crossbar arrangement. One major advantage of spin neurons is that current through the synaptic crossbars can be directly fed to the current controlled spin neurons. As discussed earlier, spin devices suffer from very low ON/OFF resistance ratios compared to other technologies. Hence, despite experimental demonstration of isolated synaptic spin devices,[119] large-scale crossbar-level neuromorphic implementations have been mostly limited to simulation studies. Such simulation studies have been based on reasonable ON/OFF ratios considering a predictive roadmap.[120] To that effect, multi-level DWM-based synapses have been arranged in a

Published under license by AIP Publishing

crossbar fashion to emulate large-scale neural networks, both in a fully connected form[114] and as convolutional networks.[121] In addition to inferencing frameworks based on spin synapses, STDP based learning[112] has also been explored at an array-level, as shown in Fig. 14, to perform feature recognition and image classification tasks. As discussed earlier, MTJ-based binary synapses require stochasticity for effective learning. They can leverage the inherent stochasticity in the network, and a population of such synapses can perform on-line learning, which not only achieves energy-efficiency but also enables extremely compressed networks.[116]

These simulation-based designs and results show significant promise for spin based neuromorphic systems. However, several technological challenges need to be overcome to realize large-scale systems with spin. As alluded to earlier, the ON/OFF ratio between the two extreme resistance states is governed by the TMR of the MTJ, which has been experimentally demonstrated to reach 600% (Ref. 122), leading to an ON/OFF ratio of 7. This is significantly lower than other competitive technologies and poses a limitation on the range of synaptic weight representation at an array level. Second, MTJs can only represent binary information. For multi-bit representation, it is necessary to use domain wall devices or multiple binary MTJs at the cost of area density. However, since synapses in the neural networks usually encode information in an analog fashion, the lack of multi-state representation in MTJs can potentially limit the area-efficiency of non-volatile spin devices for neuromorphic applications. The lack of multi-bit precision can be alleviated with architectural design facets such as "bit-slicing." This involves multiple crossbars with binary devices to represent multiple bits of storage. Despite such provisions, improved sensing circuits along with material exploration to achieve higher TMR is necessary to truly realize the potential of spin devices as a viable option to emulate synaptic behavior for large-scale neuromorphic systems.

D. Ferroelectric FETs

Similar to the phase change and ferromagnetic materials, another member of functional material family is ferroelectric (FE) materials. In addition to being electrically insulating, ferroelectric materials exhibit non-zero spontaneous polarization (P), even in the absence of an applied electric field (E). By applying an external electric field (more than a threshold value, called the coercive field), the polarization direction can be reversed. Such an electric field driven polarization switching behavior of FE is highly non-linear (compared to di-electric materials) and exhibits non-volatile hysteretic characteristics. Due to the inherent non-volatile nature, FE based capacitors have been historically investigated for non-volatile memory elements. However, in ferroelectric field effect transistors (FEFETs), an FE layer is integrated at the gate stack of a standard transistor and thus offers all the benefits of CMOS technology in addition to several unique features offered by FE. The FE layer electrostatically couples the underlying transistor. Due to such coupling, FEFETs offer non-volatile memory states by virtue of polarization retention of FE. Beside CMOS process compatibility, one of the most appealing features of FEFET based memory is the ability of voltage based READ/WRITE operation, which is unlike the current based READ/WRITE schemes in other non-volatile memory devices (spin based memory and phase change memory). Due to the non-volatility and the intricate polarization switching dynamics of FE, FEFETs have garnered immense interest in recent times as a potential candidate for neuron-mimicking and multi-bit synaptic devices. In

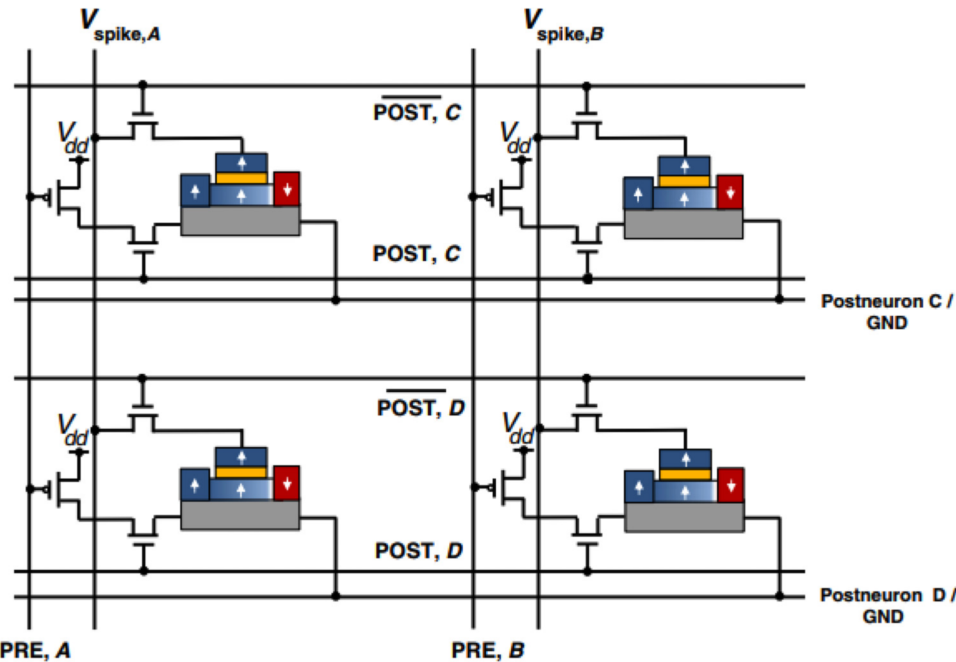


FIG. 14. A crossbar arrangement of spintronic synapses connected between pre-neurons A and B and post-neurons C and D, showing peripheral circuits for enabling STDP learning.[112] Reproduced with permission from Sengupta *et al.*, Phys. Rev. Appl. **6**(6), 064003 (2016). Copyright 2017 American Physical Society.

Published under license by AIP Publishing

this section, we will briefly discuss the recent progress in FEFET based neuro-mimetic devices.

1. FEFETs as neurons

The dynamics in a ferroelectric FET device can be used to mimic the functionality of a biological neuron. In a scaled FEFET, if identical sub-threshold pulses ("sub-coercive" in the context of FE) are applied at the gate terminal [shown in Fig. 15(a) (leftmost)], the device remains in the OFF state (since the sub-threshold pulses are insufficient for polarization switching). However, after a certain number of pulses are received, the FEFET abruptly switches to the highly conductive state [Fig. 15(a) (rightmost)]. Such phenomena can be understood as the initial nucleation of nano-domains followed by an abrupt polarization reversal of the entire grain connecting the source and drain of FEFETs. Before the critical threshold is reached, the nucleated nano-domains are not capable of inducing a significant charge inversion in the channel, leading to the absence of the conduction path (OFF state). The accumulative P-switching presented in Ref. 125 appears to be invariant with respect to the time difference between the consecutive excitation pulses, and therefore, the device acts as an integrator. Moreover, the firing dynamics of such FEFET based neurons can be tuned by modulating the amplitude and duration of the voltage pulse.[123,125] However, to implement the leaky behavior, a proposed option is to modulate the depolarization field or insertion of a negative inhibit voltage in the intervals between consecutive excitation pulses. Apart from this externally emulated leaky process, an intrinsically leaky (or spontaneous polarization relaxation) process has been theoretically predicted in Ref. 126. Such spontaneous polarization relaxation has been attributed as the cause of domain wall instability,[126] and such a process has recently been experimentally demonstrated in an $Hf_xZr_{1-x}O_2$ (HZO) thin-film.[127] By harnessing such a quasi-leaky behavior along with the accumulative and abrupt polarization switching in FE, a quasi-leaky-integration-fire (QLIF) type FEFET based neuron can offer an intrinsic homeostatic plasticity. Network level simulations utilizing the QLIF neuron showed a 2.3× reduction in the firing rate compared to the traditional LIF neuron while maintaining the accuracy of 84%–85% across varying network sizes.[127] Such an energy-efficient spiking neuron can potentially enable ultra-low-power data processing in energy constrained environments.

2. FEFETs as synapses

We have seen how the switching behavior of a FEFET can mimic the behavior of a biological neuron. The switching behavior also produces bi-stability in FEFETs, which makes them particularly suitable for synaptic operations. The bi-stable nature of spontaneous polarization of ferroelectric materials causes voltage induced polarization switching characteristics to be intrinsically hysteretic. The device structure of a FEFET based synapse is similar to a neuronal device as shown in Fig. 15(b) (leftmost). The FE electrostatically couples with the underlying transistor. Due to such coupling, FEFETs offer non-volatile memory states by virtue of polarization retention of the ferroelectric (FE) material. In a mono-domain FE (where the FE area is comparable to the domain size), two stable polarization states (−P and +P) can be achieved in the FE layer, which, in turn, yield two different channel

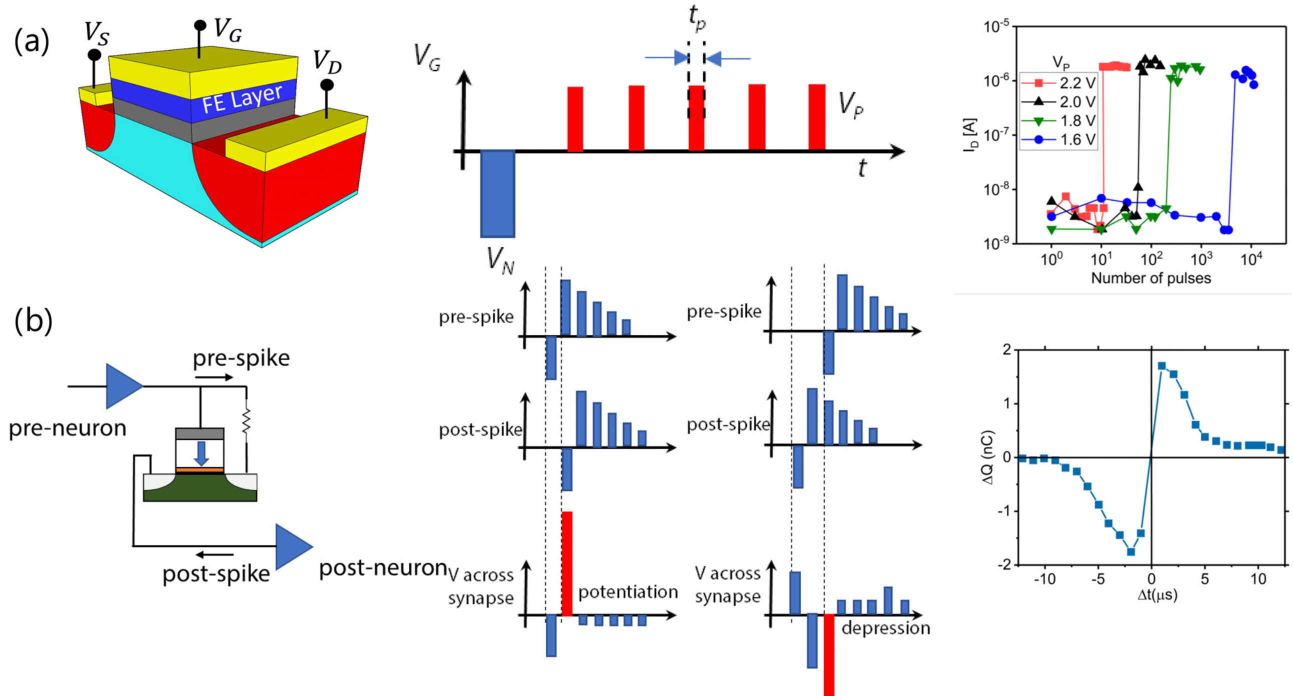


FIG. 15. (a) FEFET device structure showing an integrated ferroelectric layer in the gate stack of the transistor (leftmost). A series of pulses can be applied to emulate the integrating behavior of neurons and the eventual firing through abrupt switching of the device.[123] (b) A FEFET synaptic device (leftmost) showing programming pulsing schemes generating the STDP learning curve based on the change in charge stored in the device.[124]

Published under license by AIP Publishing

conductances for the underlying transistor. Such states can also be referred to as "low V_T" (corresponds to +P) and "high V_T" (corresponds to −P) states.[128] Even though the polarization at the lattice level (microscopic polarization) can have two values (+P or −P), in a macroscopic scenario, multi-domain nature of FE films (with the area significantly higher than the domain size), multiple levels of polarization can be achieved. Furthermore, the polycrystalline nature of the FE film offers a distribution in the polarization switching voltages (coercive voltage) and time (nucleation time) in different grains. As a result, a voltage pulse dependent polarization tuning can be obtained such that the overall polarization of the FE film can be gradually switched. This corresponds to a gradual tuning of channel conductivity (or V_T) in FEFETs and can be readily exploited to mimic multi-level synapses,[124,129] in a manner similar to what has already been reported for PCM and RRAMs. As noted above, FEFETs are highly CMOS compatible, which makes their applications as neuro-mimetic devices quite appealing.

Recently, several FEFET based analog synaptic devices have been experimentally demonstrated,[124,130,131] where the conductance potentiation and depression via a gradual V_T tuning were obtained by applying a voltage pulse at the gate terminal. However, in the case of identical voltage pulses, the observed potentiation and depression characteristics are highly non-linear and asymmetric with respect to the number of pulses. To overcome such non-ideal effects, different non-identical pulsing schemes were proposed in Ref. 130, which utilize a gradual modulation of pulse magnitude or pulse time. Such non-identical pulsing schemes demonstrate a significant improvement in potentiation/depression linearity and symmetry. However, if pulses are not identical throughout the programming process, an additional step of accessing the weight value is needed every time, and an update takes place so that an appropriate pulse can be applied. This leads to design overheads and may reduce the training efficiency. To overcome such detrimental effects, an optimum weight update scheme using identical pulses for improved linearity and asymmetry was experimentally demonstrated in a FE-Germanium-NanoWire-FET (FE-GNWFET).[131] Based on the experimentally extracted parameters of the FE-GNWFET, simulation of the multi-layer perceptron neural network over 1×10^6 MNIST images predicts an on-line learning accuracy of ∼88%. It should be noted that the underlying physics in potentiation/depression linearity and symmetry enhancement in FE-GNWFETs over the conventional FEFET is still unclear. Hence, there is a timely demand for further theoretical understanding that can enable aggressive device level engineering for achieving higher linearity and symmetry in FEFET based synaptic devices.

FEFET synapses can also be used to enable learning with the STDP based update scheme, which can also be achieved.[124] In order to utilize the single FEFET as a two-terminal synapse connected to the pre- and the post-neuron, a resistor is connected between the gate and drain [Fig. 15(b) (leftmost)] terminals. Thus, the pre-spike is applied to the gate and resistor, while the source and bulk are controlled by the post-neuron. With this synaptic scheme and the spiking waveform depicted in Fig. 15(b) (middle), the relative spike timing between the pre- and the post-neurons can be converted into voltage-drop across the FEFET. The closer the spiking in the time domain, the larger the voltage-drop, which induces a larger conductivity change in the FEFET. The corresponding STDP pattern showing the potentiation and depression is depicted in Fig. 15(b) (rightmost).

3. FEFET crossbars

FEFETs utilize the electric field driven writing scheme, and such a feature is unique when compared with the Spin-, PCM-, and RRAM-based synaptic devices. Therefore, FEFET based synaptic devices are potential candidates for low-power realization of neuro-mimetic hardware. These transistor-like devices can also be arranged in a crossbar fashion to perform dot-product operations. Simulation studies using the population of neuronal and synaptic devices have been studied for image classification tasks.[130–132] We discussed earlier that the multi-state conductance of FEFETs originates from the multi-domain behavior of the FE layer at the gate stack. However, such multi-domain features of FE (domain size and patterns) are highly dependent on the physical properties of FE (i.e., thickness, grain size, etc.).[126] As a consequence, in a FEFET synaptic array, the multi-state behavior of FEFETs may suffer from the variability of the FE layer along with the variation induced by underline transistors. Therefore, large-scale implementation of the synaptic array with identical FEFET characteristics will be challenging, which can potentially be overcome with high quality fabrication of FE films and variation aware designs. Despite the benefits offered by FEFETs, the technology is still at its nascent stage in the context of neuro-mimetic devices, and crossbar-level implementations will be potentially explored in the future.

E. Floating gate devices

Most of the aforementioned non-volatile technologies are based on non-Si platforms requiring effective integration and CMOS compatibility. Si-based non-volatile memories, such as Flash memory, use floating gate devices[134] to store data. These devices have seen considerable commercial use in universal serial bus (USB) flash drives and solid state drives. Owing to their non-volatility, floating gate devices were one of the first devices explored for emulating synaptic behavior in neuromorphic systems. Furthermore, these devices are even more promising because of their standard process technology. In this subsection, we will discuss how neuro-synaptic functionalities can be effectively mimicked using floating gate devices.

1. Floating gate devices as neurons

A floating gate (FG) transistor has the same structure as a conventional MOSFET, except for an additional electrode between the gate and the substrate, called the floating gate, shown in Fig. 16(a). The non-volatility is induced by the charge stored on the floating gate of the transistor. As the charge stored in the floating gate increases, the threshold voltage of the transistor decreases, as shown in Fig. 16(b). This charge storage dynamics can also be leveraged to emulate integrating behavior in a leaky IF neuron.[133] Such a LIF neuron circuit is shown in Fig. 17. Block A shows the integrating circuit where a charge is injected into the floating gate by the pre-synaptic current. This modulates the voltage at the floating gate, V_{FG}, which accounts for the integration. Over time, the charge decays, introducing a leaky behavior. The leak factor is dependent on the tunneling barrier thickness. The balance between charge injection and charge ejection determines the neuron operation. The rest of the circuit performs the thresholding and resetting operation as required by a LIF neuron.

Published under license by AIP Publishing

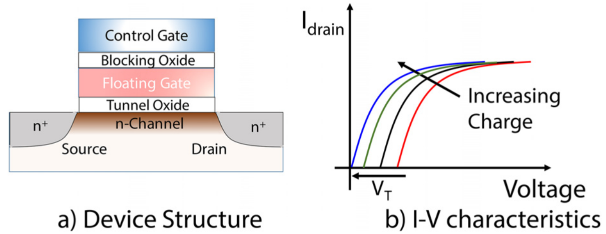


FIG. 16. (a) Basic floating gate transistor structure showing the control gate and the floating gate separated by a blocking oxide layer. (b) Increasing charge in the blocking oxide layer lowers the threshold voltage, V_T, of the transistor causing higher current at a particular voltage.

2. Floating gate devices as synapses

Unlike the neuronal behavior, which depends on the charge injection/ejection dynamics of the floating gate, the synaptic behavior depends primarily on charge storage and its ability to modulate the conductance of the device. The charge storage mechanism is governed by two phenomena known as the Fowler–Nordheim (FN) tunneling[135,136] and hot-electron injection (HEI). HEI requires a high positive voltage across the gate and the source such that electrons have enough kinetic energy to cross the insulating barrier between the floating gate and the channel. Charge gets trapped in the floating gate and remains intact even after removal of voltage due to the excellent insulating abilities of SiO_2. The other mechanism involves FN tunneling, which stores and removes charge from the floating gate in a reversible manner. A sufficiently high positive voltage between the source and control gate causes the electrons to tunnel into the floating gate, whereas an equivalent voltage of opposite polarity removes the charge. Charge in the floating gate increases the threshold voltage of the transistor, thus enabling two stable states in the FG transistor, based on the presence and absence of charge. This can be used to emulate binary synapses. In addition, due to the analog nature of charge, by manipulating the amount of charge stored in the floating gate, multi-level cells (MLC) are possible. Such multi-level storage capability of FG transistors have been heavily used in flash memory technologies.[137,138] This analog memory characteristics along with excellent stability and reliability, especially for multi-level states, make FG devices promising for emulating analog synaptic behavior. In fact, the earliest proposals of on-chip synapses with computing and learning abilities were based on FG transistors.[139–141]

3. Floating gate crossbars

Owing to the integrability with CMOS processes, floating gate transistors have been used to implement large-scale arrays of programmable synapses to perform synaptic computations between populations of neurons. The exponential dependence of injection and tunneling currents on the gate and tunneling voltages can be further used to perform STDP based weight update in such "single transistor" synapses.[142,143]

FG transistors overcome most of the major challenges encountered by the previously discussed non-volatile technologies including reliability and stability. Moreover, the retention time can also be modulated by varying the tunneling barrier of the gate oxide. However, this comes with a trade-off that FG transistors require high voltage for writing and reading. Moreover, unlike the high-density storage offered by PCM and RRAM technologies, FG transistors consume a larger area. The power-hungriness and area inefficiency have thus propelled research toward more energy and area-efficient solutions offered by beyond-CMOS technologies.

F. NVM architecture

So far, we have discussed how NVMs, owing to their intrinsic physics, can be exploited as neural and synaptic primitives. A comparison table of the aforementioned NVM technologies is shown in Fig. 18. Additionally, we have seen that, at a circuit level, the dense crossbar arrangement and associated analog computations present a promising way forward with respect to in-memory computing. Advantageously, beyond devices and circuits, even at an architectural

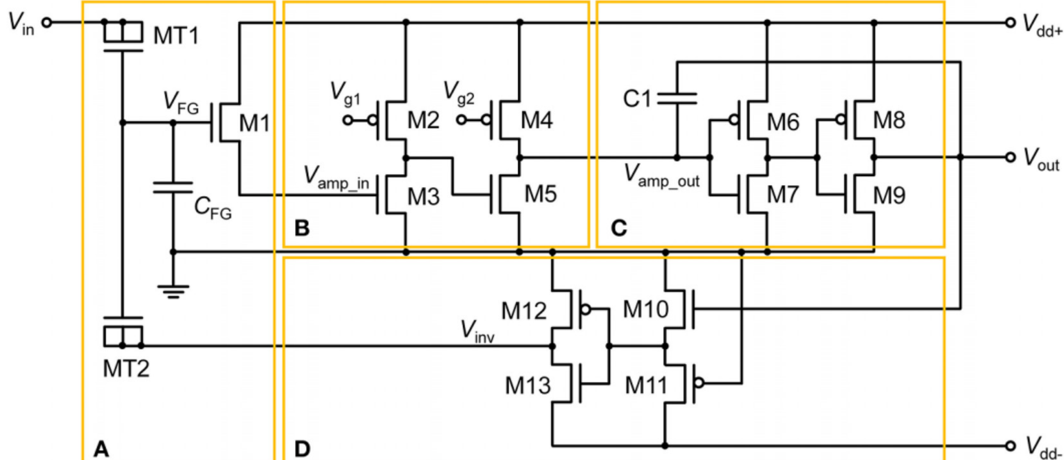


FIG. 17. Floating gate leaky-integrate-and-fire neuron[133] showing (a) the integrating circuit, (b) and (c) feedback amplifier circuits for thresholding operation, and (d) reset circuit.[133]

Technology	Device Structure		Material Stack/geometry	Metrics	Pros/Cons
	Neuron	Synapse			
PCM	Ref 29	Ref 35	**Neuron:** Stack: TEC/GST/BEC Thickness: 100 nm **Synapse:** Stack: TEC/GST/W/BEC Thickness: 100 nm	**Neuron:** Energy per timestep: 5pJ/neuron Pulsewidth: 20-200 ns **Synapse:** Set energy: <1 pJ/PCM Pulsewidth: 50-1000 ns	✓ High Density ✓ High ON/OFF ratio ✓ Scalable ✓ Multi-bit capability × Low endurance × High write latency
RRAM	Ref 63	Ref 79	**Neuron:** Stack: W/PCMO/Pt/Ti PCMO thickness: 65nm **Synapse:** Stack: Pt/AlOx/TiN/PCMO/Pt Device area: 150×150 nm^2	**Neuron:** Write voltage: 2.3-2.6V Pulsewidth: 150-200 ns **Synapse:** Write voltage: −4V to 3V Pulsewidth: 100 ns – 100us	✓ High Density ✓ High ON/OFF ratio ✓ Scalable ✓ Multi-bit capability × Low endurance × High write latency × Prone to device variations
Spin	Ref 99	Ref 112	**Neuron:** Stack: CoFeB/MgO/CoFeB/HM FL thickness: 0.8-1.5 nm MgO thickness: 2 nm Energy Barrier: 10-30kT **Synapse:** Stack: CoFeB/MgO/CoFeB/HM FL thickness: 1.2 nm HM thickness: 2 nm Energy Barrier: 20kT	**Neuron:** Energy per timestep: 1.6 fJ/neuron Pulsewidth: 0.5 ns **Synapse:** Set energy: ~ 38 fJ/synapse Pulsewidth: 1 ns	✓ Low latency ✓ Low energy ✓ High endurance × Low Density × No multi-bit capability × Low ON/OFF ratio
FeFET	Ref 119	Ref 120	**Neuron:** Stack: SiON/HfO_2/TiN HfO_2 thickness: 10 nm Transistor Length: 30 nm Transistor Width: 80 nm **Synapse:** Stack: SiON/HfO_2/TiN Transistor Length: 500 nm Transistor Width: 500 nm	**Neuron:** Programming: Program voltage: 2.2V, -3.25V Pulsewidth: 1 μs **Synapse:** Programming: Program voltage: 3.5V Pulsewidth: 1 μs	✓ Very high ON/OFF ratio ✓ High Endurance ✓ Multi-bit capability × No scalable experimental demonstration for neuromorphic application

FIG. 18. Table showing a comparison of different beyond-CMOS NVM technologies and some representative works on demonstrations and design of neuronal and synaptic elements in a spiking neural network. Note that neurons and synapses can also be designed using non-volatile floating gate transistors (discussed in Sec. III E). However, in this table, we focus on beyond-CMOS materials due to their non-standard material stack.

(or system) level, NVMs and crossbars provide interesting opportunities for energy- and area-efficiency. NVMs provide a radical departure from the state-of-the-art von-Neumann machines due to the following two factors: (1) NVM based crossbars are being looked upon by the research community as the *holy grail* for enabling in-memory massively parallel dot-product operations, and (2) the high storage density offered by NVMs allows construction of *spatial* neuromorphic architectures, leading to higher levels of energy, area, and latency improvements.[144–147] Spatial architectures differ from conventional processors in the sense that the latter rely heavily on various levels of memory hierarchy, and data have to be shuffled back and forth between various memory sub-systems over long distances (between on-chip and off-chip memory). As such, the energy and time spent in getting the data in the right level of memory hierarchy, before it can be processed, lead to the memory-wall bottleneck. Since the storage density of NVMs is much larger [a single static random access memory (SRAM) cell storing one bit of data consumes $150F^2$ area compared to an NVM that can take $4F^2$ space storing multiple bits], they lend themselves easily for distributed spatial architectures. This implies that an NVM based neuromorphic chip can have a crossbar array that stores a subset of the network weights, and such multiple crossbars can be arranged in a tiled manner, wherein weights are almost readily available within each tile for processing.

Keeping in view the aforementioned discussion, a generic NVM based distributed spatial architecture is shown in Fig. 19, enable mapping of neural network applications entirely using on-chip NVM. The

Published under license by AIP Publishing

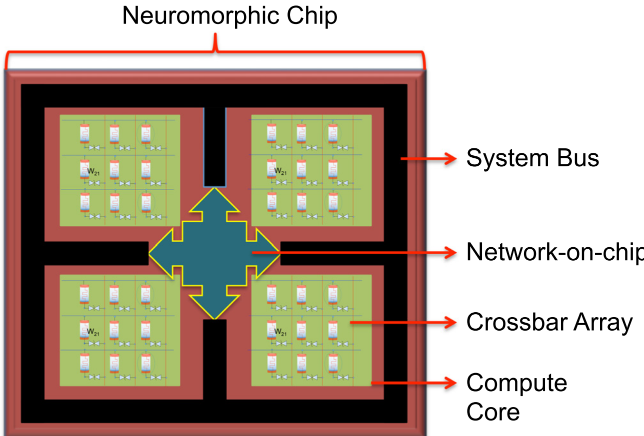


FIG. 19. A representative neuromorphic architecture based on NVM crossbars as basic compute engines.

various computing cores with their crossbar arrays are interconnected through network-on-chip (NOC). A distinct characteristic of SNN architecture is event-drivenness. SNNs communicate through spikes, i.e., binary information transfer between neurons. As such, for on-chip NOCs, spike-addresses are communicated between various compute cores rather than energy expensive transfer of actual data.[144] Furthermore, only those units are active, which have received a spike, and others remain idle, resulting in added energy-efficiency. Note that both spike-based on-chip communication and event-drivenness are direct consequences of SNN based data processing. Distributed architectures based on NVM technologies have been explored heavily to build special-purpose accelerators for both machine learning workloads such as convolutional neural networks (CNNs), multi-layer perceptrons (MLPs), and long short term memories (LSTMs),[145–148] as well as SNNs.[144,149] These works have demonstrated significant improvements over CMOS-based general purpose systems such as central processing units (CPU), graphics processing units (GPU), or application specific integrated circuits (ASICs),[150] which highlight the potential of neuromorphic computing based on NVM devices.

Until now, we have talked about inference-only accelerators that require fixed-point arithmetic, which NVM crossbars are well suited for. In addition, on-chip training based on unsupervised learning has been explored at a primitive level using low-precision devices;[151,152] however, training accelerators for large-scale tasks, which use such primitives, have not been demonstrated yet. Moreover, supervised learning, on the other hand, requires floating-point arithmetic due to small magnitude of weight updates, which is difficult to be captured by fixed-point representation. Architectures, which support training, thus face a significant challenge of incorporating such small updates to NVM crossbars. This problem is accentuated especially with limited endurance and high write latency of some NVM technologies, such as PCMs and RRAMs. Writing into crossbars in parallel using pulse width encoding schemes has been proposed although the scalability of such a technique still needs to be investigated.[153] Based on the discussion in this section, two important developments that are yet to be seen from the neuromorphic community with respect to architectures based on NVMs are (1) experimental demonstration of large-scale inference-only NVM crossbar systems that can rival their CMOS counterparts, for example, the CMOS based large-scale neuromorphic chip presented in Refs. 154 and 155, and (2) investigation and establishment of the limits of crossbar based neuromorphic systems for on-chip training keeping in mind the constrained writability of NVM technologies.

IV. PROSPECTS

A. Stochasticity—Opportunities and challenges

We have discussed about the promises of NVM technology for emulating neuro-synaptic behavior using single devices. These devices can have inherent variability embedded into their intrinsic physics, which can lead to stochastic characteristics. This is a major advantage from CMOS-based implementations where extra circuitry is required to generate stochastic behavior. Stochastic devices derive inspiration from the inherent stochasticity in biological synapses. Such synaptic uncertainty can be used in both learning and inferencing[157] in spiking neural networks. This is especially crucial for binary or ternary synapses where arbitrary weight update may result in overwriting previously learned features. Using stochasticity in binary synapses can vastly improve its feature recognition capabilities. This can be done in both a spatial manner[158] where a number of synapses are randomly chosen for weight update or a temporal manner[116] where learning in a probabilistic manner can follow the footsteps of the STDP based synaptic weight update algorithm. Stochastic STDP thus enables feature recognition with extremely low-precision synaptic efficacy, resulting in compressed networks,[152] which has the potential to achieve significant energy efficiency when implemented on hardware.[151] Stochastic learning is particularly helpful for low-precision synapses because its adds an analog probabilistic dimension, thus ensuring less degradation in accuracy in low-precision networks. For higher-precision networks where the classification accuracy does not degrade, stochasticity does not make a significant difference.

In addition to stochastic learning, we have also discussed how stochastic devices can be used to mimic the functionality of cortical neurons. In PCM devices, stochasticity has been explored in integrate-and-fire neurons[29] where multiple reset operations lead to different initial glass states. Although such stochastic IF characteristics can be exploited for robust computing, the overhead for achieving control over such stochasticity remains to be seen. On the other hand, in spin devices, stochastic neurons with sigmoidal characteristics are heavily tunable. These kinds of neurons have been explored both using high energy-barrier (10–60 kT) magnets[102] and low barrier magnets (1 kT).[107] While the resultant sigmoidal behavior looks similar, a 1 kT magnet loses its non-volatility and is more susceptible to variations, leading to more complex peripheral circuit design.[156] This results in the peripheral energy dominating the total energy consumption of such devices, which, interestingly, often makes them less energy-efficient than high barrier counterparts (Fig. 20).

B. Challenges of NVM crossbars

We have also discussed about the promises of NVM technology for emulating neuro-synaptic behavior using single devices. We have shown how these devices can be connected in an integrated crossbar network to perform large-scale neural computing. Although the promise of enabling parallel in-memory computations using crossbar arrays is attractive from the energy- and area-efficiency perspective, many non-ideal devices and circuit behaviors limit their wide scale

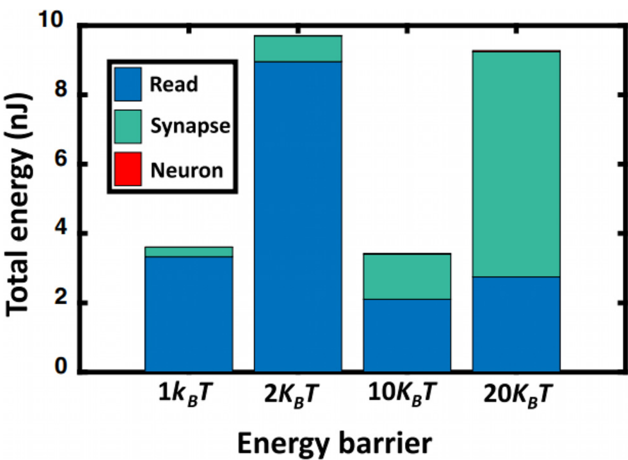


FIG. 20. A comparison in energy consumption for stochastic spin neurons for various energy-barrier heights.[156] Reproduced with permission from Liyanagedera *et al.*, Phys. Rev. Appl. **8**(6), 064017 (2017). Copyright 2017 American Physical Society.

applicability. These include the variability in RRAM states, which can detrimentally affect the verity of analog computations in synaptic elements. This is primarily due to the uncontrolled nature of the variability in filamentary RRAM or CBRAM devices.[159] PCM devices on the other hand, in spite of being less prone to variability, suffer from resistance drifting due to structural relaxations after the melt-quench amorphization of the material.[160] Resistance drifting primarily affects high resistance states in PCMs and hence adversely impacts the performance of neural networks especially for *ex situ* trained networks.[52] Carefully manipulating the highest resistance state of operation using partial resetting pulses can potentially reduce the impact of resistance drift.[52] Spintronic devices are more robust with respect to variability and endurance challenges as compared to RRAM and PCM technologies owing to their stable and controlled switching. However, practical devices suffer from low contrast in conductivity between the stable extremities. The low ON–OFF ratio severely affects the mapping of synaptic weights when implemented in neural networks and is the major technical roadblock for synaptic implementations using spin devices. Additionally, all non-volatile devices have energy and latency expensive write operations in comparison to conventional CMOS memories. This in turn limits the energy-efficiency of performing on-chip synaptic plasticity that requires frequent write operations.

Apart from device variations and limitations, building large-scale crossbars using non-volatile synaptic devices is a major hurdle toward realizing the goal of neuromorphic accelerators. Crossbar sizes are severely limited by various factors such as peripheral resistances, parasitic drops, and sneak paths. Figure 21 shows a schematic of a realistic crossbar with source, sink, and line resistances and peripherals. When training is performed on-chip taking into account the non-ideal crossbar behavior, such inaccuracies in crossbar computations can be mitigated to a large extent. However, for neuromorphic systems designed as inference-only engines, it is necessary to perform effective modeling of the crossbar array, which can potentially account for the non-idealities during off-line training and take corrective measures for accurate crossbar computations. Such modeling can either involve rigorous graph-based techniques for linear circuits,[161] simple equations involving Kirchoff's laws under certain assumptions,[162] or even data-dependent fitting.[163] Considering the minimal effect of IR-drops along the metal lines, equations of a crossbar under the effect of peripheral resistances can be simplified as

$$I_j = \frac{\sum V_{i,ni} G_{ij}}{1 + R_{sink} \sum G_{ij}}, \tag{10}$$

$$V_{i,ni} = V_i \frac{1/R_s}{1/R_s + \sum \dfrac{1}{R_{ji} + R_{sink}}}. \tag{11}$$

Here, I_j is the current of the j-th column, V_i is the input voltage to the i-th row of the crossbar, $(R_{ij} = 1/G_{ij})$ is the resistance/conductance of the synaptic element connecting the i-th row with the j-th column, $V_{i,ni}$ is the degraded input voltage due to the effect of peripheral resistances, R_s is the effective source resistance, and R_{sink} is the effective sink resistance. These resistances in relation to a crossbar are shown in Fig. 21. This modeling gives us an intuition about the behavior of crossbars, which can help preserve the computation accuracy. For example, lower synaptic resistances result in higher currents, which results in larger parasitic drops across the metal line. On the other hand, higher operating resistances might lead to low sensing margins, necessitating the need for expensive peripheral circuitry. The presence of sneak paths in synaptic crossbars can also adversely affect the programming process, thus harming the performance of on-chip learning systems.

In addition to non-ideal elements in NVM crossbars, the design of peripheral components such as Digital-to-Analog Converters

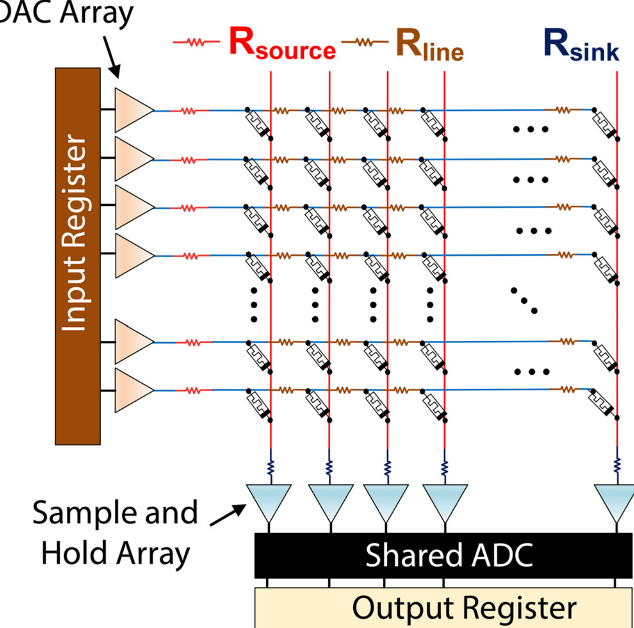


FIG. 21. A realistic crossbar system showing the peripheral circuits including digital-to-analog converters (DACs) at the input to the crossbar and analog-to-digital converters (ADCs) at the output. Crossbars can possess non-ideal resistance elements such as the source resistance (R_{source}), line resistance (R_{line}), and sink resistance (R_{sink}).

(DACs) and Analog-to-Digital Converters (ADCs) is essential toward building large-scale neuromorphic systems. As shown in Fig. 21, DACs are used to convert bit-streamed data to voltages, whereas the ADCs convert back the analog voltage outputs from a sample-and-hold array into digital bits. These converters are especially necessary as the sizes of neural network models are much higher than the size of a single crossbar. As a result, multiple crossbars are required to represent the entire neural network, which necessitates digital communication between the outputs of individual crossbars. As the crossbar size increases, the precision requirements for ADCs become higher, leading to enormous power consumption, which can potentially reduce the benefits in terms of energy consumption that NVM crossbars inherently offer. However, the inherent robustness of neural networks toward computation errors may allow us to design approximate peripheral circuitry based on ADCs with lower precision requirements. Moreover, efficient mapping of crossbars and introducing programmability in peripheral precision requirements can potentially preserve the benefits offered by NVM technology. In light of these challenges such as device variations, non-ideal resistances, sneak paths, and peripheral design, careful design space exploration is required to identify optimum resistances for operation and crossbar sizes of synaptic elements along with efficient device-circuit-algorithm co-design for exploring effective mitigation techniques.

C. Mitigating crossbar non-idealities

NVM provides a massively parallel mode of computations using crossbars. However, as we have discussed previously, analog computing is error-prone due to the presence of circuit-level non-idealities and device variations. Various mitigation techniques have been explored to address these computing inaccuracies. Although some of these techniques have been demonstrated for artificial neural networks, the methodologies still hold true for spike-based neuromorphic computing. The most commonly used methodology to recover the performance of neural networks due to crossbar-level computing errors is to re-train the network using software models of resistive crossbars. The re-training approach involves updating the weights of the network based on information of non-idealities in crossbars. This has been explored for both stuck-at-faults[164] and device variations[165] where it has been observed that re-training the network with awareness about the defect or variation distribution can minimize the effects of these non-idealities on classification performance. Re-training, however, does not recover the performance of an ideal neural network without any non-idealities. The presence of non-idealities in the forward path of a neural network may require a modified backpropagation algorithm to closely resemble the ideal neural network.[162] For unsupervised learning algorithms such as STDP, the impact of non-idealities may be significantly lower due to the ease of enabling on-line learning, which can automatically account for the errors. In addition to static non-idealities in the crossbars, the effect of non-linearity and asymmetry of programming characteristics of NVM devices can also be detrimental to the performance of the network. Reliable mitigation due to such programming errors can be performed by novel pulsing schemes.[166,167] These pulsing schemes involve modulation of pulse-widths based on the current conductance state, which help restore linearity.

Beyond re-training, other static compensation techniques can also be used to recover some system level inaccuracies. For example, the limited ON/OFF ratio and precision of NVM synaptic devices can result in computational errors, which can be taken care of by effective mapping of weight matrices to synaptic conductance.[168] Static transformations of weight matrices have been explored to alleviate circuit-level non-idealities.[169] This methodology performs gradient search to identify weight matrices with non-idealities that resemble ideal weight matrices. Most of the compensation techniques adopted to account for computation inaccuracies in NVM crossbars address very specific problems. A more complete and holistic analysis, modeling, and mitigation of crossbar non-idealities are necessary to completely understand the impact and explore appropriate solutions.

D. Multi-memristive synapses

Multi-memristive synapses are examples, wherein device limitations have been countered by the use of circuit techniques, albeit at additional area overhead. Figure 22 depicts two illustrations, which use multiple NVM devices to represent one synaptic weight. In Fig. 22(a), two separate PCM devices were used to implement LTD and LTP separately. Incrementing the PCM device corresponding to LTP increased the neuronal input, whereas incrementing the device corresponding to LTP decreased the neuronal input. By this scheme, the authors in Ref. 35 were able to simply the peripheral write circuits since only increments in device resistances were required for representing both LTP and LTD plasticity. Note that conventionally using one single device would have required write circuits for both incrementing and

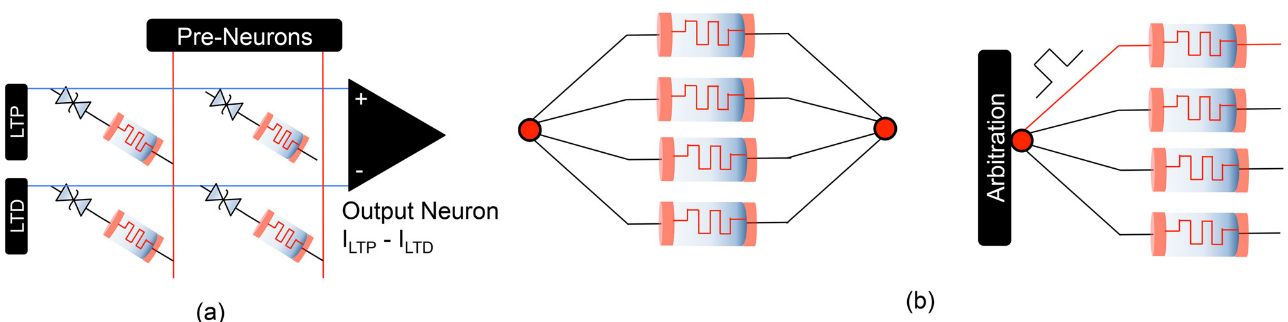


FIG. 22. (a) Two separate NVM devices used for LTP and LTD, and the resulting output of the synapse is fed to the neuron. (b) Multiple NVM devices connected in parallel to increase the current range of the synapse. (c) Through the use of an arbitrator, any one of the devices is selected for learning.

Published under license by AIP Publishing

decrementing the PCM device resistance, and given the complex nature of waveforms required to write into PCM devices, this would have led to additional area overhead. In yet another work, more than one memristors were connected in parallel [Fig. 22(b)][170] to allow the increased current range of the overall synaptic cell. For learning, an arbitration scheme was used to select one memristor and program in accordance with the learning scheme as shown in Fig. 22(c). With reference to these examples, we believe that such schemes, wherein device level constraints can be mitigated through the use of clever circuit techniques, can be a key enabler for NVMs in neuromorphic computing without solely relying on better material stack and manufacturing processes for improved device characteristics.

E. Beyond neuro-synaptic devices and STDP

As would be apparent by now, the state-of-the-art in neuromorphic hardware using non-volatile devices can be characterized in two broad categories of works—(1) those that tend to mimic the LIF dynamics of a neuron using device characteristics and (2) others that are geared toward synaptic functionalities and associated learning through STDP in shallow SNNs. On the other hand, the state-of-the-art on the algorithmic side of neuromorphic computing has taken a step forward beyond LIF dynamics and STDP learning. We have discussed briefly about how supervised learning such as gradient descent can also be used for spike-based systems. Previously, supervised learning has been performed in the artificial neural networks (ANN) domain, and trained networks have been converted to SNNs.[27] Although this method has been scaled to complex image recognition datasets such as ImageNet, one particular drawback of this scheme is high inference latency. To circumvent that, researchers have explored learning schemes, which incorporate such gradient descent algorithms in the spiking domain itself.[28,171,172] Moreover, combining unsupervised and supervised learning techniques have also been widely explored.[173] This kind of hybrid learning technique has shown better scalability (to deeper networks) and improved accuracy.

We believe that it is important for the hardware community to move beyond mimicking neurons and synapses on shallow SNNs and find ways and means of executing more dynamic learning schemes on hardware for deeper spiking networks. Such improved learning schemes would inevitably require complex compute operations, which could be beyond the intrinsic device characteristics of non-volatile devices. As such, there is a need to explore systems, wherein computations can be segregated between non-volatile sub-arrays and CMOS based compute engines, allowing the overall system to benefit both from parallelism offered by NVMs and the compute complexity offered by CMOS engines. This would also be a key enabler in building end-to-end deployable neuromorphic systems (wherein a spike-based sensor is directly interfaced to a neuromorphic processor) that can cater to real life task as in ultra-low energy IoT systems. Such IoT systems not only are important from a research perspective but can also provide a possible commercial niche-application for neuromorphic processors based on non-volatile technologies.

F. NVM for digital in-memory computing

Most of the current works involving neuromorphic computing and emerging devices have concentrated on analog-mixed-signal computing. However, the inherent approximations associated with analog computing still remain a major technical roadblock. In contrast, one could use digital in-memory computing for implementing on-chip robust SNN networks. These implementations can use various digital techniques, as in use of read only memory (ROM) embedded RAM in NVM arrays[174] or peripheral circuits based on in-memory digital computations.[175] Interestingly, these works do not require heavy re-engineering of the devices themselves. As such, they can easily benefit from the recent technological and manufacturing advancements driven by industry for commercialization of various non-volatile technologies as memory solutions.

Furthermore, in a large neural network, NVM can be used as *significance driven* in-memory compute accelerators. For example, layers of the neural network, which are less susceptible to noise, can be accelerated using analog in-memory computing, while those layers that need more accurate computations can be mapped on NVM arrays rendering digital in-memory computing. Thus, fine-grained heterogeneous in-memory computing (both digital and analog) can be used in unison to achieve both lower energy consumption and higher application accuracy. It is also well known that NVMs that store data digitally are easier to program as opposed to analog storage, which requires multiple "read-verify" cycles. Thus, on-chip learning, which requires frequent weight updates, is more amenable to digital or heterogeneous (digital + analog) computing arrays as opposed to analog storage of data. Additionally, bit errors induced due to digital computing can be easily rectified using error correction codes. Thereby, resorting to digital processing for critical or error susceptible computation could help widen the design space for use of NVMs as SNN accelerators.

G. Physical integrability of NVM technology with CMOS

There are several works on experimental demonstration of in-memory computing primitives based on non-volatile memories, especially RRAM and PCM technologies.[45,84,95] NVM devices in most state-of-art RRAM and PCM crossbars are accompanied by a CMOS selector device (like a transistor). Such a 1T-1R crossbar configuration resolves sneak paths during read and write operations.[176] Crossbars based on NVM technologies such as RRAM,[177] PCM,[178] and Spintronics[179] are fully compatible with the CMOS back end of the line (BEOL) integration process. There are some issues that need to be considered. For example, PCM is fabricated in crystalline form, as BEOL integration involves high temperature processes. Although there have been large-scale demonstrations on RRAM and PCM crossbars with CMOS peripherals, work on CMOS integration of spintronic devices has been limited to small scale Boolean logic circuits.[179] It is to be noted that the limited use of spin devices for the crossbar structure is a result of the low ON–OFF ratio for spintronic devices and not because of compatibility issues pertaining to integration of spin devices with CMOS technology. In fact, the current advancement in process integration for spin based devices with CMOS technology has led to recent widespread interest for commercial use of spin based read-write memories.[180] FEFETs, on the other hand, follow the standard Front End of Line (FEOL) CMOS process. Thus, all the NVM technologies being explored can be physically integrated with CMOS.

V. CONCLUSION

The growing complexity of deep learning models and the humongous power consumption of standard von-Neumann

computers while implementing such models have led to a three decade long search for bio-plausible computing paradigms. They draw inspiration from the elusive energy-efficiency of the brain. To that effect, non-volatile technologies offer a promising solution toward realizing such computing systems. In this review article, we discuss how the rich intrinsic physics of non-volatile devices, based on various technologies, can be exploited to emulate bio-plausible neuro-synaptic functionalities in spiking neural networks. We delve into the generic requirements of the basic functional units of SNNs and how they can be realized using various non-volatile devices. These devices can be connected in an intricate arrangement to realize a massively parallel in-memory computing crossbar structure representing a radical departure from the existing von-Neumann computing model. A huge number of such computing units can be arranged in a tiled architecture to realize extremely area and energy-efficient large-scale neuromorphic systems. Finally, we discuss the challenges and possible solution of realizing neuromorphic systems using non-volatile devices. We believe that non-volatile technologies show significant promise and immense potential as the building blocks in neuromorphic systems of the future. In order to truly realize that potential, a joint research effort is necessary, right from the materials that would achieve better trade-offs between higher stability and programming speeds and exhibit more linear and symmetric characteristics. This material investigation should be complemented with effective device-circuit co-design to alleviate problems of variations and other non-idealities that introduce errors into neuromorphic computations. Finally, there must be efficient hardware-algorithm amalgamation to design more hardware-friendly algorithms and vice versa. With these challenges in mind and possible avenues of research, the dream of achieving truly integrated non-volatile technology based neuromorphic systems should not be far into the future.

AUTHORS' CONTRIBUTION

I.C. and A.J. contributed equally to this work.

ACKNOWLEDGMENTS

This work was supported in part by the Center for Brain-inspired Computing Enabling Autonomous Intelligence (C-BRIC), a DARPA sponsored JUMP center, the Semiconductor Research Corporation, the National Science Foundation, Intel Corporation, the DoD Vannevar Bush Fellowship, the Office of Naval Research Multidisciplinary University Research Initiative, the U.S. Army Research Laboratory, and the U.K. Ministry of Defense under Agreement No. W911NF-16-3-0001.

REFERENCES

[1]F. Allen, G. Almasi, W. Andreoni, D. Beece, B. J. Berne, A. Bright, J. Brunheroto, C. Cascaval, J. Castanos, P. Coteus *et al.*, "Blue gene: A vision for protein science using a petaflop supercomputer," IBM Syst. J. **40**, 310–327 (2001).
[2]N. Howard, "Energy paradox of the brain," Brain Sci. **1**, 35 (2012).
[3]R. Ananthanarayanan, S. K. Esser, H. D. Simon, and D. S. Modha, "The cat is out of the bag: Cortical simulations with 10^9 neurons, 10^{13} synapses," in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis* (ACM, 2009), p. 63.
[4]W. Maass, "Networks of spiking neurons: The third generation of neural network models," Neural Networks **10**, 1659–1671 (1997).
[5]S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, and classification," IEEE Trans. Neural Networks **3**, 683–697 (1992).
[6]V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in Proceedings of the 27th International Conference on Machine Learning (ICML-10) (2010), pp. 807–814.
[7]D. E. Rumelhart, G. E. Hinton, R. J. Williams *et al.*, "Learning representations by back-propagating errors," Nature **323**(6088), 533–536 (1986).
[8]Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436 (2015).
[9]K. A. Boahen, "Point-to-point connectivity between neuromorphic chips using address events," IEEE Trans. Circuits Syst. II **47**, 416–434 (2000).
[10]G. M. Shepherd, *The Synaptic Organization of the Brain* (Oxford University Press, 2003).
[11]M. Mahowald and R. Douglas, "A silicon neuron," Nature **354**, 515 (1991).
[12]E. T. Rolls and G. Deco, *The Noisy Brain: Stochastic Dynamics as a Principle of Brain Function* (Oxford University Press, Oxford, 2010), Vol. 34.
[13]B. Nessler, M. Pfeiffer, L. Buesing, and W. Maass, "Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity," PLoS Comput. Biol. **9**, e1003037 (2013).
[14]A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," J. Physiol. **117**, 500–544 (1952).
[15]L. F. Abbott, "Lapicque's introduction of the integrate-and-fire model neuron (1907)," Brain Res. Bull. **50**, 303–304 (1999).
[16]A. V. Hill, "Excitation and accommodation in nerve," Proc. R. Soc. London, Ser. B **119**, 305–355 (1936).
[17]C. D. Geisler and J. M. Goldberg, "A stochastic model of the repetitive activity of neurons," Biophys. J. **6**, 53–69 (1966).
[18]Y.-H. Liu and X.-J. Wang, "Spike-frequency adaptation of a generalized leaky integrate-and-fire model neuron," J. Comput. Neurosci. **10**, 25–45 (2001).
[19]G. Qiang Bi and M. Ming Poo, "Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type," J. Neurosci. **18**, 10464–10472 (1998).
[20]T. V. Bliss and G. L. Collingridge, "A synaptic model of memory: Long-term potentiation in the hippocampus," Nature **361**, 31 (1993).
[21]R. S. Zucker and W. G. Regehr, "Short-term synaptic plasticity," Annu. Rev. Physiol. **64**, 355–405 (2002).
[22]C. F. Stevens and Y. Wang, "Facilitation and depression at single central synapses," Neuron **14**, 795–802 (1995).
[23]R. Atkinson and R. Shiffrin, "Human memory: A proposed system and its control processes," in *Psychology of Learning and Motivation* (Elsevier, 1968), pp. 89–195.
[24]B. Rueckauer, I.-A. Lungu, Y. Hu, M. Pfeiffer, and S.-C. Liu, "Conversion of continuous-valued deep networks to efficient event-driven networks for image classification," Front. Neurosci. **11**, 682 (2017).
[25]P. U. Diehl, D. Neil, J. Binas, M. Cook, S.-C. Liu, and M. Pfeiffer, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *2015 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2015).
[26]S. K. Esser, P. A. Merolla, J. V. Arthur, A. S. Cassidy, R. Appuswamy, A. Andreopoulos, D. J. Berg, J. L. McKinstry, T. Melano, D. R. Barch, C. di Nolfo, P. Datta, A. Amir, B. Taba, M. D. Flickner, and D. S. Modha, "Convolutional networks for fast, energy-efficient neuromorphic computing," Proc. Natl. Acad. Sci. **113**, 11441–11446 (2016).
[27]A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going deeper in spiking neural networks: VGG and residual architectures," Front. Neurosci. **13**, 95 (2019).
[28]Y. Jin, W. Zhang, and P. Li, "Hybrid macro/micro level backpropagation for training deep spiking neural networks," in *Advances in Neural Information Processing Systems* (Curran Associates, 2018), pp. 7005–7015.
[29]T. Tuma, A. Pantazi, M. Le Gallo, A. Sebastian, and E. Eleftheriou, "Stochastic phase-change neurons," Nat. Nanotechnol. **11**, 693 (2016).
[30]I. Chakraborty, G. Saha, A. Sengupta, and K. Roy, "Toward fast neural computing using all-photonic phase change spiking neurons," Sci. Rep. **8**, 12980 (2018).
[31]M. Stanisavljevic, H. Pozidis, A. Athmanathan, N. Papandreou, T. Mittelholzer, and E. Eleftheriou, "Demonstration of reliable triple-level-cell (TLC) phase-change memory," in *2016 IEEE 8th International Memory Workshop (IMW)* (IEEE, 2016), pp. 1–4.

[32]T. Tuma, M. Le Gallo, A. Sebastian, and E. Eleftheriou, "Detecting correlations using phase-change neurons and synapses," IEEE Electron Device Lett. **37**, 1238–1241 (2016).
[33]S. Ambrogio, N. Ciocchini, M. Laudato, V. Milo, A. Pirovano, P. Fantini, and D. Ielmini, "Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses," Front. Neurosci. **10**, 56 (2016).
[34]D. Kuzum, R. G. Jeyasingh, B. Lee, and H.-S. P. Wong, "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing," Nano Lett. **12**, 2179–2186 (2012).
[35]M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction," in *2011 International Electron Devices Meeting* (IEEE, 2011), p. 4.
[36]Y. Li, Y. Zhong, L. Xu, J. Zhang, X. Xu, H. Sun, and X. Miao, "Ultrafast synaptic events in a chalcogenide memristor," Sci. Rep. **3**, 1619 (2013).
[37]O. Bichler, M. Suri, D. Querlioz, D. Vuillaume, B. DeSalvo, and C. Gamrat, "Visual pattern extraction using energy-efficient "2-PCM synapse" neuromorphic architecture," IEEE Trans. Electron Devices **59**, 2206–2214 (2012).
[38]D. Kuzum, R. G. Jeyasingh, and H.-S. P. Wong, "Energy efficient programming of nanoelectronic synaptic devices for large-scale implementation of associative and temporal sequence learning," in *2011 International Electron Devices Meeting* (IEEE, 2011), pp. 30–33.
[39]Z. Cheng, C. Ríos, W. H. Pernice, C. D. Wright, and H. Bhaskaran, "On-chip photonic synapse," Sci. Adv. **3**, e1700160 (2017).
[40]J. Feldmann, N. Youngblood, C. Wright, H. Bhaskaran, and W. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," Nature **569**, 208 (2019).
[41]S. B. Eryilmaz, D. Kuzum, R. G. Jeyasingh, S. Kim, M. BrightSky, C. Lam, and H.-S. P. Wong, "Experimental demonstration of array-level learning with phase change synaptic devices," in *2013 IEEE International Electron Devices Meeting* (IEEE, 2013), p. 25.
[42]S. B. Eryilmaz, D. Kuzum, R. Jeyasingh, S. Kim, M. BrightSky, C. Lam, and H.-S. P. Wong, "Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array," Front. Neurosci. **8**, 205 (2014).
[43]S. Kim, M. Ishii, S. Lewis, T. Perri, M. BrightSky, W. Kim, R. Jordan, G. Burr, N. Sosa, A. Ray *et al.*, "NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning," in *2015 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2015), p. 17.
[44]G. W. Burr, R. M. Shelby, S. Sidler, C. Di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element," IEEE Trans. Electron Devices **62**, 3498–3507 (2015).
[45]S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. Farinha *et al.*, "Equivalent-accuracy accelerated neural-network training using analogue memory," Nature **558**, 60 (2018).
[46]I. Chakraborty, G. Saha, and K. Roy, "Photonic in-memory computing primitive for spiking neural networks using phase-change materials," Phys. Rev. Appl. **11**, 014063 (2019).
[47]T. Matsunaga, N. Yamada, and Y. Kubota, "Structures of stable and metastable $Ge_2Sb_2Te_5$, an intermetallic compound in $GeTe–Sb_2Te_3$ pseudobinary systems," Acta Crystallogr., Sect. B **60**, 685–691 (2004).
[48]M. Wuttig and N. Yamada, "Phase-change materials for rewriteable data storage," Nat. Mater. **6**, 824 (2007).
[49]A. Pirovano, A. L. Lacaita, F. Pellizzer, S. A. Kostylev, A. Benvenuti, and R. Bez, "Low-field amorphous state resistance and threshold voltage drift in chalcogenide materials," IEEE Trans. Electron Devices **51**, 714–719 (2004).
[50]S. Kim, B. Lee, M. Asheghi, F. Hurkx, J. P. Reifenberg, K. E. Goodson, and H.-S. P. Wong, "Resistance and threshold switching voltage drift behavior in phase-change memory and their temperature dependence at microsecond time scales studied using a micro-thermal stage," IEEE Trans. Electron Devices **58**, 584–592 (2011).
[51]D. Ielmini, S. Lavizzari, D. Sharma, and A. L. Lacaita, "Physical interpretation, modeling and impact on phase change memory (PCM) reliability of resistance drift due to chalcogenide structural relaxation," in *2007 IEEE International Electron Devices Meeting* (IEEE, 2007), pp. 939–942.
[52]M. Suri, D. Garbin, O. Bichler, D. Querlioz, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Impact of PCM resistance-drift in neuromorphic systems and drift-mitigation strategy," in *Proceedings of the 2013 IEEE/ACM International Symposium on Nanoscale Architectures* (IEEE Press, 2013), pp. 140–145.
[53]Y. Watanabe, J. Bednorz, A. Bietsch, C. Gerber, D. Widmer, A. Beck, and S. Wind, "Current-driven insulator–conductor transition and nonvolatile memory in chromium-doped $SrTiO_3$ single crystals," Appl. Phys. Lett. **78**, 3738–3740 (2001).
[54]A. Beck, J. Bednorz, C. Gerber, C. Rossel, and D. Widmer, "Reproducible switching effect in thin oxide films for memory applications," Appl. Phys. Lett. **77**, 139–141 (2000).
[55]W. Zhuang, W. Pan, B. Ulrich, J. Lee, L. Stecker, A. Burmaster, D. Evans, S. Hsu, M. Tajiri, A. Shimaoka *et al.*, "Novel colossal magnetoresistive thin film nonvolatile resistance random access memory (RRAM)," in *International Electron Devices Meeting, Technical Digest* (IEEE, 2002), pp. 193–196.
[56]L. Goux, P. Czarnecki, Y. Y. Chen, L. Pantisano, X. Wang, R. Degraeve, B. Govoreanu, M. Jurczak, D. Wouters, and L. Altimime, "Evidences of oxygen-mediated resistive-switching mechanism in $TiN\backslash HfO_2\backslash Pt$ cells," Appl. Phys. Lett. **97**, 243509 (2010).
[57]C. Rohde, B. J. Choi, D. S. Jeong, S. Choi, J.-S. Zhao, and C. S. Hwang, "Identification of a determining parameter for resistive switching of TiO_2 thin films," Appl. Phys. Lett. **86**, 262907 (2005).
[58]Z. Wei, Y. Kanzawa, K. Arita, Y. Katoh, K. Kawai, S. Muraoka, S. Mitani, S. Fujii, K. Katayama, M. Iijima *et al.*, "Highly reliable TaO_x ReRAM and direct evidence of redox reaction mechanism," in *2008 IEEE International Electron Devices Meeting* (IEEE, 2008), pp. 1–4.
[59]M. Al-Shedivat, R. Naous, G. Cauwenberghs, and K. N. Salama, "Memristors empower spiking neurons with stochasticity," IEEE J. Emerging Sel. Top. Circuits Syst. **5**, 242–253 (2015).
[60]J.-W. Jang, B. Attarimashalkoubeh, A. Prakash, H. Hwang, and Y.-H. Jeong, "Scalable neuron circuit using conductive-bridge ram for pattern reconstructions," IEEE Trans. Electron Devices **63**, 2610–2613 (2016).
[61]D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," Nature **453**, 80 (2008).
[62]L. Chua, "Memristor-The missing circuit element," IEEE Trans. Circuit Theory **18**, 507–519 (1971).
[63]S. Lashkare, S. Chouhan, T. Chavan, A. Bhat, P. Kumbhare, and U. Ganguly, "PCMO RRAM for integrate-and-fire neuron in spiking neural networks," IEEE Electron Device Lett. **39**, 484–487 (2018).
[64]A. Mehonic and A. J. Kenyon, "Emulating the electrical activity of the neuron using a silicon oxide RRAM cell," Front. Neurosci. **10**, 57 (2016).
[65]H. Lee, P. Chen, T. Wu, Y. Chen, C. Wang, P. Tzeng, C. Lin, F. Chen, C. Lien, and M.-J. Tsai, "Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO_2 based Rram," in *2008 IEEE International Electron Devices Meeting* (IEEE, 2008), pp. 1–4.
[66]X. Cao, X. Li, X. Gao, W. Yu, X. Liu, Y. Zhang, L. Chen, and X. Cheng, "Forming-free colossal resistive switching effect in rare-earth-oxide Gd_2O_3 films for memristor applications," J. Appl. Phys. **106**, 073723 (2009).
[67]M.-J. Lee, S. Han, S. H. Jeon, B. H. Park, B. S. Kang, S.-E. Ahn, K. H. Kim, C. B. Lee, C. J. Kim, I.-K. Yoo *et al.*, "Electrical manipulation of nanofilaments in transition-metal oxides for resistance-based memory," Nano Lett. **9**, 1476–1481 (2009).
[68]C. Yoshida, K. Kinoshita, T. Yamasaki, and Y. Sugiyama, "Direct observation of oxygen movement during resistance switching in NiO/Pt film," Appl. Phys. Lett. **93**, 042106 (2008).
[69]S. Yu and H.-S. P. Wong, "A phenomenological model for the reset mechanism of metal oxide RRAM," IEEE Electron Device Lett. **31**, 1455–1457 (2010).
[70]S. R. Lee, Y.-B. Kim, M. Chang, K. M. Kim, C. B. Lee, J. H. Hur, G.-S. Park, D. Lee, M.-J. Lee, C. J. Kim *et al.*, "Multi-level switching of triple-layered TaO_x RRAM with excellent reliability for storage class memory," in *2012 Symposium on VLSI Technology (VLSIT)* (IEEE, 2012), pp. 71–72.
[71]M. Prezioso, F. M. Bayat, B. Hoskins, K. Likharev, and D. Strukov, "Self-adaptive spike-time-dependent plasticity of metal-oxide memristors," Sci. Rep. **6**, 21331 (2016).

Published under license by AIP Publishing

[72]B. Rajendran, Y. Liu, J.-S. Seo, K. Gopalakrishnan, L. Chang, D. J. Friedman, and M. B. Ritter, "Specifications of nanoscale devices and circuits for neuromorphic computational systems," IEEE Trans. Electron Devices **60**, 246–253 (2013).
[73]K. Seo, I. Kim, S. Jung, M. Jo, S. Park, J. Park, J. Shin, K. P. Biju, J. Kong, K. Lee *et al.*, "Analog memory and spike-timing-dependent plasticity characteristics of a nanoscale titanium oxide bilayer resistive switching device," Nanotechnology **22**, 254023 (2011).
[74]I.-T. Wang, Y.-C. Lin, Y.-F. Wang, C.-W. Hsu, and T.-H. Hou, "3D synaptic architecture with ultralow sub-10 fJ energy per spike for neuromorphic computation," in *2014 IEEE International Electron Devices Meeting* (IEEE, 2014), p. 28.
[75]Z. Wang, S. Ambrogio, S. Balatti, and D. Ielmini, "A 2-transistor/1-resistor artificial synapse capable of communication and stochastic learning in neuromorphic systems," Front. Neurosci. **8**, 438 (2015).
[76]S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H.-S. P. Wong, "Stochastic learning in oxide binary synaptic device for neuromorphic computing," Front. Neurosci. **7**, 186 (2013).
[77]A. Serb, J. Bill, A. Khiat, R. Berdan, R. Legenstein, and T. Prodromakis, "Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses," Nat. Commun. **7**, 12611 (2016).
[78]S. Park, H. Kim, M. Choo, J. Noh, A. Sheri, S. Jung, K. Seo, J. Park, S. Kim, W. Lee *et al.*, "RRAM-based synapse for neuromorphic system with pattern recognition function," in *2012 International Electron Devices Meeting* (IEEE, 2012), pp. 10–12.
[79]S. Park, A. Sheri, J. Kim, J. Noh, J. Jang, M. Jeon, B. Lee, B. Lee, B. Lee, and H.-J. Hwang, "Neuromorphic speech systems using advanced ReRAM-based synapse," in *2013 IEEE International Electron Devices Meeting* (IEEE, 2013), pp. 25–26.
[80]N. Panwar, B. Rajendran, and U. Ganguly, "Arbitrary spike time dependent plasticity (STDP) in memristor by analog waveform engineering," IEEE Electron Device Lett. **38**, 740–743 (2017).
[81]H. Lim, I. Kim, J.-S. Kim, C. S. Hwang, and D. S. Jeong, "Short-term memory of TiO_2-based electrochemical capacitors: Empirical analysis with adoption of a sliding threshold," Nanotechnology **24**, 384005 (2013).
[82]R. Yang, K. Terabe, Y. Yao, T. Tsuruoka, T. Hasegawa, J. K. Gimzewski, and M. Aono, "Synaptic plasticity and memory functions achieved in a WO_{3-x}-based nanoionics device by using the principle of atomic switch operation," Nanotechnology **24**, 384003 (2013).
[83]R. Berdan, E. Vasilaki, A. Khiat, G. Indiveri, A. Serb, and T. Prodromakis, "Emulating short-term synaptic dynamics with memristive devices," Sci. Rep. **6**, 18639 (2016).
[84]F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, and W. D. Lu, "A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations," Nat. Electron. **2**, 290–299 (2019).
[85]X. Sun, S. Yin, X. Peng, R. Liu, J.-S. Seo, and S. Yu, "XNOR-RRAM: A scalable and parallel resistive synaptic architecture for binary neural networks," in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (IEEE, 2018), pp. 1423–1428.
[86]P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, "Fully hardware-implemented memristor convolutional neural network," Nature **577**, 641–646 (2020).
[87]C. Xue, T. Huang, J. Liu, T. Chang, H. Kao, J. Wang, T. Liu, S. Wei, S. Huang, W. Wei, Y. Chen, T. Hsu, Y. Chen, Y. Lo, T. Wen, C. Lo, R. Liu, C. Hsieh, K. Tang, and M. Chang, "15.4 A 22 nm 2Mb ReRAM compute-in-memory macro with 121-28TOPS/W for multibit mac computing for tiny AI edge devices," in 2020 IEEE International Solid-State Circuits Conference (ISSCC) (2020), pp. 244–246.
[88]S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," Nano Lett. **10**, 1297–1301 (2010).
[89]D. Mahalanabis, M. Sivaraj, W. Chen, S. Shah, H. J. Barnaby, M. N. Kozicki, J. B. Christen, and S. Vrudhula, "Demonstration of spike timing dependent plasticity in CBRAM devices with silicon neurons," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2016), pp. 2314–2317.
[90]M. Suri, D. Querlioz, O. Bichler, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Bio-inspired stochastic computing using binary CBRAM synapses," IEEE Trans. Electron Devices **60**, 2402–2409 (2013).
[91]T. Ohno, T. Hasegawa, A. Nayak, T. Tsuruoka, J. K. Gimzewski, and M. Aono, "Sensory and short-term memory formations observed in a Ag_2S gap-type atomic switch," Appl. Phys. Lett. **99**, 203108 (2011).
[92]K.-H. Kim, S. Gaba, D. Wheeler, J. M. Cruz-Albrecht, T. Hussain, N. Srinivasa, and W. Lu, "A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications," Nano Lett. **12**, 389–395 (2012).
[93]Y. Wang, T. Tang, L. Xia, B. Li, P. Gu, H. Yang, H. Li, and Y. Xie, "Energy efficient RRAM spiking neural network for real time classification," in *Proceedings of the 25th Edition on Great Lakes Symposium on VLSI* (ACM, 2015), pp. 189–194.
[94]G. Pedretti, V. Milo, S. Ambrogio, R. Carboni, S. Bianchi, A. Calderoni, N. Ramaswamy, A. Spinelli, and D. Ielmini, "Memristive neural network for on-line learning and tracking with brain-inspired spike timing dependent plasticity," Sci. Rep. **7**, 5288 (2017).
[95]C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves *et al.*, "Analogue signal and image processing with large memristor crossbars," Nat. Electron. **1**, 52 (2018).
[96]M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," in *European Conference on Computer Vision* (Springer, 2016), pp. 525–542.
[97]P. Wijesinghe, A. Ankit, A. Sengupta, and K. Roy, "An all-memristor deep spiking neural computing system: A step toward realizing the low-power stochastic brain," IEEE Trans. Emerging Top. Comput. Intell. **2**, 345–358 (2018).
[98]G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen, "A novel architecture of the 3D stacked MRAM l2 cache for CMPS," in *2009 IEEE 15th International Symposium on High Performance Computer Architecture* (IEEE, 2009), pp. 239–249.
[99]J. C. Slonczewski, "Current-driven excitation of magnetic multilayers," J. Magn. Magn. Mater. **159**, L1–L7 (1996).
[100]S. Emori, U. Bauer, S.-M. Ahn, E. Martinez, and G. S. Beach, "Current-driven dynamics of chiral ferromagnetic domain walls," Nat. Mater. **12**, 611 (2013).
[101]M. Bibes and A. Barthélémy, "Multiferroics: Towards a magnetoelectric memory," Nat. Mater. **7**, 425 (2008).
[102]A. Sengupta, P. Panda, P. Wijesinghe, Y. Kim, and K. Roy, "Magnetic tunnel junction mimics stochastic cortical spiking neurons," Sci. Rep. **6**, 30039 (2016).
[103]A. Jaiswal, S. Roy, G. Srinivasan, and K. Roy, "Proposal for a leaky-integrate-fire spiking neuron based on magnetoelectric switching of ferromagnets," IEEE Trans. Electron Devices **64**, 1818–1824 (2017).
[104]A. Sengupta and K. Roy, "Encoding neural and synaptic functionalities in electron spin: A pathway to efficient neuromorphic computing," Appl. Phys. Rev. **4**, 041105 (2017).
[105]X. Fong, Y. Kim, R. Venkatesan, S. H. Choday, A. Raghunathan, and K. Roy, "Spin-transfer torque memories: Devices, circuits, and systems," Proc. IEEE **104**, 1449–1488 (2016).
[106]W. F. Brown, Jr., "Thermal fluctuations of a single-domain particle," Phys. Rev. **130**, 1677 (1963).
[107]O. Hassan, R. Faria, K. Y. Camsari, J. Z. Sun, and S. Datta, "Low barrier magnet design for efficient hardware binary stochastic neurons," IEEE Magn. Lett. **10**, 1 (2019).
[108]A. Sengupta and K. Roy, "A vision for all-spin neural networks: A device to system perspective," IEEE Trans. Circuits Syst., I **63**, 2267–2277 (2016).
[109]A. Jaiswal, A. Agrawal, P. Panda, and K. Roy, "Voltage-driven domain-wall motion based neuro-synaptic devices for dynamic on-line learning," arXiv:1705.06942 (2017).
[110]A. Thiaville, Y. Nakatani, J. Miltat, and Y. Suzuki, "Micromagnetic understanding of current-driven domain wall motion in patterned nanowires," Europhys. Lett. **69**, 990 (2005).
[111]M.-C. Chen, A. Sengupta, and K. Roy, "Magnetic skyrmion as a spintronic deep learning spiking neuron processor," IEEE Trans. Magn. **54**, 1–7 (2018).
[112]A. Sengupta, A. Banerjee, and K. Roy, "Hybrid spintronic-CMOS spiking neural network with on-chip learning: Devices, circuits, and systems," Phys. Rev. Appl. **6**, 064003 (2016).

Published under license by AIP Publishing

[113]M. Sharad, C. Augustine, G. Panagopoulos, and K. Roy, "Spin-based neuron model with domain-wall magnets as synapse," IEEE Trans. Nanotechnol. **11**, 843–853 (2012).
[114]A. Sengupta, Y. Shim, and K. Roy, "Proposal for an all-spin artificial neural network: Emulating neural and synaptic functionalities through domain wall motion in ferromagnets," IEEE Trans. Biomed. Circuits Syst. **10**, 1152–1160 (2016).
[115]A. F. Vincent, J. Larroque, N. Locatelli, N. B. Romdhane, O. Bichler, C. Gamrat, W. S. Zhao, J.-O. Klein, S. Galdin-Retailleau, and D. Querlioz, "Spin-transfer torque magnetic memory as a stochastic memristive synapse for neuromorphic systems," IEEE Trans. Biomed. Circuits Syst. **9**, 166–174 (2015).
[116]G. Srinivasan, A. Sengupta, and K. Roy, "Magnetic tunnel junction based long-term short-term stochastic synapse for a spiking neural network with on-chip stdp learning," Sci. Rep. **6**, 29545 (2016).
[117]D. Zhang, L. Zeng, Y. Zhang, W. Zhao, and J. O. Klein, "Stochastic spintronic device based synapses and spiking neurons for neuromorphic computation," in *2016 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)* (IEEE, 2016), pp. 173–178.
[118]A. Sengupta and K. Roy, "Short-term plasticity and long-term potentiation in magnetic tunnel junctions: Towards volatile synapses," Phys. Rev. Appl. **5**, 024012 (2016).
[119]M. Romera, P. Talatchian, S. Tsunegi, F. A. Araujo, V. Cros, P. Bortolotti, J. Trastoy, K. Yakushiji, A. Fukushima, H. Kubota *et al.*, "Vowel recognition with four coupled spin-torque nano-oscillators," Nature **563**, 230 (2018).
[120]A. Hirohata, H. Sukegawa, H. Yanagihara, I. Žutić, T. Seki, S. Mizukami, and R. Swaminathan, "Roadmap for emerging materials for spintronic device applications," IEEE Trans. Magn. **51**, 1–11 (2015).
[121]A. Sengupta, M. Parsa, B. Han, and K. Roy, "Probabilistic deep spiking neural systems enabled by magnetic tunnel junction," IEEE Trans. Electron Devices **63**, 2963–2970 (2016).
[122]S. Ikeda, J. Hayakawa, Y. Ashizawa, Y. Lee, K. Miura, H. Hasegawa, M. Tsunoda, F. Matsukura, and H. Ohno, "Tunnel magnetoresistance of 604% at 300 K by suppression of ta diffusion in Co Fe B/Mg O/Co Fe B pseudo-spin-valves annealed at high temperature," Appl. Phys. Lett. **93**, 082508 (2008).
[123]H. Mulaosmanovic, E. Chicca, M. Bertele, T. Mikolajick, and S. Slesazeck, "Mimicking biological neurons with a nanoscale ferroelectric transistor," Nanoscale **10**, 21755–21763 (2018).
[124]H. Mulaosmanovic, J. Ocker, S. Müller, M. Noack, J. Müller, P. Polakowski, T. Mikolajick, and S. Slesazeck, "Novel ferroelectric FET based synapse for neuromorphic systems," in *2017 Symposium on VLSI Technology* (IEEE, 2017), pp. T176–T177.
[125]H. Mulaosmanovic, T. Mikolajick, and S. Slesazeck, "Accumulative polarization reversal in nanoscale ferroelectric transistors," ACS Appl. Mater. Interfaces **10**, 23997–24002 (2018).
[126]A. K. Saha, K. Ni, S. Dutta, S. Datta *et al.*, "Phase field modeling of domain dynamics and polarization accumulation in ferroelectric HZO," arXiv:1901.07121 (2019).
[127]S. Dutta, A. K. Saha, P. Panda, W. Chakraborty, J. Gomez, A. Khanna, S. Gupta, K. Roy, and S. Datta, "Biologically plausible energy-efficient ferroelectric quasi-leaky integrate and fire neuron," in Symposium on VLSI Technology (2019).
[128]S. Dünkel, M. Trentzsch, R. Richter, P. Moll, C. Fuchs, O. Gehring, M. Majer, S. Wittek, B. Müller, T. Melde *et al.*, "A FEFET based super-low-power ultra-fast embedded NVM technology for 22 nm FDSOI and beyond," in *2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2017), p. 19.
[129]A. K. Saha and S. K. Gupta, "Modeling and comparative analysis of hysteretic ferroelectric and anti-ferroelectric FETs," in *2018 76th Device Research Conference (DRC)* (IEEE, 2018), pp. 1–2.
[130]M. Jerry, P.-Y. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu, and S. Datta, "Ferroelectric FET analog synapse for acceleration of deep neural network training," in *2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2017), p. 6.
[131]W. Chung, M. Si, and D. Y. Peide, "First demonstration of GE ferroelectric nanowire FET as synaptic device for online learning in neural network with high number of conductance state and Gmax/Gmin," in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2018), p. 15.
[132]B. Obradovic, T. Rakshit, R. Hatcher, J. Kittl, R. Sengupta, J. G. Hong, and M. S. Rodder, "A multi-bit neuromorphic weight cell using ferroelectric FETs, suitable for SOC integration," IEEE J. Electron Devices Soc. **6**, 438–448 (2018).
[133]V. Kornijcuk, H. Lim, J. Y. Seok, G. Kim, S. K. Kim, I. Kim, B. J. Choi, and D. S. Jeong, "Leaky integrate-and-fire neuron circuit based on floating-gate integrator," Front. Neurosci. **10**, 212 (2016).
[134]D. Kahng and S. M. Sze, "A floating gate and its application to memory devices," Bell Syst. Tech. J. **46**, 1288–1295 (1967).
[135]R. H. Fowler and L. Nordheim, "Electron emission in intense electric fields," Proc. R. Soc. London, Ser. A **119**, 173–181 (1928).
[136]M. Lenzlinger and E. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2," J. Appl. Phys. **40**, 278–283 (1969).
[137]T.-S. Jung, Y.-J. Choi, K.-D. Suh, B.-H. Suh, J.-K. Kim, Y.-H. Lim, Y.-N. Koh, J.-W. Park, K.-J. Lee, J.-H. Park *et al.*, "A 3.3 V 128 Mb multi-level NAND flash memory for mass storage applications," in *1996 IEEE International Solid-State Circuits Conference. Digest of Technical Papers, ISSCC* (IEEE, 1996), pp. 32–33.
[138]M. Bauer, R. Alexis, G. Atwood, B. Baltar, A. Fazio, K. Frary, M. Hensel, M. Ishac, J. Javanifard, M. Landgraf *et al.*, "A multilevel-cell 32 Mb flash memory," in *Proceedings 30th IEEE International Symposium on Multiple-Valued Logic (ISMVL 2000)* (IEEE, 2000), pp. 367–368.
[139]M. Holler, S. Tam, H. Castro, and R. Benson, "An electrically trainable artificial neural network (ETANN) with 10240 floating gate synapses," in International Joint Conference on Neural Networks (1989), Vol. 2, pp. 191–196.
[140]B. W. Lee, B. J. Sheu, and H. Yang, "Analog floating-gate synapses for general-purpose VLSI neural computation," IEEE Trans. Circuits Syst. **38**, 654–658 (1991).
[141]P. E. Hasler, C. Diorio, B. A. Minch, and C. Mead, "Single transistor learning synapses," in *Advances in Neural Information Processing Systems* (Curran Associates, 1995), pp. 817–824.
[142]S.-C. Liu and R. Mockel, "Temporally learning floating-gate VLSI synapses," in *2008 IEEE International Symposium on Circuits and Systems* (IEEE, 2008), pp. 2154–2157.
[143]S. Ramakrishnan, P. E. Hasler, and C. Gordon, "Floating gate synapses with spike-time-dependent plasticity," IEEE Trans. Biomed. Circuits Syst. **5**, 244–252 (2011).
[144]A. Ankit, A. Sengupta, P. Panda, and K. Roy, "RESPARC: A reconfigurable and energy-efficient architecture with memristive crossbars for deep spiking neural networks," in *Proceedings of the 54th Annual Design Automation Conference 2017* (ACM, 2017), p. 27.
[145]A. Ankit, I. E. Hajj, S. R. Chalamalasetti, G. Ndu, M. Foltin, R. S. Williams, P. Faraboschi, W.-M. W. Hwu, J. P. Strachan, K. Roy *et al.*, "Puma: A programmable ultra-efficient memristor-based accelerator for machine learning inference," in *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems* (ACM, 2019), pp. 715–731.
[146]A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," ACM SIGARCH Comput. Archit. News **44**, 14–26 (2016).
[147]P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, "Prime: A novel processing-in-memory architecture for neural network computation in ReRAM-based main memory," in *Proceedings of the 43rd International Symposium on Computer Architecture* (IEEE Press, 2016), pp. 27–39.
[148]S. G. Ramasubramanian, R. Venkatesan, M. Sharad, K. Roy, and A. Raghunathan, "Spindle: Spintronic deep learning engine for large-scale neuromorphic computing," in *Proceedings of the 2014 international symposium on Low power electronics and design* (ACM, 2014), pp. 15–20.
[149]Y. Kim, Y. Zhang, and P. Li, "A reconfigurable digital neuromorphic processor with memristive synaptic crossbar for cognitive computing," ACM J. Emerging Technol. Comput. Syst. **11**, 1 (2015).
[150]N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)* (IEEE, 2017), pp. 1–12.

Published under license by AIP Publishing

[151]A. Yousefzadeh, E. Stromatias, M. Soto, T. Serrano-Gotarredona, and B. Linares-Barranco, "On practical issues for stochastic STDP hardware with 1-bit synaptic weights," Front. Neurosci. **12**, 665 (2018).
[152]G. Srinivasan and K. Roy, "Restocnet: Residual stochastic binary convolutional spiking neural network for memory-efficient neuromorphic computing," Front. Neurosci. **13**, 189 (2019).
[153]S. Agarwal, T.-T. Quach, O. Parekh, A. H. Hsia, E. P. DeBenedictis, C. D. James, M. J. Marinella, and J. B. Aimone, "Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding," Front. Neurosci. **9**, 484 (2016).
[154]P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," Science **345**, 668–673 (2014).
[155]M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," IEEE Micro **38**, 82–99 (2018).
[156]C. M. Liyanagedera, A. Sengupta, A. Jaiswal, and K. Roy, "Stochastic spiking neural networks enabled by magnetic tunnel junctions: From nontelegraphic to telegraphic switching regimes," Phys. Rev. Appl. **8**, 064017 (2017).
[157]E. O. Neftci, B. U. Pedroni, S. Joshi, M. Al-Shedivat, and G. Cauwenberghs, "Stochastic synapses enable efficient brain-inspired learning machines," Front. Neurosci. **10**, 241 (2016).
[158]W. Senn and S. Fusi, "Convergence of stochastic learning in perceptrons with binary synapses," Phys. Rev. E **71**, 061907 (2005).
[159]S. Yu, X. Guan, and H.-S. P. Wong, "On the stochastic nature of resistive switching in metal oxide RRAM: Physical modeling, Monte Carlo simulation, and experimental characterization," in *2011 International Electron Devices Meeting* (IEEE, 2011) p. 17.
[160]D. Ielmini, A. L. Lacaita, and D. Mantegazza, "Recovery and drift dynamics of resistance and threshold voltages in phase-change memories," IEEE Trans. Electron Devices **54**, 308–315 (2007).
[161]S. Jain, A. Sengupta, K. Roy, and A. Raghunathan, "Rx-Caffe: Framework for evaluating and training deep neural networks on resistive crossbars," arXiv:1809.00072 (2018).
[162]I. Chakraborty, D. Roy, and K. Roy, "Technology aware training in memristive neuromorphic systems for nonideal synaptic crossbars," IEEE Trans. Emerging Top. Comput. Intell. **2**, 335–344 (2018).
[163]Y. Jeong, M. A. Zidan, and W. D. Lu, "Parasitic effect analysis in memristor-array-based neuromorphic systems," IEEE Trans. Nanotechnol. **17**, 184–193 (2018).
[164]C. Liu, M. Hu, J. P. Strachan, and H. H. Li, "Rescuing memristor-based neuromorphic design with high defects," in *Proceedings of the 54th Annual Design Automation Conference 2017* (ACM, 2017), p. 87.
[165]L. Chen, J. Li, Y. Chen, Q. Deng, J. Shen, X. Liang, and L. Jiang, "Accelerator-friendly neural-network training: Learning variations and defects in RRAM crossbar," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)* (IEEE, 2017).
[166]P.-Y. Chen, B. Lin, I.-T. Wang, T.-H. Hou, J. Ye, S. Vrudhula, J. S. Seo, Y. Cao, and S. Yu, "Mitigating effects of non-ideal synaptic device characteristics for on-chip learning," in *2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (IEEE, 2015).
[167]I. Kataeva, F. Merrikh-Bayat, E. Zamanidoost, and D. Strukov, "Efficient training algorithms for neural networks based on memristive crossbar circuits," in *2015 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2015).
[168]M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila, C. Graves, S. Lam, N. Ge, J. J. Yang, and R. S. Williams, "Dot-product engine for neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication," in *Proceedings of the 53rd Annual Design Automation Conference* (ACM, 2016), p. 19.
[169]B. Liu, H. Li, Y. Chen, X. Li, T. Huang, Q. Wu, and M. Barnell, "Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems," in *2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (IEEE, 2014).
[170]I. Boybat, M. Le Gallo, S. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian, and E. Eleftheriou, "Neuromorphic computing with multi-memristive synapses," Nat. Commun. **9**, 2514 (2018).
[171]J. H. Lee, T. Delbruck, and M. Pfeiffer, "Training deep spiking neural networks using backpropagation," Front. Neurosci. **10**, 508 (2016).
[172]Y. Wu, L. Deng, G. Li, J. Zhu, and L. Shi, "Spatio-temporal backpropagation for training high-performance spiking neural networks," Front. Neurosci. **12**, 331 (2018).
[173]C. Lee, P. Panda, G. Srinivasan, and K. Roy, "Training deep spiking convolutional neural networks with STDP-based unsupervised pre-training followed by supervised fine-tuning," Front. Neurosci. **12**, 435 (2018).
[174]D. Lee, X. Fong, and K. Roy, "R-MRAM: A ROM-embedded STT MRAM cache," IEEE Electron Device Lett. **34**, 1256–1258 (2013).
[175]S. Jain, A. Ranjan, K. Roy, and A. Raghunathan, "Computing in memory with spin-transfer torque magnetic RAM," IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **26**, 470–483 (2018).
[176]J. Zhou, K.-H. Kim, and W. Lu, "Crossbar RRAM arrays: Selector device requirements during read operation," IEEE Trans. Electron Devices **61**, 1369–1376 (2014).
[177]Y.-W. Chin, S.-E. Chen, M.-C. Hsieh, T.-S. Chang, C. J. Lin, and Y.-C. King, "Point twin-bit RRAM in 3D interweaved cross-point array by cu BEOL process," in *2014 IEEE International Electron Devices Meeting* (IEEE, 2014), p. 6.
[178]T. Ohyanagi, N. Takaura, M. Kitamura, M. Tai, M. Kinoshita, K. Akita, T. Morikawa, and J. Tominaga, "Superlattice phase change memory fabrication process for back end of line devices," Jpn. J. Appl. Phys., Part 1 **52**, 05FF01 (2013).
[179]H. Ohno, T. Endoh, T. Hanyu, N. Kasai, and S. Ikeda, "Magnetic tunnel junction for nonvolatile CMOS logic," in *2010 International Electron Devices Meeting* (IEEE, 2010), p. 9.
[180]H. Noguchi, K. Ikegami, S. Takaya, E. Arima, K. Kushida, A. Kawasumi, H. Hara, K. Abe, N. Shimomura, J. Ito *et al.*, "7.2 4Mb STT-MRAM-based cache with memory-access-aware power optimization and write-verify-write/read-modify-write scheme," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2016), pp. 132–133.

Published under license by AIP Publishing