Memristive Devices and Networks for Brain-Inspired Computing

Teng Zhang, Ke Yang, Xiaoyan Xu, Yimao Cai, Yuchao Yang, and Ru Huang**

As the era of big data approaches, conventional digital computers face increasing difficulties in performance and power efficiency due to their von Neumann architecture. As a result, there is recently a tremendous upsurge of investigations on brain-inspired neuromorphic hardware with high parallelism and improved efficiency. Memristors are considered as promising building blocks for the realization of artificial synapses and neurons and can therefore be utilized to construct hardware neural networks. Here, a review is provided on existing approaches for the implementation of artificial synapses and neurons based on memristive devices; and the respective advantages and disadvantages of these approaches are evaluated. This is followed by a discussion of hardware accelerators and neuromorphic computing systems that exploit the parallel, in-memory and analog characteristics of memristive crossbar arrays as well as the intrinsic dynamics of memristors. Finally, the outstanding challenges are addressed that have not yet been resolved in the present studies, and future advances are discussed that might be needed for building intelligent and energy efficient neuromorphic systems.

1. Introduction

Conventional digital computers are based on von Neumann architecture, where the logic and memory units are physically separated from each other. As a consequence, the time and energy spent on transporting data lead to a fundamental bottleneck in terms of performance and power efficiency.[1] To address this issue and obtain a new-generation of computers with improved efficiency, an in-memory computing architecture with high parallelism is desired, and a typical example existing in nature falling into this category is the human brain. The human brain is a highly intelligent and efficient computing system capable of performing complicated cognitive tasks such as learning, inference, abstraction and generalization, etc. that cannot be found in existing digital computers, yet operated at a low frequency of ≈10 Hz and low power of ≈20 W.[2] The information processing capability in the human brain is based upon biological neural networks containing ≈10^{11} neurons and ≈10^{15} synapses. The neurons can alter their membrane potential in response to the inputs from the dendrites and fire spikes through the axon when exceeding certain thresholds, thus acting as distributed computing units.[3] These neurons are in turn connected via synapses whose strengths can be varied by neuronal activities, and such plasticity is responsible for the formation of memory and learning.[3] Taking inspiration from the architecture and principle of the human brain, neuromorphic computing aims to construct energy efficient and fault tolerant computing systems and may have dramatic potential in future artificial intelligence (AI), big data, Internet of Things (IoT), etc.[4] In particular, devices that can emulate the functionalities of biological synapses and neurons are of fundamental importance in the construction of such neuromorphic systems.

Recently, there have been encouraging advances in artificial intelligence. In particular, convolutional neural networks (CNN)[5] and recurrent neural networks (RNN)[6] have achieved high accuracy in visual recognition, speech recognition, and other applications, even surpassing the performance of human in many tasks.[7] It should be noted that at the heart of most machine learning algorithms is the implementation of some very simple operations. For example, 666 million multiply and accumulate operations are involved in the inference pass for a single 227 × 227 image in AlexNet,[5] while 15.3 billion multiply and accumulate operations are required for a single 224 × 224 image in VGG16.[8] Such operations consume enormous time and energy in conventional von Neumann computers,[1] and neural network accelerators with high throughput and high energy efficiency will be the first but significant step toward the realization of neuromorphic systems.

The emergence of memristors, as the fourth passive circuit element,[9] offered a promising candidate for the construction of neural network accelerators and neuromorphic systems. The general definition of a memristor refers to a device displaying pinched hysteresis loops in the voltage–current plot under periodic stimulations, including devices based on ionic resistive switching, phase change and spin-transfer torque mechanisms, etc.[10] Since memristors can be analog and nonvolatile,[11–13] the

T. Zhang, K. Yang, Dr. X. Xu, Prof. Y. Cai, Prof. Y. Yang, Prof. R. Huang
Key Laboratory of Microelectronic Devices and Circuits (MOE)
Institute of Microelectronics
Peking University
Beijing 100871, P. R. China
E-mail: yuchaoyang@pku.edu.cn; ruhuang@pku.edu.cn

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/pssr.201900029.

DOI: 10.1002/pssr.201900029

vector matrix multiplication (VMM) or weighted summation operation can be physically implemented in memristor arrays with a crossbar configuration, as illustrated in **Figure 1**. In this case, the analog weights of connections in neural networks are mapped into the memristor conductance values G_{ij} at each crosspoint. The input vector x is converted into input voltages and applied to the rows of the crossbar, where x can be represented either in the amplitude or the width of the input voltage pulses. In the former case, namely, when x is represented by the amplitude of the pulses, current flowing through each device is x_iG_{ij} based on the Ohm's law, and thus the total current collected at each column is $I_j = \sum_i x_iG_{ij}$ according to the Kirchhoff's current law. If x is represented by the pulse width instead, the total charge collected at each column will also be proportional to $\sum_i x_iG_{ij}$. Therefore, in both cases the VMM operation can be conveniently achieved in a parallel, in-memory, and analog manner in the memristor crossbar within a single read cycle, which could largely benefit data-centric and VMM intensive algorithms, including a large majority of artificial neural network (ANN) algorithms and many other arithmetic calculations.

Aside from the nonvolatile and analog nature that can be used for hardware acceleration, the resistive switching in memristors is in fact a complicated phenomenon that can incorporate a variety of physical, electrochemical or thermochemical processes, mediated by dynamic ion transport and redox reactions under high electric field and/or elevated temperature.[12,14–19] Such intrinsic properties and dynamics can be exploited to emulate more biologically plausible functionalities of synapses and neurons, for instance, short-term plasticity (STP),[20,21] spike timing-dependent plasticity,[22] heterosynaptic plasticity,[23,24] and leaky integrate and fire dynamics of neurons,[25] etc. These rich functionalities form the basis for encoding and processing complex information, e.g., spiking information based on the rate or sequence of spikes. This approach bears better similarity with

Yuchao Yang received his Ph.D. from Tsinghua University, China. After that he joined University of Michigan, Ann Arbor, USA, as a postdoctoral Research Fellow and was promoted to Senior Research Fellow in 2013. He is now an assistant professor and Boya Young Scholar in Peking University. His research interests include microscopic physics and dynamics of memristors, synaptic/neuronal devices and their neuromorphic as well as in-memory computing applications. He is a member of IEEE, MRS, RSC, and ACS. He is a recipient of awards including the Qiu Shi Outstanding Young Scholar Award and MIT Technology Review Innovators Under 35 in China.

Ru Huang is currently a professor at Peking University, China. She is an elected academician of Chinese Academy of Science and IEEE Fellow. Her research interests include nano-scaled CMOS devices, ultra-low-power new devices, new devices for neuromorphic computing, emerging memory technology and device variability/reliability. She has authored or coauthored 5 books, 3 book chapters and more than 300 papers, including more than 80 papers in Electron Devices Meet. (IEDM), VLSI Technol. Symp., IEEE Electron Device Lett. and IEEE Trans. Electron Devices, and gave more than 40 invited talks at international conferences. She is the holder of more than 200 granted patents.

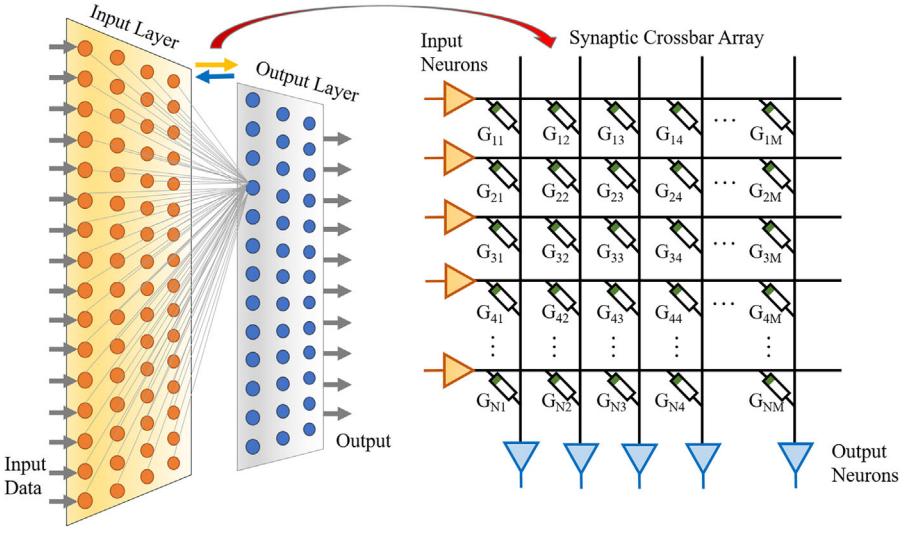

Figure 1. Schematic illustration showing that the vector-matrix multiplication can be performed in memristive crossbar array on a single read cycle, while the conductance of each cell at the crosspoint could be updated by applying programming voltages from row/column at the same time. Both forward inference and backpropagation of neural networks can thus be implemented in hardware with high efficiency.

the human brain and could be important for the construction of neuromorphic systems with intelligence and high efficiency.

This review is organized as follows. Firstly, we will provide a summary and comparison on existing approaches for the implementation of artificial synapses and neurons based on memristive devices. Subsequently, we will give an overview on hardware accelerators as well as neuromorphic computing systems that have been built based on memristors, by exploiting the parallelism of memristive crossbar and the intrinsic dynamics of memristors. In the end, we will shed light on outstanding challenges that have yet been addressed at present and discuss future advances that might be pertinent for memristor-based neuromorphic systems.

2. Memristive Synapses

In biological nervous systems, synapse refers to the structure through which the impulse of one neuron passes to another,[3] as shown in **Figure 2**, which plays a key role of information transmission based on its unique plasticity. When an action potential arrives at the presynaptic terminal, voltage-gated Ca^{2+} channels are open, which produces a high concentration of Ca^{2+} inside the membrane and causes vesicles there containing neurotransmitters to fuse with the membrane, and hence the neurotransmitters can be released into the synaptic cleft. These neurotransmitter molecules then diffuse across the synaptic cleft and bind with specific receptors on the postsynaptic membrane, causing ion channels to open or close. As a result, the membrane conductance and membrane potential of the postsynaptic cell are changed. The receptor activation can translate electrical signals into biochemical signals, among which some lead to long lasting changes (hours or days) in synaptic strength,[3] so called long-term plasticity (LTP) that is thought to be crucial during synaptic development and for regulating neural circuits in the adult brain. A typical form of LTP is the so-called spike timing-dependent plasticity (STDP), where the synaptic efficacy is enhanced or depressed depending on the relative timing between presynaptic and postsynaptic spikes, and this learning rule plays an essential role in building causality correlations.[3] As a result, LTP is a basic function to implement in artificial synapses. This also meets the requirement in the training of neural networks, where the weights need to be constantly adjusted when training the neural networks. In addition to LTP, the change in membrane conductance might only last for a short time (tens of milliseconds to minutes) instead, which is called STP and has great significance for processing temporal information on the relevant time scales.[3,26–28] Realization of STP in artificial synapses is also very important depending on detailed applications.

To date, encouraging progress has been achieved in the implementation of both LTP and STP using memristive devices, which offers several clear advantages compared with existing CMOS approaches.[29,30] First, owing to the simple two-terminal structure and compact size, memristive devices can potentially achieve a much higher synaptic density compared with pure CMOS-based synaptic circuits, since usually a large number of transistors are required (>10) per synapse.[31] Second, the two-terminal structure decides that the synaptic array can be arranged in a crossbar structure, which allows high degree of connectivity as well as high parallelism and desired for improved computing efficiency. Furthermore, memristive devices also provide the much needed nonvolatility to the neuromorphic systems, and this allows memristor-based neural networks to constantly learn from the data and is also beneficial in reducing static power consumption. Last but not the least, analog synaptic weights can be achieved using memristive devices with continuously variable conductance, and this might be used to achieve online learning, where the synaptic weights are continuously adjusted on the fly in hardware when new data samples are applied. Compared with offline learning where the synaptic weights are calculated in software and loaded into

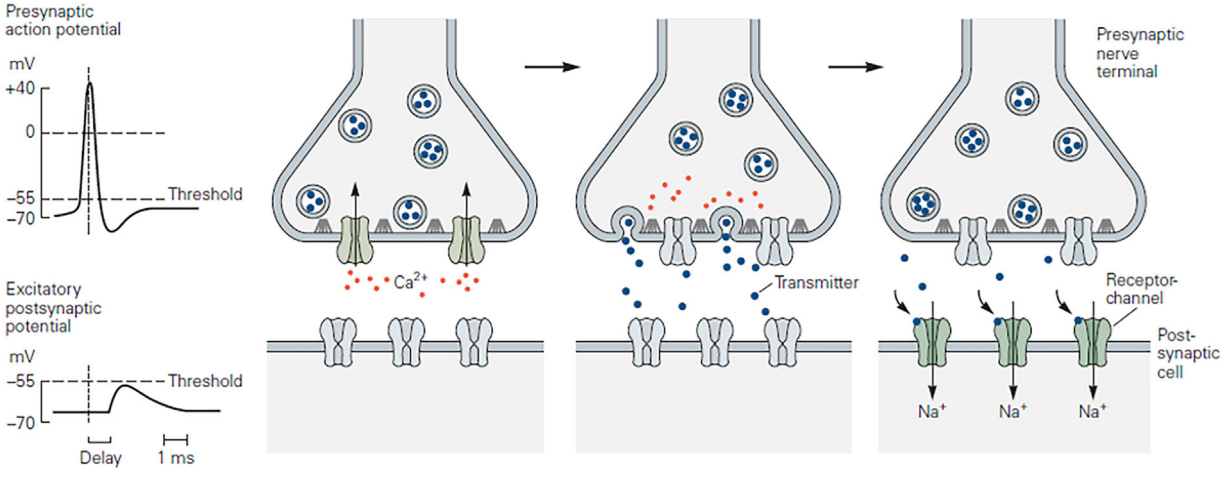

Figure 2. Synaptic transmission in biological nervous systems. First, an action potential arriving at the terminal of a presynaptic terminal causes voltage-gated Ca^{2+} channels at the active zone to open. A high concentration of intracellular Ca^{2+} thus causes vesicles containing neurotransmitter to fuse with the membrane and release their contents into the synaptic cleft. The neurotransmitter molecules diffuse across the synaptic cleft and bind with specific receptors on the postsynaptic membrane, which causes ion channels to open (or close), thereby changing the membrane conductance and membrane potential of the postsynaptic cell. Reproduced with permission.[3] Copyright 2013, The McGraw-Hill Companies, Inc.

hardware afterwards, implementation of online learning is obviously more demanding for device performance and memristive synapses could have a great prospect in building online learning systems.

Table 1 gives a rough estimation on desired performance metrics for artificial synapses. In particular, a dynamic range of 100, weight precision of 6 bit and low energy consumption that is comparable with biological synapses (≈10 fJ) might be desirable for many applications.[32] It is worthwhile pointing out that other requirements on device performance may strongly depend on detailed applications,[33] and Table 1 only suggests typical values. For example, an endurance of 10^5 may be required to train state-of-the-art neural networks in an online fashion,[34] but it should be noted that the device conductance only needs to be modified by an incremental amount instead of being switched across the whole dynamic range in each iteration, making the requirement for endurance a bit relaxed. Furthermore, device endurance is less of a concern in offline learning, since the synaptic weights do not need to be frequently updated. Similarly, high weight update linearity is generally important and favorable for online learning but is not a big concern in offline learning. Nevertheless, offline learning applications places a high requirement on retention (>10 years) to allow reliable inference after the synaptic weights in the arrays are programmed. However, such retention requirement can be relaxed in online learning, where the weights are updated frequently.[33]

Below we discuss the artificial synapses based on memristive devices reported recently. A crude classification based on the physical mechanism and material system of these synaptic devices can divide them into metal ion transport-based synapses, oxygen ion transport-based synapses, phase change synapses, and others approaches.

2.1. Metal Ion Transport-Based Synapses

Conductive bridge random access memroy (CBRAM), also called electrochemical metallization (ECM) cells or programmable metallization cells (PMC), have been widely used for the realization of artificial synapses. The resistive switching in such devices is based on the transport of metal cations, such as Ag and Cu cations, and the accompanying electrochemical redox processes.[12,15,17,18,35] In general, a CBRAM structure includes an active electrode (Ag or Cu), an inert electrode (Pt, Au or W), and a solid electrolyte layer sandwiched in between.[36] A positive voltage applied to the active electrode leads to the oxidation of metal atoms and growth of a metal filament in the switching layer, which subsequently forms a conducting channel bridging the two electrodes and increases the device conductance. This physical process can be mapped to achieve synaptic facilitation or increase the connection weight. As a reverse process, the filament can be dissolved by applying a negative voltage, which returns the cell to a low conductance state and can be utilized to achieve synaptic depression or decrease the connection weight.

Table 1. Desired performance metrics for artificial synapses.

Parameters	Desired value
Dynamic range	>100
Weight precision	>64
Linearity	<4 for set and >−4 for reset
Energy consumption	<10 fJ/spike
Endurance	$>10^9$
Retention	>10 years

The advantage of metal ion transport-based synapses lies in their high scalability and large dynamic range. A sub-20 nm device has been demonstrated using a Ag–Ge–Se electrolyte,[37] while extremely high scalability down to atomic level was experimentally verified using similarly metal ion transport-based atomic switches,[21,38] as shown in **Figure 3**(a), where chalcogenides-based electrolytes (such as Ag_2S) are usually adopted. The dynamic range of the overall conductance change can reach a few orders of magnitude due to the high conductance offered by the metal filament(s) and the insulating nature of solid electrolytes.[39] Besides, high endurance, long retention, and high operation speed have also been obtained in metal ion transport-based synapses.[40] Notably, important biologically inspired learning rules such as STDP, have also been achieved.[41]

One of the disadvantages of CBRAM-based artificial synapses is the usually abrupt nature of the filament formation process, which in turn results in limited number of weight states and low weight update linearity. Although the subsequent filament broadening process after filament formation can be gradual in nature (Figure 3(a)), the dynamic range originating from this process is usually small and the states are relatively conductive (Figure 3(b)), leading to large operation current that is undesirable. This issue might be addressed by tailoring the metal ion transport in CBRAM, for example, by co-sputtering Ag and amorphous Si and forming a more gradual Ag concentration gradient in the electrolyte, which gave rise to more analog resistive switching.[41] Another approach to mitigating the abrupt switching as a result of metal filament formation in CBRAM-based synapses might be optimization of the filament composition. For example, a subquantum CBRAM cell based on the migration and redistribution of Te ions was proposed recently, where the device conductance during programming tends to increase in increments of 1-atom conductance, which is much smaller than the conductance quantum due to the semiconducting nature of Te (≈0.03 G_0) and can potentially provide more gradual and linear weight updates.[42] However, since the filament formation process is stochastic in nature, CBRAM-based synapses usually show inevitable variability.[43]

2.2. Oxygen Ion Transport-Based Synapses

In addition to metal ion transport, resistive switching in memristors can also be induced by the migration and redistribution of oxygen ions or equivalently positively charged oxygen vacancies (V_Os) in a large variety of oxide systems, especially transition metal oxides.[12,15,19] Such oxide-based systems usually have good compatibility with existing CMOS

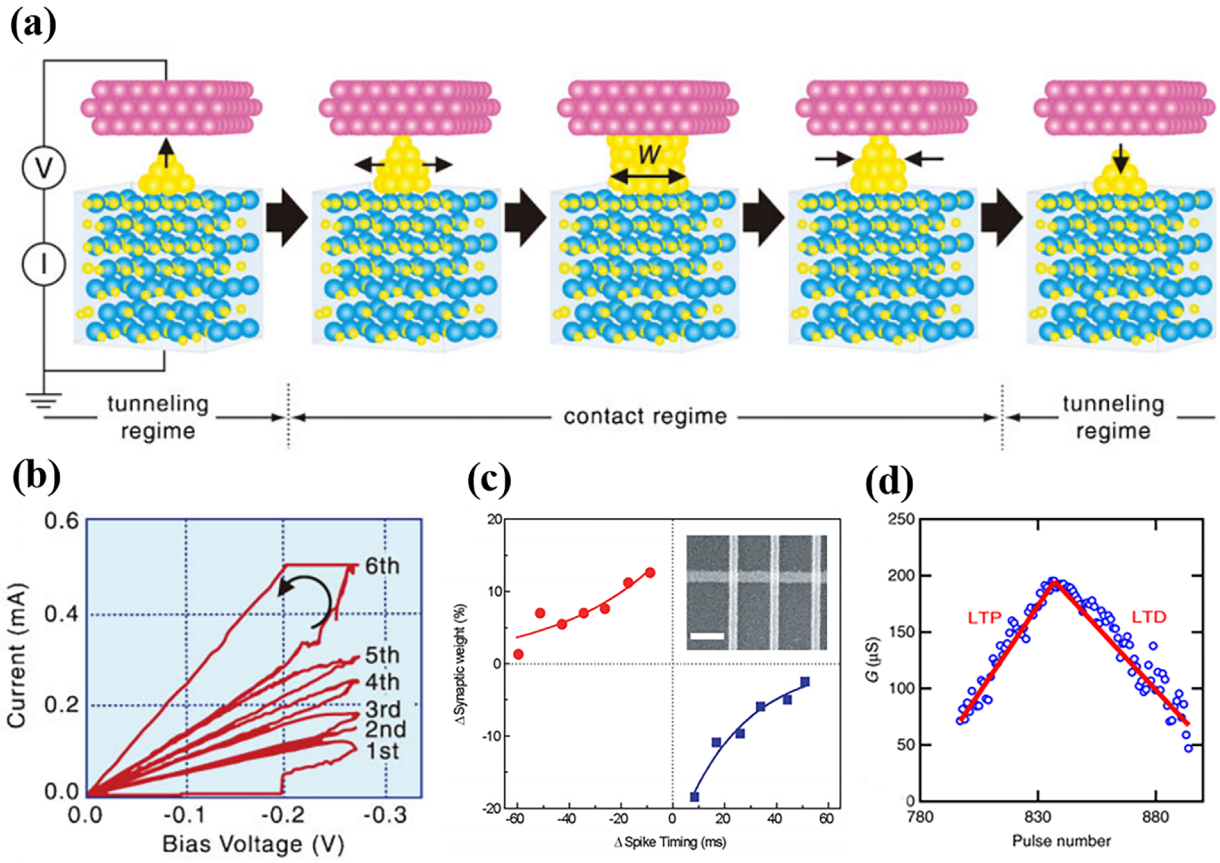

Figure 3. a) Schematic illustration of an artificial synapse based on Ag_2S atomic switch. b) Experimental result of a gradual increase under DC voltage sweep from the device in (a). Reproduced with permission.[38] Copyright 2012, Wiley-VCH. c) Demonstration of STDP behavior using Ag/Si based memristive synapse. Reproduced with permission.[41] Copyright 2010, American Chemical Society. d) Implementation of LTP and LTD in a subquantum CBRAM synapse using stepwise voltage pulses. Reproduced with permission.[42] Copyright 2018, Springer Nature Publishing AG.

processes. The resistive switching therein can be dominated either by localized conducting filaments, electrode/dielectric interfaces, or bulk effects such as charge trapping/detrapping. Below we discuss oxygen ion transport-based artificial synapses by categorizing them into filamentary and non-filamentary devices.

2.2.1. Filamentary Type

Similar with CBRAM devices, a conducting filament can also be formed in oxides via the migration and accumulation of oxygen vacancies, and the resultant filaments are usually composed of a sub-oxide phase that has higher conductivity than the original matrix.[44] Such filamentary switching has been realized in a large number of oxide systems, such as HfO_x,[45] TaO_x,[46] AlO_x,[47] WO_x,[48] and TiO_x[49] as well as oxide bilayers/multilayers.[50] Analog conductance modulation that is crucial to imitate synaptic weight change has been demonstrated in most of these oxide systems, and detailed characteristics of the analog switching, including the dynamic range, number of resistance levels, linearity, symmetry and power consumption etc., significantly affect the network performance and require careful optimizations.

Among the above performance metrics, the weight update linearity when applying identical voltage pulses is important for realizing online learning in neural networks,[51] but filamentary oxide memristors without optimization typically exhibits low linearity where an initial abrupt increase (decrease) in device conductance is followed by a more gradual modulation regime (**Figure 4**(a)). Such linearity in conductance modulation has been optimized by introducing an oxygen ion diffusion limiting layer (DLL) into TaO_x-based devices,[46] i.e., SiO_2, so that a SiO_2/TaO_x bilayer structure is formed. Compared with the typically nonlinear conductance modulation in $TiN/TaO_x/Pt$ devices shown in Figure 4(a), $TiN/SiO_2/TaO_x/Pt$ devices clearly demonstrate improved linearity after introduction of the DLL, as can be seen in Figure 4(b–d) with different DLL thicknesses. Because of the sub-stoichiometric nature of the TaO_x layer, the applied voltage is mainly dropped across the SiO_2 film due to its more insulating property, and as a result the switching event tends to occur in the SiO_2 layer with filament formation and rupture processes. This therefore allows the inserted DLL layer with low oxygen ion mobility to effectively reduce the amount of oxygen

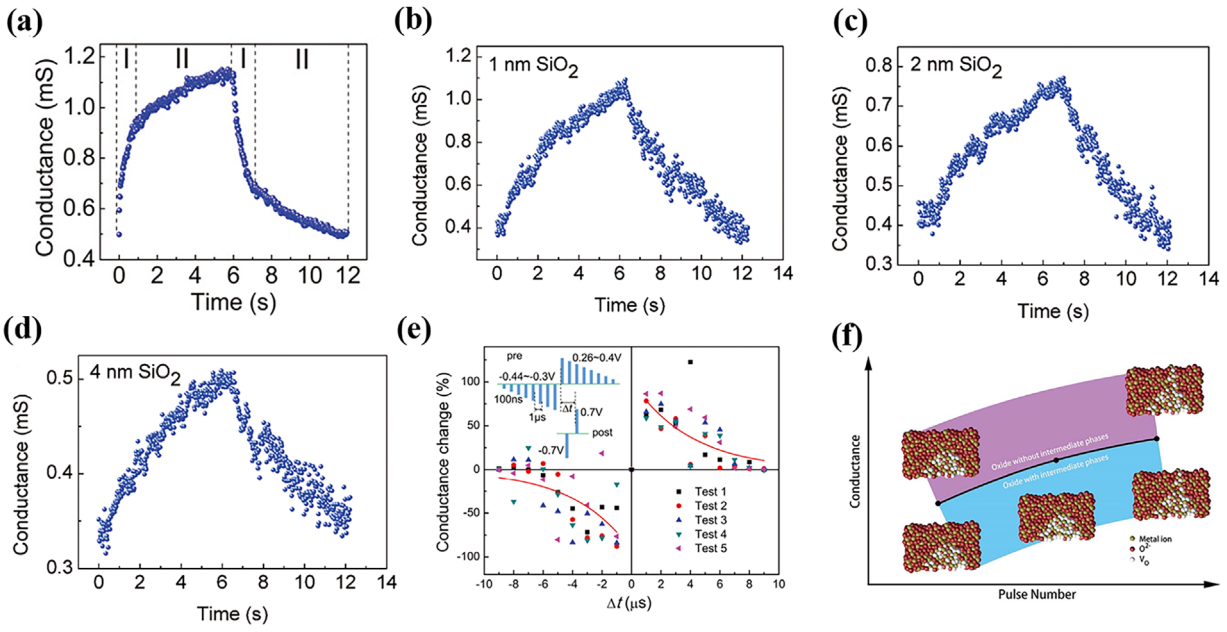

Figure 4. a) Typical nonlinear conductance modulation in TiN/TaO_x/Pt memristive synapse. b–d) Highly linear conductance modulation processes for TiN/SiO_2/TaO_x/Pt memristive synapse with oxygen diffusion limiting layer (SiO_2) thickness of 1 nm (b), 2 nm (c), and 4 nm (d). e) STDP behavior in TiN/SiO_2-1 nm/TaO_x/Pt memristive synapse. The inset shows the applied spikes to the pre-synaptic and post-synaptic terminals. Reproduced with permission.[46] Copyright 2016, The Royal Society of Chemistry. f) Schematic of the conductance change mechanism in oxide with and without intermediate phases, leading to different number of weight states. Reproduced with permission.[53] Copyright 2017, The Royal Society of Chemistry.

ions/vacancies that take part in the initial growth/dissolution so as to suppress regime I, hence leading to overall improved linearity in Figure 4(b–d).[46] Besides material optimization, the weight update linearity might also be improved by using different pulse schemes. Jeong et al.[52] proposed a pulse scheme including a long and low-amplitude heating pulse in addition to the regular SET pulses to improve the analog switching of TaO_x-based synaptic devices. Moreover, some complex learning rule like STDP (Figure 4(e)) can also be achieved by well-designed pulse overlaps or by exploiting inherent device dynamics, for example the internal temperature decay in second state memristors.[22]

Besides weight update linearity, the number of weight states is another important performance metric for both training and inference applications of hardware neural networks. In particular, a high weight precision is usually indispensable in order to achieve online training.[54] Recently, the weight precision in filamentary oxide memristors was found to be related to the phase diagram of the switching oxides used.[53] It is well understood that the facilitation or depression of filamentary synapses usually originates from geometric changes of the conducting filament, including its length or width. This imposes a constraint in weight precision for oxide synapses with a limited number of intermediate phases in their phase diagrams, e.g., TaO_x and HfO_x. In contrast, additional dimensionality might be exploited to increase the weight precision in oxides with rich intermediate oxide phases e.g., WO_x. Aside from the geometric changes of filaments, it is possible for some localized parts of the conduction channel to undergo phase transitions between the intermediate oxide phases upon application of facilitation and depression pulses (Figure 4(f)), which thus improves the incremental switching and increases the number of weight states.[53] However, it should be pointed out that different aspects of synaptic performance could require different material properties and therefore form a tradeoff. For instance, a simple phase diagram with a limited number of intermediate phases could be helpful for improving the endurance, and detailed selection of oxide systems should be considered based on the applications.

Similar with CBRAM-based synapses, the filament formation and rupture processes in filamentary oxide memristors can also result in abrupt conductance modulation and thus low linearity, and the changes in the shape, composition, and location of conducting filaments naturally lead to device variations. However, such stochasticity and abruptness can be mitigated by only switching the devices in binary fashion and adopting the probability of resistive switching instead as the synaptic weight. This has been found applicable to certain types of neural networks.[55] In order to address the device variation caused by filamentary switching, a fundamental means is to build artificial synapses based on non-filamentary mechanism, as be discussed next.

2.2.2. Non-Filamentary Type

Non-filamentary resistive switching typically relies on ion transport induced redox processes occurring at the interface,[56] metal–insulator transition[57] or purely electronic mechanism such as charge trapping/detrapping in the bulk,[58–60] etc. Since

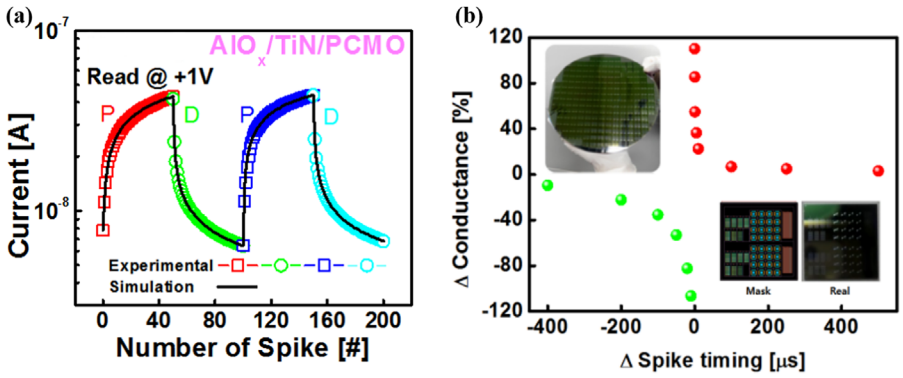

Figure 5. a) Experimental and simulation result for long-term facilitation and depression with identical spike in AlO_x/TiN/PCMO synaptic device. Reproduced with permission.[61] Copyright 2013, IEEE. b) STDP characteristic of 1k-bit W/Al/PCMO/Pt crossbar array. Reproduced with permission.[62] Copyright 2012, IEEE.

these processes are governed by the whole device area instead of a localized region, non-filamentary oxide memristors are expected to show higher switching uniformity compared with filamentary devices and therefore address the abovementioned issue. **Figure 5**(a) shows an example of the synaptic behavior in Pt/AlO_x/TiN/$Pr_{0.7}Ca_{0.3}MnO_3$ (PCMO)/Pt devices where the resistive switching can be attributed to the redox processes occurring at the interface between manganites and oxidizable electrodes,[61] and Figure 5(b) further presents the STDP learning rule achieved in the devices.[62] Besides, the authors of the present work[63] achieved transition from filamentary type to non-filamentary type switching by creating a continuous gradient distribution and fine regulation of oxygen vacancies in oxidized TaO_x film, thereby improving the uniformity of synapses. Besides the switching layer, the performance of non-filamentary synapses also largely depends on electrode materials. For example, Moon et al.[64] obtained improved device uniformity, analog characteristics and retention in Mo/PCMO-based synaptic device, where the improved retention can be explained by the high activation energy of the oxidation process and the high electronegativity of the Mo electrode. Similarly, other characteristics such as the current density and dynamic range can be improved in TiN/PCMO devices by varying sputtering conditions of the TiN electrode.[65] The weight update symmetry is another important performance metric in online learning, and it has been shown that the accuracy in pattern recognition can be improved by mitigating the asymmetric behavior in TiN/PCMO devices,[66] where a new PCMO-based synaptic structure[66] consisting of two PCMO devices was used to avoid abrupt conductance changes. It should be noted that although non-filamentary devices have shown advantages in addressing a few issues existing in filamentary synapses especially the device uniformity, they may have a smaller dynamic range and also face limitations in weight update linearity and symmetry, which requires further device optimizations.

2.3. Phase Change Synapses

In addition to the above memristive devices based on ionic transport and redox reactions, there exists another class of memristors whose resistive switching is induced by the phase transition between amorphous (high resistivity) and crystalline (low resistivity) states, which is named phase change memory (PCM). **Figure 6**(a) shows a typical structure of PCM cells.[67] A pulse with a small amplitude and large width will be needed to anneal the switching region (Figure 6(b)) during the set process,[67] thus leading to a phase transition to the crystalline state. During the reset process, a pulse with a large amplitude and small width is applied, and the switching region is melted and then quenched quickly (Figure 6(b)), leaving a region of amorphous state that is highly resistive. The PCM cells are typically based on chalcogenides, such as $Ge_2Sb_2Te_5$ (GST),[68,69] GeTe/Sb_2Te_3,[69,70] $Ag_5In_5Sb_{60}Te_{30}$ (AIST),[69] etc.

To date, long-term facilitation (LTF) and long-term depression (LTD) processes have been achieved using phase change synapses. Similarly, by using carefully designed pulse shapes, the STDP learning rule can also be implemented using phase change synapses.[71] One of the advantages for phase change synapses is their large dynamic range due to the significantly different resistance of amorphous and crystalline states, e.g., over five orders of magnitude in some systems.[72] Besides, switching of a PCM cell occurs in tens of nanoseconds, or even faster.[73,74] The scaling limit of PCM depends on the minimum volume in which the material can exist stably in both amorphous and crystalline phases. It has been demonstrated that GeTe-based devices can be scaled down to 1.8 nm,[75] suggesting a high scaling limit of phase change-based artificial synapses.

However, phase change synapses also have several disadvantages. One is the inherent unipolar switching property due to their thermal effect-based switching mechanism, and the reset current and thus the power consumption is usually high.[68,76] Besides, it can be seen that compared with the gradual LTF process (Figure 6(c)), the LTD process involving melt and quench is inherently an abrupt process (Figure 6(d)), giving rise to low weight update linearity. A differential synaptic structure based on two PCM cells was thus proposed[77,78] to mitigate this issue, where the actual weight is represented by the conductance difference between the two PCM cells (Figure 6(e)). However, this method is largely affected by the mismatch within the differential pair of memristive devices, and refresh operations are needed to avoid conductance saturation. Recently, a synaptic structure based on multiple PCM cells with a counter-based

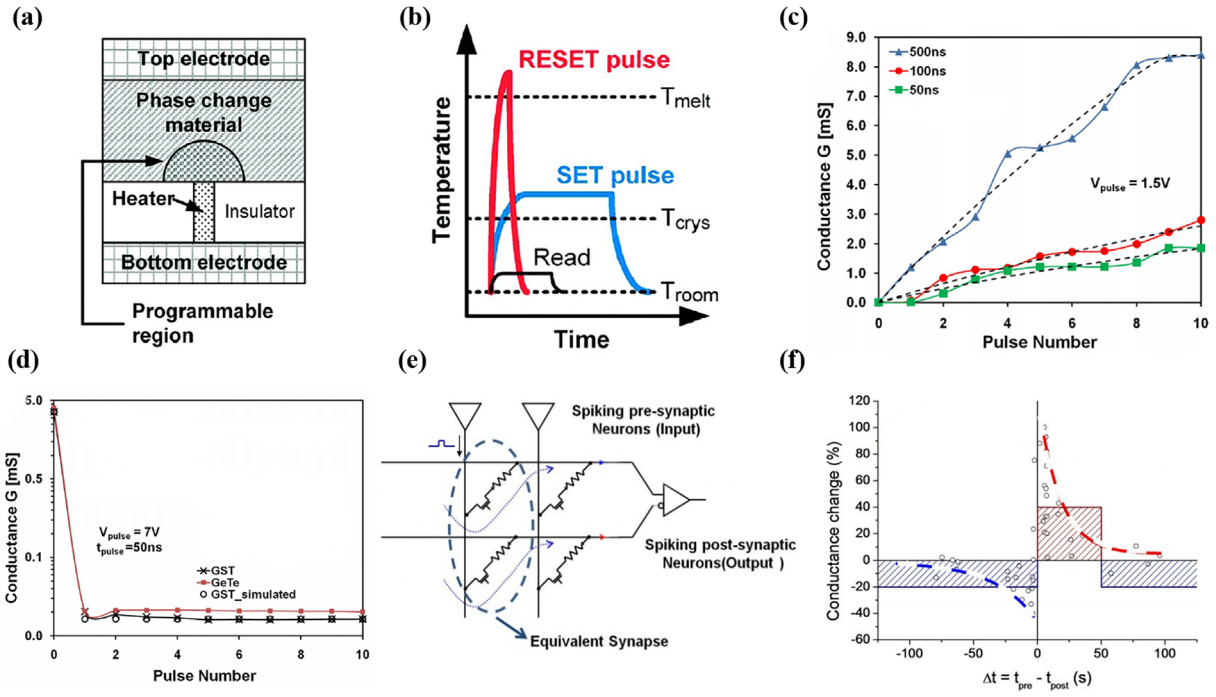

Figure 6. a) Typical "mushroom" structure of PCM cells, where the phase change material is sandwiched between the top electrode and the heater in contact with the bottom electrode. b) Typical programming and read conditions of PCM cells. Reproduced with permission.[67] Copyright 2010, IEEE. c) Experimental LTF characteristics of GeTe PCM cell. d) Experimental LTD characteristics in GST and GeTe PCM cells. e) Schematic of the differential synaptic structure based on two PCM cells. f) Biological STDP and simplified STDP achieved using PCM. In the simplified rule, a synapse receiving a post-synaptic spike with no pre-synaptic spike in the LTF window undergoes a LTD regardless of the existence of a pre-synaptic spike. Reproduced with permission.[77] Copyright 2011, IEEE.

arbitration architecture was proposed, where the synaptic weight was represented by the combined conductance of all the memristive devices in parallel, thus enhancing the overall weight precision.[79] Besides the analog STDP learning rule implemented,[71] a simplified binary STDP was also implemented in phase change synapses (Figure 6(f)), which can dramatically reduce the complexity in circuit design while still allow the system to learn sophisticated inference.[77]

2.4. Other Artificial Synapses

Besides the abovementioned artificial synapses, a variety of other approaches have been proposed to realize artificial synapses, such as spintronic devices,[80,81] carbon nanotubes,[82] ferroelectric memristors,[83] Mott memristors,[84] tunnel junction-based memristive devices,[85] and two-dimensional materials-based devices.[86,87] Electrical conductance of such devices can be modulated as a result of magnetic reversal, phase transition, movement of ferroelectric domain wall or ion transport etc., making them suitable for realizing the variable weight state of artificial synapses. It should be noted that in more complicated cases some memristive devices can be based on not only one mechanism, for example, memristive switching as a result of both ferroelectric polarization switching and oxygen migration have been reported.[88,89] In addition to long-term weight changes, the dynamic ionic transport and redox reactions occurred in memristors actually provide interesting possibilities to emulate more biologically plausible synaptic functionalities, such as STP[21] including both short-term facilitation (STF) and short-term depression (STD). For instance, Ohno et al.[21] reported short-term plasticity in Ag_2S-based synapses when applying electrical stimulations with low frequency, while application of high-frequency stimulation results in a transition from STP to LTP. Lu and co-workers also reported second order memristors with an additional state variable, namely the internal temperature, where the short-term temperature decay plays a role similar to Ca^{2+} level in biological synapses and thus provides a native mechanism to encode timing.[22] Different synaptic behaviors including rate dependent and spike timing-dependent plasticity can thus be implemented using such devices due to the emulation of synaptic dynamics.[22] More recently, a diffusive memristor was developed,[20] where the dynamics of Ag migration and diffusion faithfully emulate synaptic Ca^{2+} dynamics, leading to a direct and natural emulation of multiple synaptic functions, including both STP and LTP. Such short-term plasticity forms the basis for the construction of neural networks capable of processing complicated information such as spiking sequences, which will be discussed afterwards.

The above artificial synapses based on emerging devices including CBRAM, filamentary oxide memristors, non-filamentary oxide memristors and phase change devices have shown high scalability, while CBRAM based synapses have

demonstrated a scaling limit down to atomic level.[21,38] The weight update linearity and symmetry are challenging for all these device technologies, and this issue is particularly serious for PCM synapses compared with ion transport based synapses, since the depression process in PCM is inherently abrupt. The Joule heating-based switching mechanism imposes a limit in the energy consumption of PCM synapses. Despite a number of issues, the mass production for PCM is more mature compared with ionic memristors at this stage. Compared with CBRAM and filamentary oxide-based synapses, non-filamentary oxide synapses have generally displayed higher uniformity and weight precision, but they usually showed a smaller dynamic range and shorter retention, although encouraging progresses have been reported in some recent studies on ferroelectric tunnel memristors-based synapses.[88] In the meantime, it should be noted that there may exist trade-offs in the optimization of device performance. For instance, it might be possible to enlarge the dynamic range and improve retention by applying strong operating conditions, which however may deteriorate the endurance and result in higher power consumption.[90] As a result, a perfect synaptic device that combines all the favorable properties is yet to be demonstrated so far, which apparently demands continued device optimizations, and the device optimization should also be performed according to the requirements in detailed applications. It should be noted that biological neurons and synapses are similarly imperfect elements, but after being organized as neural networks with high parallelism they can perform complicated tasks with high robustness. This thus gives a hint to co-design of devices, circuits, and algorithms.

3. Memristive Neurons

Besides synapses, another fundamental element in biological neural networks is neuron. Every neuron in biological nervous systems is made of a cell body (soma), an axon, and dendrites.[3] The information encoding function of neurons is mainly realized by the change in their membrane potential. The resting equilibrium membrane potential is determined by the concentration gradients of ions across the membrane. The ion channels on the membrane are much more permeable to K^+ than to Na^+, so the resting potential is close to the equilibrium potential of K^+ that equals to about −70 mV. The action potential is generated by the flow of ions through voltage-gated channels, as shown in **Figure 7**. That is, the depolarization of the membrane as caused by external stimulations will change the configuration of the Na^+ channels, resulting in higher permeability to Na^+. As a result, there will be Na^+ influx through voltage-gated Na^+ channels, which forms the rising edge of the action potential and reverses the polarity of the plasma membrane. The voltage-gated Na^+ channels will then be closed and the K^+ channels are activated, hence leading to an outward flux of K^+ ions. This forms the dropping edge of the action potential and returns the membrane potential to the resting state.[3]

The above neuronal dynamics can be described by models with different complexity and accuracy, and existing studies on memristor-based neurons can be roughly divided into two types depending on the neuron model used in the implementation of

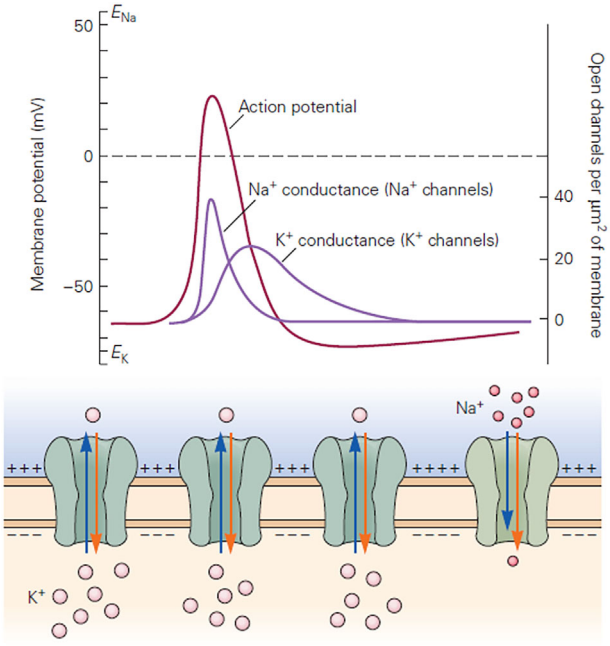

Figure 7. Generation of an action potential as a result of the activities of voltage-gated Na^+ and K^+ channels. Reproduced with permission.[3] Copyright 2013, The McGraw-Hill Companies, Inc.

devices, namely, leaky-integrate-and-fire (LIF) neuron and Hodgkin-Huxley (HH) neuron.[25,77] To date, a variety of methods have been adopted to implement LIF neurons based on memristive devices, which is popular due to the ease of circuit implementation and high scaling-up potential. For example, the amorphous to crystalline phase transition in a PCM cell can be used to physically emulate the integrate-and-fire dynamics of artificial neurons, provided that a reset process is applied to realize the decrease of the membrane potential, as shown in **Figure 8**(a).[25] Since the melt-quench induced reconfiguration in chalcogenide materials in inherently stochastic, it emulates the stochasticity of biological neurons and plays an important role in neuronal populations that represent and transmit sensory and motor signals.

Besides nonvolatile devices like the PCM cells, artificial neurons can also be constructed by connecting a volatile switch in parallel with a capacitor, where the volatile device can be a metal–insulator transition device[91,92] or a diffusive memristor,[93] as schematically depicted in Figure 8(b). The capacitor here is used for integration of charge, and the neuron fires once the voltage on the capacitor exceeds the threshold voltage of the volatile switch. The charge accumulated on the capacitor will thus be depleted and returns the neuron to its resting state automatically, and a reset process is thus unnecessary.

While LIF model captures basic functionalities of biological neurons, it is largely simplified and misses many detailed ion dynamics. The HH model is more biologically plausible and widely accepted in computational neuroscience.[94] Interestingly, a HH neuron capturing complex neural dynamics, including the all-or-nothing spiking, bifurcation threshold to a continuous spiking regime, signal gain and refractory period, can also be

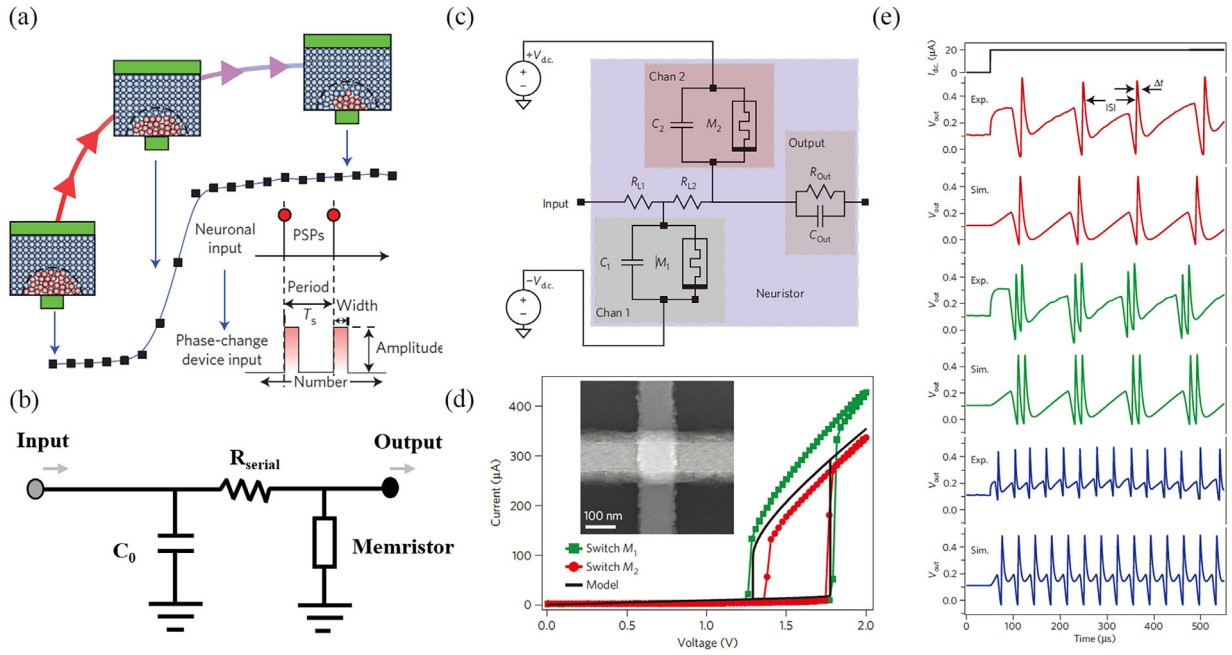

Figure 8. a) Schematic of a PCM neuron. Reproduced with permission.[25] Copyright 2016, Springer Nature Publishing AG. b) Schematic of a neuron consisting of a volatile switch and a capacitor in parallel. c–e) Hodgkin–Huxley neuron based on Mott memristors. The channels consist of Mott memristors with characteristic parallel capacitores. As the channel capacitances are adjusted, the neuron exhibits regular spiking, chattering, and fast spiking. Reproduced with permission.[84] Copyright 2013, Springer Nature Publishing AG.

achieved based on Mott memristors,[84] as shown in Figure 8(c–e), indicating the potential of building neuromorphic systems with rich neural and synaptic dynamics. One can see from Figure 8(c) that both Mott memristors have a capacitor in parallel and are powered by DC sources in opposite polarities, which are used to emulate the dynamics of the sodium and potassium ion channels, respectively. The HH neuron thus realizes more functionalities with the cost of additional device elements and chip areas.[94]

4. Memristive Crossbar Array

The artificial synapses and neurons discussed above provide building blocks for neuromorphic systems. A prerequisite to the construction of memristor-based hardware neural networks is the integration of memristive arrays, especially in a crossbar topology, which can maximize the integration density and computing parallelism.[95] **Figure 9**(a) exhibits a scanning electron microscopy (SEM) image of a 32×32 passive crossbar array fabricated by the authors, with a cell size <100 nm. Figure 9(b) further shows a memristive crossbar array vertically integrated on top of a CMOS chip,[96] showing the possibility of hybrid crossbar/CMOS systems, where the peripheral functionalities can be achieved by the CMOS circuits. An inherent issue to these passive crossbar arrays is the existence of sneak paths, which is a significant challenge in memory applications leading to possible read errors and increased energy consumption.[97] Although the sneak path issue is mitigated for neuromorphic applications especially during inference, since the output electrodes are grounded,[98] it may still increase the energy consumption during the training process and limit the array size. To address the sneak path issue, a select device connected in series with memristor is required, such as a transistor or a selector, which therefore forms a 1 transistor 1 resistor (1T1R) or 1 selector 1 resistor (1S1R) structure, or on-state nonlinearity needs to be introduced into the memristive device itself.[99–102] Figure 9(c) shows structure and die micrograph of a 2 Mb 1T1R memristor array using 28 nm technology.[103] However, the introduction of transistors increased the feature size compared with passive arrays. Figure 9(d) further shows an 8×8 Ni/TiO_2/Ni/HfO_2/Pt 1S1R memristor array, which can allow the same integration density with passive arrays.[104] Another advantage for 1T1R structure is that the design and fabrication for transistors is quite mature. Thus this structure is mostly adopted by researchers.[105–108] The challenge for the integration of 1S1R arrays lies in acquiring high-performance selectors, which have been extensively studied recently.[109–111] It should be noted all the passive crossbar, 1T1R and 1S1R arrays face some common issues, for example, the IR drop caused by long electrode wires in large-scale arrays,[112] requiring further considerations and optimizations at the array level.

5. Neural Network Accelerators Based on Memristors

Based on the neuromorphic devices and arrays shown above, neuromorphic computing systems can be further constructed. Since the VMM operation can be accelerated in memristor crossbar thanks to the parallel, in-memory, and analog characteristics of memristors, a large variety of neural networks

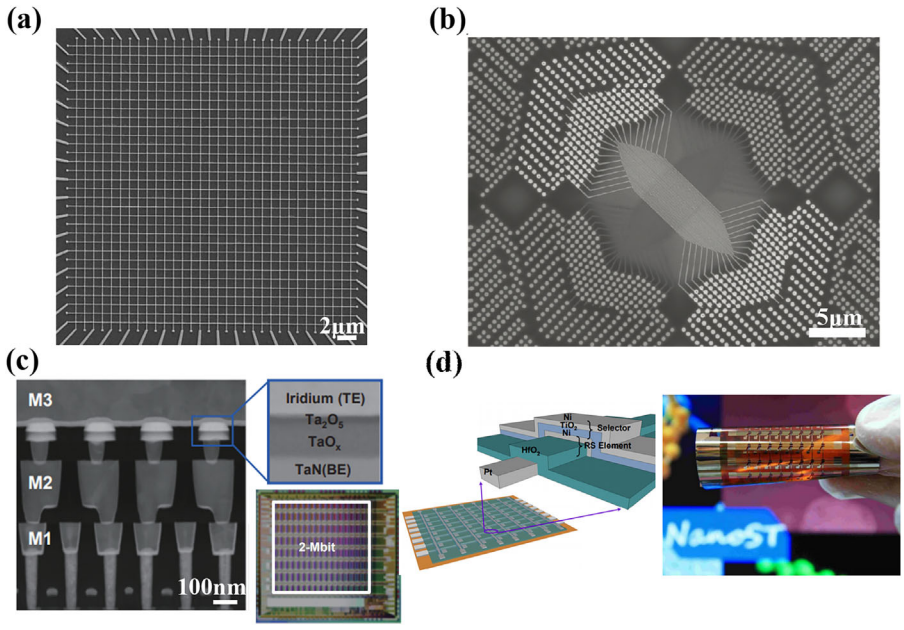

Figure 9. a) Scanning electron microscopy (SEM) image of 1 kbit passive memristor crossbar array. b) SEM image of a crossbar array fabricated on top of a CMOS chip. Scale bar: 5 μm. Reproduced with permission.[96] Copyright 2012, American Chemical Society. c) Structure and die micrograph of 2 Mbit 1T1R memristor array. Reproduced with permission.[103] Copyright 2015, IEEE. d) Cross-sectional schematic of 1S1R memristor array and photograph of 8 × 8 1S1R flexible array. Reproduced with permission.[104] Copyright 2011, IEEE.

can be accelerated,[93,113–116] and neuromorphic systems capable of encoding and processing spatiotemporal information might also be built by exploiting device dynamics such as short-term plasticity.[117,118] **Figure 10** summarizes typical network-level demonstrations to date that have been achieved experimentally, grouped according to the type of devices and algorithms used. The training methods in these studied are indicated in color, including both online and offline training in a supervised or unsupervised fashion. One can see that a majority of the present studies are focused on memristor-based ANN accelerators, with few studies on other types of neuromorphic systems. From a device perspective, ionic memristor (or RRAM) is considered as a highly promising device candidate and receives extensive research interests compared with other device technologies. Both passive memristor arrays (0T1R) and 1T1R arrays have been widely adopted in existing studies. Below we first discuss ANN accelerators based on memristors, and will cover other arithmetic accelerator and neuromorphic systems enabled by memristor dynamics in Sections 6 and 7, respectively.

5.1. Perceptron Networks

The first experimental ANN hardware based on memristor crossbar was reported in 2013,[115] where a simple pattern classification task was demonstrated using a single layer perceptron. As shown in **Figure 11**(a), the synaptic weights are mapped into a 2 × 10 titanium dioxide passive memristive crossbar. Since the synaptic weight can be negative, it is represented by the conductance difference between a pair of

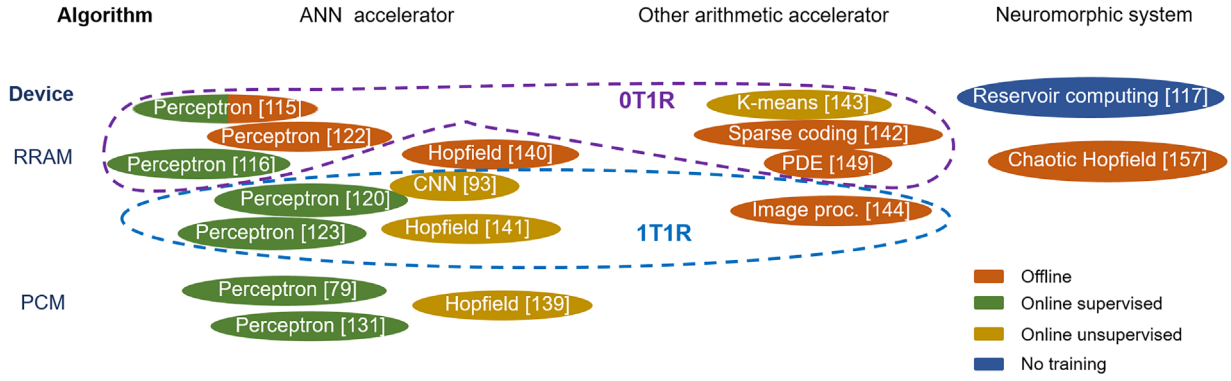

Figure 10. Overview of representative works on memristor-based neuromorphic computing systems.

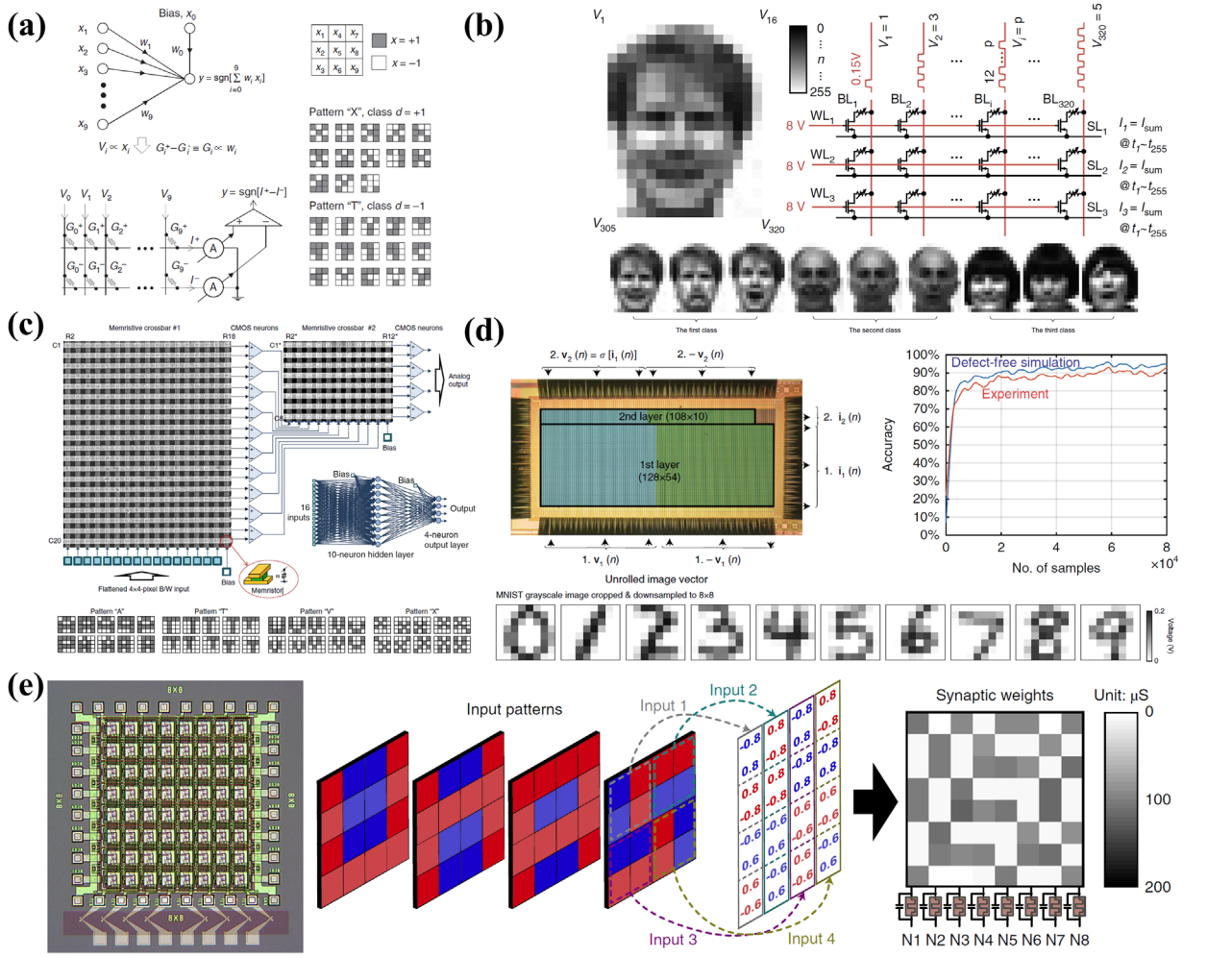

Figure 11. a) Single-layer perceptron based on memristor crossbar for letter classification. Reproduced with permission.[115] Copyright 2013, Springer Nature Publishing AG. b) Single-layer perceptron based on 1T1R memristor crossbar for face classification. Reproduced with permission.[120] Copyright 2017, Springer Nature Publishing AG. c) Multilayer perceptron classifier using two memristive crossbar arrays. Reproduced with permission.[122] Copyright 2018, Springer Nature Publishing AG. d) Multilayer perceptron network based on partitioned memristor array. The blue and green false-colored areas are the positive and negative parts of the differential pairs. Reproduced with permission.[123] Copyright 2018, Springer Nature Publishing AG. e) Fully integrated memristive neural network for pattern classification. Reproduced with permission.[93] Copyright 2018, Springer Nature Publishing AG.

devices in the neural network, and this differential approach has been widely adopted in matrix value mapping.[77,115] Both offline and online training in memristive crossbar-based hardware were demonstrated in this work. In a subsequent study by the same group, another perceptron network trained in situ with 2×30 Al_2O_3/TiO_2 memristors were used to solve a pattern classification problem,[116] where the Manhattan learning rule[119] was employed. The simplicity of the learning rule could be favorable for reducing the complexity of periphery circuit designs.

Obviously, the size of the crossbar array is important for the functionality and complexity of the network. A 1k 1T1R array was fabricated in ref. [120], with which a gray-scale face classification was demonstrated (Figure 11(b)). Compared with passive crossbar arrays, the 1T1R structure can improve the network performance in two aspects. Firstly, a proper current compliance determined by the transistor allows more controllable conductance update compared with that in 0T1R arrays. Secondly, the gate transistor can prevent disturbance to the states of unselected devices during programming. The 1T1R structure can also eliminate the sneak path problem of the passive crossbar,[121] although the sneak path is not troubling that much as in a memory array.[13,44] By comparing online training realized by both write-verify scheme and update scheme without verification, it was found that the seemingly more complicated write-verify scheme resulted in higher accuracy and required less training cycles, therefore leading to reduced energy consumption.[120]

However, it is well known single-layer perceptrons can only figure out linearly separable problems. In contrast, perceptron networks with hidden layers can be used to address this issue. A multi-layer perceptron network with one hidden layer integrated on board was reported recently, which comprised two separated

passive memristor crossbars corresponding to synaptic weights in different layers, as shown in Figure 11(c),[122] where the periphery circuits and neurons were implemented by discrete CMOS components. Besides such multiple crossbar method, the synaptic weights in different layers can also be realized by a large memristor array with proper partition.[123] As shown in Figure 11 (d), a two-layer perceptron can be constructed by partitioning the monolithic 128 × 64 1T1R array, which is subsequently trained online with standard MNIST dataset and achieves high classification accuracy of 91.71%,[123] showing the potential of online learning in tolerating device imperfections.

Although emerging memristive device can accelerate both the inference and training processes of artificial neural networks, some of the device non-idealities place essential constraints on network performance. These non-ideal characteristics include limited dynamic range, low weight precision, update nonlinearity, asymmetry as well as device variations, etc.[13,124–128] The insufficient dynamic range or weight precision limits the representation capability of the weight values, which are usually set to high precision in software, e.g., 32-bit floating point numbers in TensorFlow,[129] and hence the network performance. Update nonlinearity indicates that different conductance changes are obtained when identical set (reset) pulses are applied,[46] while symmetry is an important criterion for synaptic elements that refers to the relative rates of weight modulations during the facilitation and depression processes, where equal rates in facilitation and depression are defined as high symmetry.[51] Both nonlinearity and asymmetry will dramatically affect the online learning performance of ANN networks, otherwise a complicated periphery circuit design might be needed. Variation of memristive synapses mainly originates from the intrinsic stochasticity of memristors. While a low level of variations can be favorable for neural networks, it may also affect the network performance if device variations are serious. Further device optimizations are obviously needed to address the device variations, but it might be necessary to investigate on material-device-algorithm co-optimization as well in order to tackle these problems.

For instance, it is well known that PCM devices suffer from severe update asymmetry due to the abrupt conductance change in the reset process.[130] A multi-PCM synaptic architecture with a counter-based arbitration training scheme (**Figure 12**) was recently proposed to balance the mismatch in facilitation and depression.[79] In this multi-memristive synapse, *N* devices were connected in parallel to represent one weight which improved the weight precision of the synaptic element. Besides, an arbitration strategy was adopted when updating the synaptic weights, which contained three components – the global counter (Figure 12(a)) as well as independent facilitation and depression counters (Figure 12(b and c)).[79] The maximum value of the global counter is the number of devices *N* in the synapse unit. At any instance of synaptic update, only one of the *N* devices that was pointed to by the global counter is programmed. After that, the global counter is incremented by a fixed amount to ensure that the devices in one unit can get similar number of updates. In addition to the global counter, independent potentiation and

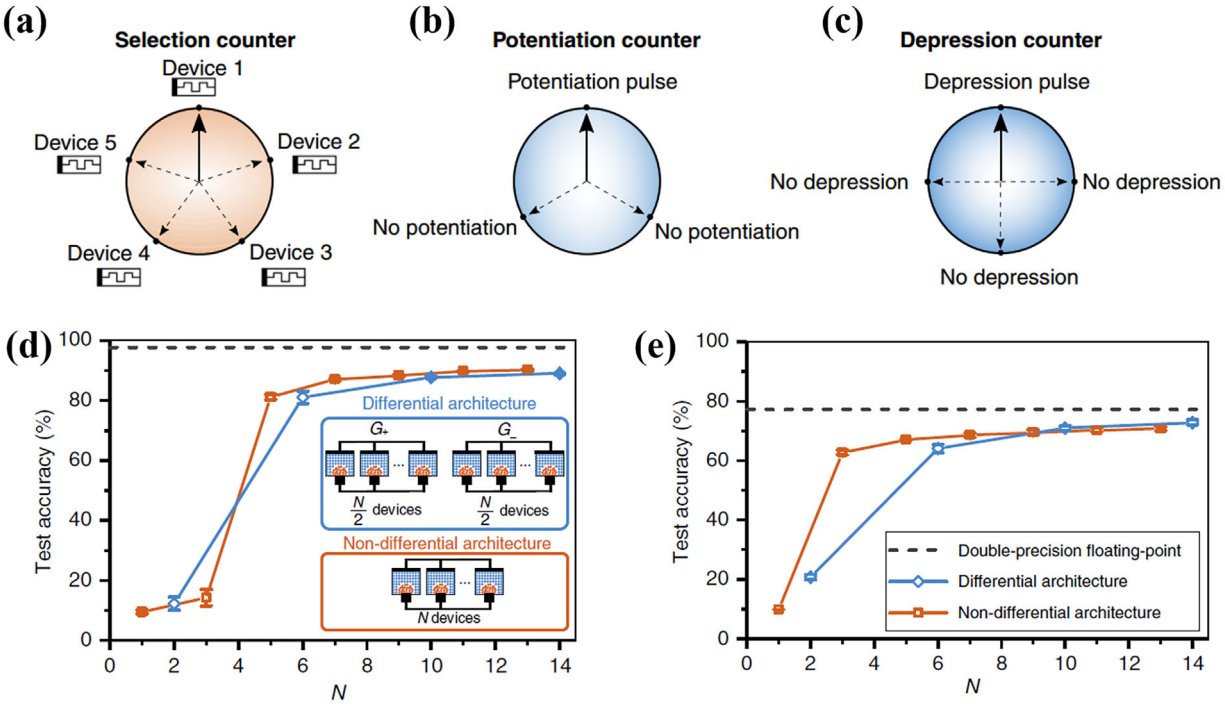

Figure 12. The arbitration scheme of the multi-PCM synapse. a) Schematic of the global counter, choosing which one device in the synapse to be updated during training. b) Schematic of the potentiation counter, which determines the frequency of the potentiation operations. c) Schematic of the depression counter, which controls the frequency of the depression operations. d) Performance of the multi-memristive synapses-based artificial neural network in handwritten digit classification. e) Performance of the multi-memristive synapses-based spiking neural network in handwritten digit classification. Reproduced with permission.[79] Copyright 2018, Springer Nature Publishing AG.

depression counters are also used to control the frequency of the potentiation or depression events in order to compensate device asymmetry in the programming process.[79] Such multi-PCM synaptic architecture can therefore tolerate the asymmetry in PCM, along with other non-ideal characteristics such as limited weight precision, as verified by simulations results from both artificial neural network trained by back propagation and a SNN trained by STDP (Figure 12(d and e)).

Another approach to addressing the weight update asymmetry and nonlinearity in synaptic elements is to combine the nonvolatile memory character of the memristive devices, which usually suffer from a certain level of asymmetry and nonlinearity, with the linear and symmetric updating of other devices, taking the PCM + CMOS architecture as an example.[131] In this case, the whole synaptic weight consists of two parts with different significances (**Figure 13**(a)). The part of lower significance is mapped to the effective conductance of a read transistor. A capacitor is connected to the gate electrode of the read transistor offering a proper gate voltage to determine the conductance of the transistor, which has good linearity. An extra PMOS is applied to add charge on the capacitor, whereas another NMOS is used to subtract charge, hence leading to bidirectional, symmetric and gradual updating of lower significance weight. However, the effective conductance of the transistor is volatile due to the leakage of the capacitor, requiring an extra nonvolatile part to hold the weight information. As a result, the higher significance weight is implemented by a pair of nonvolatile PCM cells. During the training process, only the transistor conductance is updated to achieve precise weight modulation. Once a preset number of examples have been trained, the weight temporarily stored in the read transistor will be transferred into the PCM pair. Based on this synapse architecture, the test accuracy reaches 97.94% for the MNIST dataset, which is already equivalent to software-based training results (Figure 13(b,c)),[131] implying the great prospect of memrister-based ANN accelerators for practical applications.

5.2. Convolutional Neural Networks

Memristor-based CNN accelerator has also generated extensive interest, due to the excellent performance of CNN in image recognition and many other applications. A CNN generally consists of multiple convolutional layers and fully connected layers for feature extractions and subsequent classifications, respectively.[5] In a convolution operation, the convolution kernel (usually a 2D matrix) is overlapped with the receptive field in the input image. Each pixel in the receptive field is multiplied by the corresponding value in the kernel, and the products are then summed up to form the pixel value in the corresponding feature map. This is once again weighted sum or VMM operations in essence and can be conveniently mapped onto memristor arrays, where the pixels of the receptive field serve as the inputs and each convolution kernel is stored in a column of the memristor array.[132] Such convolution operations represent a majority of calculation in CNN, and the above implementation in memristive arrays thus offers high efficiency. Aside from the convolution layers, a fully connected layer basically performs VMM, similar with a perceptron network, where the operations can also be directly implemented and accelerated in memristor crossbar arrays, as discussed in Section 4.1.

Besides the accelerated in-memory VMM operations realized by the memristive array, the whole architecture is worthy of a careful design in order to achieve good synergy between the analog and digital parts. ISSAC provides an efficient pipelined architecture with a new data encoding scheme to ensure the high throughout and reduce the cost of analog-to-digital converters (ADCs), digital-to-analog converters (DACs) as well as eDRAMs in the meantime.[133] PRIME includes configurable memory arrays with periphery circuits reused, which can be adopted as either network accelerators or memory based on requirements.[134] In addition, PRIME is equipped with a software/hardware interface to map different neural networks into the hardware. Aside from the accelerators that only target inference, memristor based architectures that support training have also been proposed, such as PipeLayer[135] and TIME.[136]

In a majority of existing studies, memristive devices mainly serve as synapses that hold the weight values and display plasticity during training, while the neurons are realized by CMOS components or software. Nevertheless, it should be noted that memristors have also been exploited to implement artificial neurons. Since one or a few memristive devices with (without) a capacitor can replace the CMOS neuron containing tens to hundreds of devices in the LIF neuron or HH neuron as

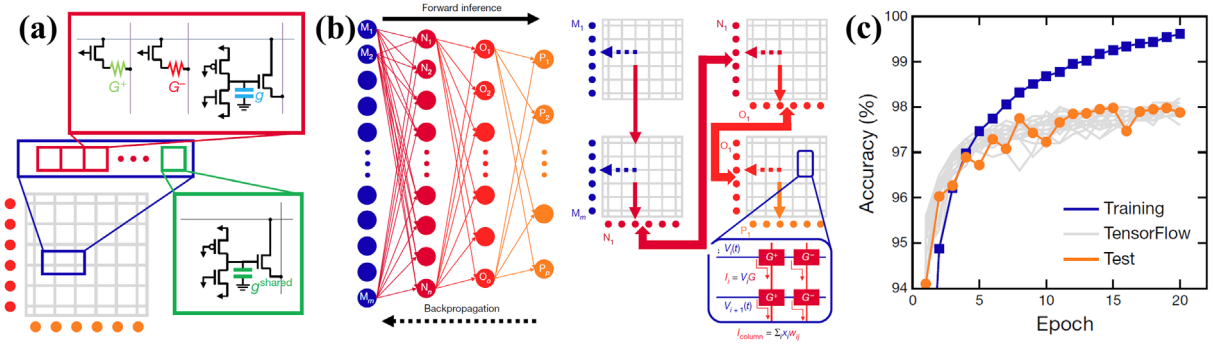

Figure 13. a) Schematic of a PCM+CMOS synapse unit cell. b) Schematic of mapping a fully connected neural network onto NVM arrays. c) Training (blue) and test (orange) accuracies for the mixed hardware–software experiment, which combines hardware-based PCM devices and SPICE-modeled 3T1C devices with full CMOS variability, on the MNIST dataset closely match those achieved for the same size network using TensorFlow (grey). Reproduced with permission.[131] Copyright 2018, Springer Nature Publishing AG.

discussed above, it opens up a new opportunity for fully memristor-based neural networks with high area and power efficiency. A fully memristive CNN was recently reported by Yang and co-workers,[93] where the leaky integrate-and-fire neurons were realized by volatile diffusive memristors and the synapse array was achieved by a 1T1R Pd/HfO_x/Ta memristor crossbar. Such network can be used to realize both the convolution layer and the fully connected layer, which in turn demonstrates a pattern classification task using unsupervised learning scheme,[93] as shown in Figure 11(e).

5.3. Recurrent Neural Networks

In addition to the above feedforward neural networks including perceptron and CNN, RNN is another important type of artificial neural networks and has shown good performance in processing temporal sequences, thus having potential applications in handwritten recognition and speech recognition thanks to its ability in maintaining internal states. Hopfield neural network is a typical class of RNN and has proven capable of performing content-addressable memories,[137] associative memories, and combinatorial optimization problems.[138]

The structure of a Hopfield network is a non-layered fully connected network. Unlike feedforward networks, the outputs in Hopfield networks are not used as inputs to the next layer but are processed as the inputs to the next cycle iteratively, thus falling into the category of recurrent neural networks.[139–141] Since the update of neuronal outputs in Hopfield networks also involves VMM, the synaptic connections can also be implemented using memristor crossbars, similar to the abovementioned networks. Different patterns can thus be remembered in memristive Hopfield networks by tuning the resistance states of the devices in the array via offline or online learning.[139–141]

Several Hopfield neural networks have been proposed based on memristors, for example, associative learning was achieved in a Hopfield network based on PCM array.[139] During the learning stage, the voltages applied on the wordlines and bitlines created a desirable overlap, which in turn programmed the synapse connecting two active neurons, following the Hebbian learning rule. Subsequently, the pre-trained pattern can be recovered in the recall stage, even if incomplete patterns are presented. Another memristive Hopfield network for associative memory was based on HfO_2 memristors,[140] where the weight matrix was programmed into the memristor crossbar with stepwise voltages in an offline fashion and both single and multiple patterns can be successfully retrieved in the recall stage. Besides, STDP learning rule can also be realized in HfO_2 devices in series with a transistor.[141] By implementing inhibitory synapses together with the excitatory ones, the network can have a competition mechanism between different patterns.

6. Other Arithmetic Accelerators

Although many recent studies have focused on ANN accelerators, the VMM operation is also heavily involved in many other algorithms, whose efficiency can also be significantly improved if taking advantage of the in-memory computing, parallel and analog characteristics of the memristor array, as shown in **Figure 14**. Similar with the application in ANN accelerators, herein the key purpose of using a memristor crossbar is to accelerate the heavily involved VMM as well as its deformation in these algorithms. Specifically, the matrix in the VMM is mapped element-wise to the conductance of devices in the memristor crossbar. Thereafter, voltages corresponding to the input vector are applied to the row or column of the memristor array. Based on the Ohm's law and Kirchhoff's current law, the VMM result can once again be obtained in a single read operation and used in subsequent calculations of the algorithms (see Figure 1). Therefore, in the following examples we first point out the mathematical operations related to VMM

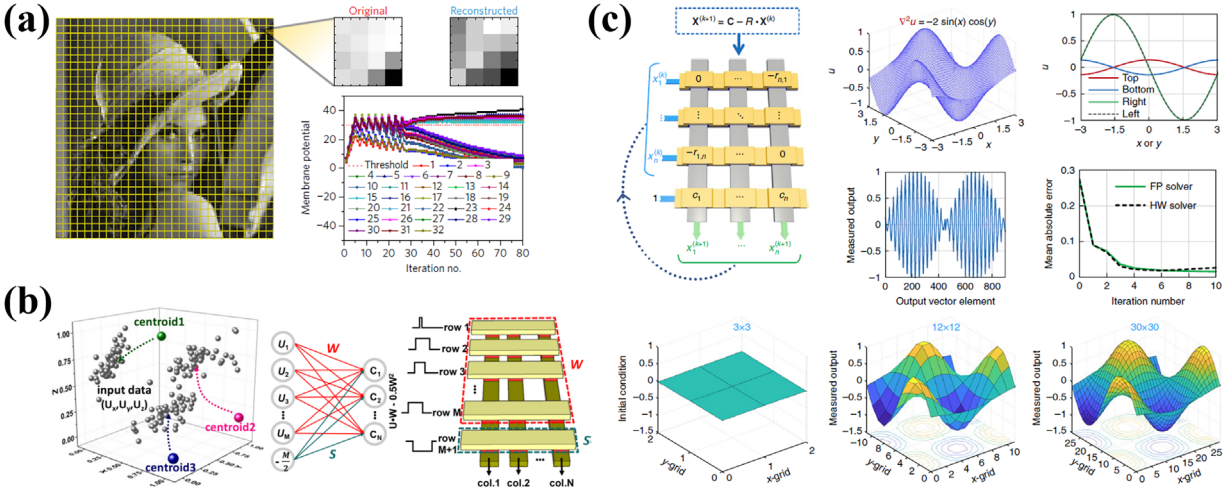

Figure 14. a) Patched processing of sparse coding using memristor crossbar array. Reproduced with permission.[142] Copyright 2017, Springer Nature Publishing AG. b) Experimental setup and mapping of K-means implementation using memristor crossbar array. Reproduced with permission.[143] Copyright 2018, American Chemical Society. c) Implementation of the Jacobi method using memristor crossbar array which can be used to iteratively solve Poisson's partial differential equation. Reproduced with permission.[149] Copyright 2018, Springer Nature Publishing AG.

that can be accelerated by the memristor array according to the mathematical form of the algorithms, followed by the detailed processes of the algorithm implementations in memristive hardware.

A representative example is sparse coding known as an efficient mechanism for encoding a vast number of sensory inputs in a sparse way in biological neural systems. From an application point of view, sparse coding offers a powerful means to perform feature extraction from high-dimensional data and has attracted extensive interest in computer vision, signal processing and object recognition, etc.[142] In addition to the general goal aiming at reducing the difference between the input information and its representation, another key idea in sparse coding is to implement lateral inhibition among neurons in order to realize the encoding sparsity. Locally competitive algorithm (LCA) is one of the specific approaches to accomplish sparse coding, where the neuron potential update equation is

$$\frac{du}{dt} = \frac{1}{\tau}(-u + x^{T}D - a(D^{T}D - I_{n}))$$
$$a_{i} = \begin{cases} u_{i} & \text{if } u_{i} > \lambda \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where x represents the input signal, u is the membrane potential changing over time t, D represents the feature dictionary, a is the coefficient describing the input signal using a linear combination of the features stored in the dictionary, τ is a time constant, I_{n} represents the $n \times n$ identity matrix, and λ refers to the threshold determining whether to update a. However, the inhibition term $a(D^{T}D - I_{n})$ containing matrix multiply matrix needs intensive computations and is hard to be directly mapped onto the memristive crossbar. Fortunately, with some simple mathematic transformation, the equation can be rewritten as

$$\frac{du}{dt} = \frac{1}{\tau}(-u + (x - \hat{x})^{T}D + a) \quad (2)$$

where $\hat{x} = Da^{T}$ is the sparse representation. Experimentally, by storing the dictionary into the crossbar, the expression can be realized by a memristor crossbar with two separate phases.[142] In the first phase, the input vector is applied to the crossbar rows to perform a "forward read" and the neuron potential can therefore be obtained to determine a. Whereas in the second phase, a is input through the column to calculate $\hat{x} = Da^{T}$ with a single "backward read" operation. After that, $x - \hat{x}$ is applied to the rows again to obtain the updated membrane potential. The above forward and backward processes are iterated until a stabilized output is obtained, and the final sparsely encoded data are acquired. Natural image compression has been demonstrated following this approach,[142] as shown in Figure 14(a). Other image processing algorithms such as discrete cosine transformation (DCT) and convolutional filtering have also been experimentally accomplished[144] using a 128×64 1T1R array.

Another typical example of arithmetic accelerators is the calculation of Euclidean distance based on memristive crossbar (Figure 14(b)). The calculation of Euclidean distance is important in many algorithms such as K-means data clustering, which is an unsupervised algorithm that can pre-cluster unlabeled data sets based upon the Euclidean distances between the input data and certain centroid locations before classifications are performed.[145] In the case of K-means, the Euclidean distance between a cluster center W_{n} and an input data U is mathematically defined by $\| U - W_{n} \|^{2} = U^{2} - 2U \cdot W_{n} + {W_{n}}^{2}$, once again including extensive VMM operations due to the term $U \cdot W_{n}$. By storing the extra term ${W_{n}}^{2}$ directly in an additional row, the Euclidean distance can be compared directly in hardware with high efficiency.[143]

Due to the limitations in device performance, memristive hardware is commonly considered more suitable for circumstances where the algorithm itself can tolerate device variations or relative lower precisions.[93,113,116,139,146–148] In stark contrast to ANN, there are still many tasks that involve high precision computing, such as numerical simulations. The finite scale of memristive crossbar places further constraint on practical applications. Despite the above issues at the device level, the potential of using memristor-based hardware in high-precision computing tasks is demonstrated by a memristor-based partial differential equation (PDE) solver,[149] as shown in Figure 14(c). Solving PDEs is of generalized significance in simulation, prediction and optimization problems.[150–153] To solve the PDEs, the zero elements in the very sparse coefficient matrix are ignored and the remaining matrix is then divided into equally sized submatrix that matches with the capacity of the crossbar. Furthermore, a precision expansion strategy is adopted to meet the precision requirement, where multiple crossbars are needed with each one only storing part of the bits. The partial products from such separate crossbars can be summed later to get the final VMM result. The validity of this approach was verified by experimentally solving both static Poisson's equation and time-evolving wave equation, thus showing that the application of memristors-based hardware can be extended to high precision and accurate computation tasks.

7. Neuromorphic Systems Enabled by Memristor Dynamics

It should be pointed out that the above implementations of memristor-based ANN and arithmetic accelerators mainly rely on the long-term storage of analog weight states in the memristor array. However, that is not the way how information is processed in biological neural systems. Instead, the brain is found to be an intricate, noisy and dynamic system having real-time interaction with environments and can process complex spatiotemporal information. While neurons work near the edge of chaos,[154] the synaptic plasticity in biological systems is also much more complex compared with the simply long-term plasticity as used in ANN and arithmetic accelerators. There have been clear evidence in neuroscience showing that other forms of synaptic behaviors, such as short-term plasticity[26] and heterosynaptic plasticity,[155,156] also play significant roles in biological systems. Recently, there is a growing interest in building neuromorphic systems with more complex information processing capabilities, by exploiting the intrinsic dynamics of memristors.[117,118,157]

7.1. Reservoir Computing

Reservoir computing is considered to be an important framework inspired by biological neural systems. A reservoir computing system essentially consists of two parts, a reservoir and a readout module. The reservoir maps input data nonlinearly to a higher dimensional space, while the readout module subsequently extracts the state of the reservoir and obtains the final output result. In contrast to CNN containing numerous parameters in both convolutional and fully connected layers, whose training requires big data, the reservoir needs no training and thus attracts a lot of interest recently.

Figure 15(a) exhibits a hardware reservoir system based on memristive devices with short-term memory.[117] Since the states of the devices are not only dependent on the presence or absence of electric pulses but also determined by the timing of the pulses arriving, the different temporal inputs can be translated into different device states due to the STP behavior, which can in turn be read out as an abstract feature for further processing. A single device can therefore effectively serve as a reservoir, and both digit recognition and second-order nonlinear tasks have been successfully solved using such reservoir networks.[117]

Besides nanoionic memristors, a reservoir computing system has also been implemented by a magnetic tunnel junction oscillator,[118] which can similarly endow the computing system with nonlinearity and short-term memory. In this case, the amplitude of the spin-torque oscillator responds nonlinearly to the input current and is related to the input history, which can thus be utilized as the reservoir. Application of such reservoir system in spoken-digit recognition has been successfully demonstrated, as shown in Figure 15(b).

7.2. Memristors as Chaotic Sources

The rich dynamics in memristive devices can also be exploited to realize a chaotic source, which is deemed important for solving combinatorial or global optimization problems. In a recent study by Williams and co-workers,[157] the two state variables of NbO_2 based Mott memristors, i.e., the internal device temperature and

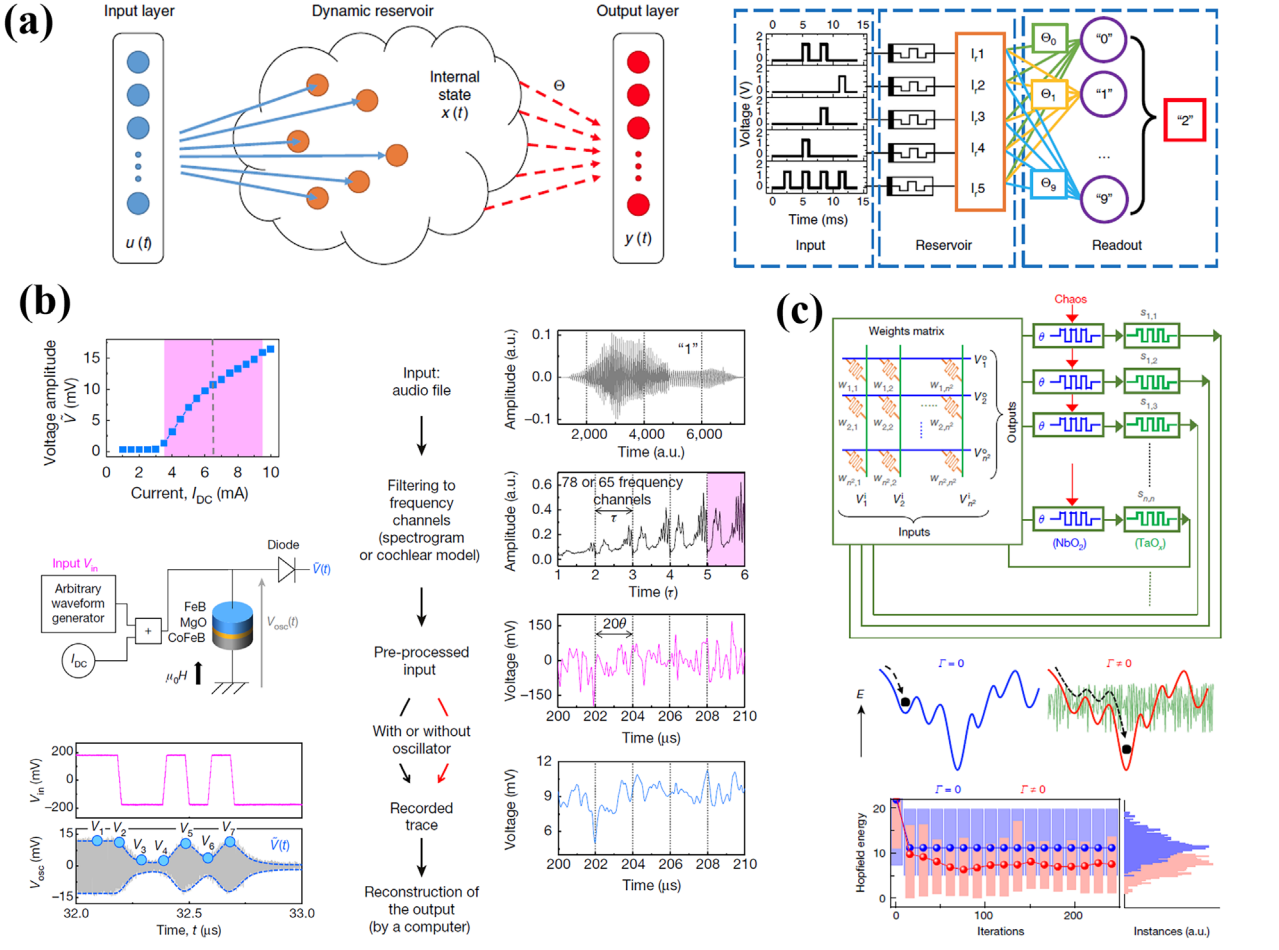

Figure 15. a) Reservoir computing network based on memristive devices with short-term plasticity. Reproduced with permission.[117] Copyright 2017, Springer Nature Publishing AG. b) Reservoir computing system implemented by a magnetic tunnel junction oscillator. Reproduced with permission.[118] Copyright 2017, Springer Nature Publishing AG. c) Implementation of Hopfield network using Mott memristors, where the chaotic dynamics can help the Hopfield network jump out of local minima and converge to a global minimal state. Reproduced with permission.[157] Copyright 2017, Springer Nature Publishing AG.

the charge on its intrinsic capacitance, led to a relaxation oscillator. When coupled with thermal oscillations, the coupled oscillator naturally exhibited a chaotic behavior. It was demonstrated that introduction of such chaotic dynamics into a Hopfield network can prevent the network from falling into local minima and lead to convergence to a global minimization state (Figure 15(c)).[157]

8. Conclusion and Outlook

Here we have given an overview on artificial synapses and neurons based on memristive devices and evaluated the respective advantages and disadvantages of these approaches. Neural network and arithmetic accelerators based on memristive crossbars are also discussed, along with neuromorphic computing systems that exploit the intrinsic dynamics of memristors. Despite the encouraging progresses so far, memristor-based neuromorphic system is still in its infancy, and there are a number of outstanding challenges that have yet been resolved in the present studies. First, an ideal synaptic element that combines high scalability, high weight precision, large dynamic range, high weight update linearity/symmetry and low energy consumption is still missing, which are highly desirable for the inference and especially online training of many neural networks. In order to address this issue, further device engineering is obviously needed, but co-optimization studies between devices, algorithms and circuits might also be necessary. Preliminary efforts along this direction have shown encouraging results in achieving generally equivalent performance with software based on non-ideal devices in certain tasks. Considerable effort should be devoted to balance and benchmark the network performance, chip area, power consumption as well as manufacturing complexity of the computing system, and such system-level evaluations will require electronic design automation (EDA) tools. Furthermore, it may be imperative to develop new algorithms or neural networks that can take advantage of the unique properties of memristors, in particular those are more akin to the operation principles of the brain. It should also be noted that the studies on artificial neurons are still very limited so far, and further investigations are certainly needed to obtain a spiking neuron element with high scalability, low energy consumption and biologically plausible functionalities for the construction of bio-inspired computing systems. Memristor-based neuromorphic hardware has been a rapidly developing area in the last decade. We expect this subject is going to receive increasing attention from the academia and industry and make dramatic advancements in the near future, potentially leading to good complements to exiting von Neumann computers and finding extensive applications in big data and artificial intelligence. This, nevertheless, requires close collaborations between scientists that cross a wide disciplines including neuroscience, electrical engineering, computer science, materials science, and mathematics.

Acknowledgments

T.Z. and K.Y. contributed equally to this work. This work was supported by National Key R&D Program of China (2017YFA0207600), National Natural Science Foundation of China (61674006, 61421005), and the 111 Project (B18001).

Conflict of Interest

The authors declare no conflict of interest.

Keywords

artificial neurons, artificial synapses, intelligence, memristors, neuromorphic computing

Received: January 14, 2019
Revised: February 22, 2019
Published online:

[1] S. A. McKee, *Proc. 1st Conf. on Computing Frontiers (CF'04)*, ACM, **2004**, p. 162.
[2] L. A. Hart, *How the Brain Works: New Understanding of Human Learning, Emotion, and Thinking*. Basic Books, New York, USA **1975**.
[3] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, A. J. Hudspeth, *Principles of Neural Science*, 5th ed., McGraw-Hill, New York, USA **2013**.
[4] C. A. Mead, *Proc. IEEE* **1990**, *78*, 1629.
[5] A. Krizhevsky, I. Sutskever, G. E. Hinton, *Advances in Neural Information Processing Systems*, Vol. 25, Curran Associates, Red Hook **2012**, p. 1097.
[6] A. Graves, A.-R. Mohamed, G. Hinton, *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* **2013**, p. 6645.
[7] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[8] K. Simonyan, A. Zisserman, *arXiv preprint*, arXiv:1409.1556, **2014**.
[9] L. Chua, *IEEE Trans. Circuit Theory* **1971**, *18*, 507.
[10] L. Chua, *Semicond. Sci. Technol*. **2014**, *29*, 104001.
[11] D. B. Strukov, G. S. Snider, D. R. Stewart, R. S. Williams, *Nature* **2008**, *453*, 83.
[12] R. Waser, R. Dittmann, G. Staikov, K. Szot, *Adv. Mater.* **2009**, *21*, 2632.
[13] J. J. Yang, D. B. Strukov, D. R. Stewart, *Nat. Nanotechnol*. **2013**, *8*, 13.
[14] I. Valov, E. Linn, S. Tappertzhofen, S. Schmelzer, J. van den Hurk, F. Lentz, R. Waser, *Nat. Commun*. **2013**, *4*, 1771.
[15] Y. Yang, R. Huang, *Nat. Electron*. **2018**, *1*, 274.
[16] I. Valov, I. Sapezanskaia, A. Nayak, T. Tsuruoka, T. Bredow, T. Hasegawa, G. Staikov, M. Aono, R. Waser, *Nat. Mater.* **2012**, *11*, 530.
[17] Y. Yang, P. Gao, S. Gaba, T. Chang, X. Pan, W. D. Lu, *Nat. Commun*. **2012**, *3*, 732.
[18] Y. Yang, P. Gao, L. Li, X. Pan, S. Tappertzhofen, S. Choi, R. Waser, I. Valov, W. D. Lu, *Nat. Commun*. **2014**, *5*, 4232.
[19] Y. Yang, X. Zhang, L. Qin, Q. Zeng, X. Qiu, R. Huang, *Nat. Commun*. **2017**, *8*, 15173.
[20] Z. Wang, S. Joshi, S. E. Savel'ev, H. Jiang, R. Midya, P. Lin, M. Hu, N. Ge, J. P. Strachan, Z. Li, Q. Wu, M. Barnell, G.-L. Li, H. L. Xin, R. S. Williams, Q. Xia, J. J. Yang, *Nat. Mater.* **2017**, *16*, 101.
[21] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. K. Gimzewski, M. Aono, *Nat. Mater.* **2011**, *10*, 591.
[22] S. Kim, C. Du, P. Sheridan, W. Ma, S. Choi, W. D. Lu, *Nano. Lett.* **2015**, *15*, 2203.
[23] Y. Yang, M. Yin, Z. Yu, Z. Wang, T. Zhang, Y. Cai, W. D. Lu, R. Huang, *Adv. Electron. Mater.* **2017**, *3*, 1700032.
[24] Y. Yang, B. Chen, W. D. Lu, *Adv. Mater.* **2015**, *27*, 7720.
[25] T. Tuma, A. Pantazi, M. L. Gallo, A. Sebastian, E. Eleftheriou, *Nat. Nanotechnol*. **2016**, *11*, 693.

[26] R. S. Zucker, W. G. Regehr, *Annu. Rev. Physiol.* **2002**, *64*, 355.
[27] S. J. Martin, P. D. Grimwood, R. G. M. Morris, *Annu. Rev. Neurosci.* **2000**, *23*, 649.
[28] H. L. Atwood, J. M. Wojtowicz, *Int. Rev. Neurobiol.* **1986**, *28*, 275.
[29] S. Li, F. Zeng, C. Chen, H. Liu, G. Tang, S. Gao, C. Song, Y. Lin, F. Pan, D. Guo, *J. Mater. Chen. C* **2013**, *34*, 5292.
[30] M.-S. Lee, J.-W. Lee, C.-H. Kim, B.-G. Park, J.-H. Lee, *IEEE Trans. Electron Devices* **2015**, *62*, 569.
[31] S. Mitra, S. Fusi, G. Indiveri, *IEEE International Symposium on Circuits and Systems (ISCAS)* **2006**, p. 4.
[32] W. Xu, S.-Y. Min, H. Hwang, T.-W. Lee, *Sci. Adv.* **2016**, *2*, e1501326.
[33] R. Islam, H. Li, P.-Y. Chen, W. Wan, H.-Y. Chen, B. Gao, H. Wu, S. Yu, K. Saraswat, H.-S. P. Wong, *J. Phys. D: Appl. Phys.* **2019**, *52*, 113001.
[34] Y. Cai, Y. Lin, L. Xia, X. Chen, S. Han, Y. Wang, H. Yang, *Proc. 55th Annual Design Automation Conference*, ACM, **2018**, p. 107.
[35] I. Valov, M. N. Kozicki, *J. Phys. D* **2013**, *46*, 074005.
[36] M. N. Kozicki, M. Park, M. Mitkova, *IEEE Trans. Nanotechnol.* **2005**, *4*, 331.
[37] M. Kund, G. Beitel, C.-U. Pinnow, T. Rohr, J. Schumann, R. Symanczyk, K. Ufert, G. Muller, *IEEE Int. Electron Devices Meet.* **2005**, p. 754.
[38] T. Hasegawa, T. Ohno, K. Terabe, T. Tsuruoka, T. Nakayama, J. K. Gimzewski, M. Aono, *Adv. Mater.* **2010**, *22*, 1831.
[39] A. Calderoni, S. Sills, C. Cardon, E. Faraoni, N. Ramaswamy, *Microelectron. Eng.* **2015**, *147*, 145.
[40] A. Bricalli, E. Ambrosi, M. Laudato, M. Maestro, R. Rodriguez, D. Ielmini, *IEEE Trans. Electron Devices* **2018**, *65*, 115.
[41] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, W. D. Lu, *Nano. Lett.* **2010**, *10*, 1297.
[42] Y. Shi, L. Nguyen, S. Oh, X. Liu, F. Koushan, J. R. Jameson, D. Kuzum, *Nat. Commun.* **2018**, *9*, 5312.
[43] M. Suri, V. Parmar, *IEEE Trans. Nanotechnol.* **2015**, *14*, 963.
[44] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. T. Chen, M.-J. Tsai, *Proc. IEEE* **2012**, *100*, 1951.
[45] M. Lanza, K. Zhang, M. Porti, M. Nafría, Z. Y. Shen, L. F. Liu, J. F. Kang, D. Gilmer, G. Bersuker, *Appl. Phys. Lett.* **2012**, *100*, 123508.
[46] Z. Wang, M. Yin, T. Zhang, Y. Cai, Y. Wang, Y. Yang, R. Huang, *Nanoscale* **2016**, *8*, 14015.
[47] Y. Wu, S. Yu, B. Lee, P. Wong, *J Appl. Phys.* **2011**, *110*, 094104.
[48] W.-T. Wu, J.-J. Wu, J.-S. Chen, *ACS Appl. Mater. Interfaces* **2011**, *3*, 2616.
[49] H. Y. Jeong, Y. I. Kim, J. Y. Lee, S. Y. Choi, *Nanotechnology* **2010**, *21*, 115203.
[50] Z. Fang, H. Y. Yu, X. Li, N. Singh, G. Q. Lo, D. L. Kwong, *IEEE Electron Device Lett.* **2011**, *32*, 566.
[51] P.-Y. Chen, X. Peng, S. Yu, *IEEE Trans. Comput-Aided Des. Integr. Circuits Syst.* **2018**, *37*, 3067.
[52] Y. Jeong, S. Kim, W. D. Lu, *Appl. Phys. Lett.* **2015**, *107*, 173105.
[53] J. Li, Q. Duan, T. Zhang, M. Yin, X. Sun, Y. Cai, L. Li, Y. Yang, R. Huang, *Rsc Adv.* **2017**, *7*, 43132.
[54] P.-Y. Chen, B. Lin, I.-T. Wang, T.-H. Hou, J. Ye, S. Vrudhula, J.-S. Seo, Y. Cao, S. Yu, *IEEE/ACM Int. Conf. Comput.-Aided Des.* **2015**, 194.
[55] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, H.-S. P. Wong, *Front. Neurosci.* **2013**, *7*, 186.
[56] A. Herpers, C. Lenser, C. Park, F. Offi, F. Borgatti, G. Panaccione, S. Menzel, R. Waser, R. Dittmann, *Adv. Mater.* **2014**, *26*, 2730.
[57] A. B. K. Chen, S. G. Kim, Y. Wang, W.-S. Tung, I.-W. Chen, *Nat. Nanotech.* **2011**, *6*, 237.
[58] X. L. Shao, L W. Zhou, K. J. Yoon, H. Jiang, J. S. Zhao, K. L. Zhang, S. Yoob, C. S. Hwang, *Nanoscale* **2015**, *7*, 11063.
[59] R. Pan, J. Li, F. Zhuge, L. Zhu, L. Liang, H. Zhang, J. Gao, H. Cao, B. Fu, K. Li, *Appl. Phys. Lett.* **2016**, *108*, 013504.
[60] J. Wang, R. Pan, H. Cao, Y. Wang, L. Liang, H. Zhang, J. Gao, F. Zhuge, *Appl. Phys. Lett.* **2016**, *109*, 143505.
[61] S. Park, A. Sheri, J. Kim, J. Noh, J. Jang, M. Jeon, B. Lee, B. R. Lee, B. H. Lee, H. Hwang, *IEEE Int. Electron Devices Meet.* **2013**, p. 25.6.
[62] S. Park, H. Kim, M. Choo, J. Noh, A. Sheri, S. Jung, K. Seo, J. Park, S. Kim, W. Lee, J. Shin, D. Lee, G. Choi, J. Woo, E. Cha, J. Jang, C. Park, M. Jeon, B. Lee, B. H. Lee, H. Hwang, *IEEE Int. Electron Devices Meet.* **2012**, p. 10.2.
[63] M. Yin, Y. Yang, Z. Wang, T. Zhang, Y. Fang, X. Yang, Y. Cai, R. Huang, *ICSICT* **2016**, p. 1113.
[64] K. Moon, S. Lim, J. Park, C. Sung, S. Oh, J. Woo, J. Lee, H. Hwang, *Faraday Discuss.* **2018**, *213*, 421.
[65] S. Park, M. Siddik, J. Noh, D. Lee, K. moon, J. Woo, B. H. Lee, H. Hwang, *Semicond. Sci. Technol.* **2014**, *29*, 104006.
[66] A. M. Sheri, H. Hwang, M. Jeon, B.-G. Lee, *IEEE Trans. Ind. Electron.* **2014**, *61*, 2933.
[67] H.-S. P. Wong, S. Raoux, S. B. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, K. E. Goodson, *Proc. IEEE* **2010**, *98*, 2201.
[68] D. J. Wouters, R. Waser, M. Wuttig, *Proc. IEEE* **2015**, *103*, 1274.
[69] M. Wuttig, N. Yamada, *Nat. Mater.* **2007**, *6*, 824.
[70] D. Loke, L. Shi, W. Wang, R. Zhao, H. Yang, L.-T. Ng, K.-G. Lim, T.-C. Chong, Y.-C. Yeo, *Nanotechnology* **2011**, *22*, 254019.
[71] D. Kuzum, R. G. D. Jeyasingh, B. Lee, H.-S. P. Wong, *Nano. Lett.* **2012**, *12*, 2179.
[72] S. Raoux, C. T. Rettner, *J. Appl. Phys.* **2007**, *102*, 094305.
[73] W. J. Wang, L. P. Shi, R. Zhao, K. G. Lim, H. K. Lee, T. C. Chong, Y. H. Wu, *Appl. Phys. Lett.* **2008**, *93*, 043121.
[74] F. Rao, K. Ding, Y. Zhou, Y. Zheng, M. Xia, S. Lv, Z. Song, S. Feng, I. Ronneberger, R. Mazzarello, W. Zhang, E. Ma, *Science* **2017**, *358*, 1423.
[75] M. A. Caldwell, S. Raoux, R. Y. Wang, H.-S. P. Wong, D. J. Milliron, *J. Mater. Chem.* **2010**, *20*, 1285.
[76] S. Raoux, F. Xiong, M. Wuttig, E. Pop, *MRS Bull.* **2014**, *39*, 703.
[77] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, B. DeSalvo, *IEEE Int. Electron Devices Meet.* **2011**, p. 4.4.
[78] O. Bichler, M. Suri, D. Querlioz, D. Vuillaume, B. DeSalvo, C. Gamrat, *IEEE Trans. Electron Devices* **2012**, *59*, 2206.
[79] I. Boybat, M. L. Gallo, S. R. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian, E. Eleftheriou, *Nat. Commun.* **2018**, *9*, 2514.
[80] M. Sharad, C. Augustine, K. Roy, *IEEE Int. Electron Devices Meet.* **2012**, 11.6.
[81] A. Senguptam, M. Parsa, B. Han, K. Roy, *IEEE Trans. Electron Devices* **2016**, *63*, 2963.
[82] K. Gacem, J.-M. Retrouvey, D. Chabi, A. Filoramo, W. Zhao, J.-O. Klein, V. Derycke, *Nanotechnology* **2013**, *24*, 384013.
[83] Y. Kaneko, Y. Nishitani, M. Ueda, A. Tsujimura, *VLSI* **2013**, p. 238.
[84] M. D. Pickett, G. Medeiros-Ribeiro, R. S. Williams, *Nat. Mater.* **2013**, *12*, 114.
[85] A. Thomas, S. Niehörster, S. Fabretti, N. Shepheard, O. Kuschel, K. Küpper, J. Wollschläger, P. Krzysteczko, E. Chicca, *Front. Neurosci.* **2015**, *9*, 241.
[86] X. Zhu, D. Li, X. Liang, W. D. Lu, *Nat. Mater.* **2019**, *18*, 141.
[87] J. Zhu, Y. Yang, R. Jia, Z. Liang, W. Zhu, Z. U. Rehman, L. Bao, X. Zhang, Y. Cai, L. Song, R. Huang, *Adv. Mater.* **2018**, *30*, 1800195.
[88] R. Guo, Y. Zhou, L. Wu, Z. Wang, Z. Lim, X. Yan, W. Lin, H. Wang, H. Y. Yoong, S. Chen, Ariando, T. Venkatesan, J. Wang, G. M. Chow, A. Gruverman, X. Miao, Y. Zhu, J. Chen, *ACS Appl. Mater. Interfaces* **2018**, *10*, 12862.
[89] W. J. Hu, Z. Wang, W. Yu, T. Wu, *Nat. Commun.* **2016**, *7*, 10808.
[90] G. Molasa, G. Sassinea, C. Naila, D. A. Robayoa, J.-F. Nodina, C. Caglia, J. Coignusa, P. Blaisea, E. Nowaka, *ECS Trans.* **2018**, *86*, 35.
[91] L. Gao, P.-Y. Chen, S. Yu, *Appl. Phys. Lett.* **2017**, *111*, 103503.

[92] J. Lin, Annadi, S. Sonde, C. Chen, L. Stan, K. V. L. V. Achari, S. Ramanathan, S. Guha, *IEEE Int. Electron Devices Meet*. **2016**, *34*, 5.
[93] Z. Wang, S. Joshi, S. Savel'ev, W. Song, R. Midya, Y. Li, M. Rao, P. Yan, S. Asapu, Y. Zhuo, H. Jiang, P. Lin, C. Li, J. H. Yoon, N. K. Upadhyay, J. Zhang, M. Hu, J. P. Strachan, M. Barnell, Q. Wu, H. Wu, R. S. Williams, Q. Xia, J. J. Yang, *Nat. Electron*. **2018**, *1*, 137.
[94] E. M. Izhikevich, *IEEE Trans. Neural. Netw.* **2004**, *15*, 1063.
[95] A. Chen, *J. Comput. Electron* **2017**, *16*, 1186.
[96] K.-H. Kim, S. Gaba, D. Wheeler, J. M. C. Albrecht, T. Hussain, N. Srinivasa, W. Lu, *Nano. Lett*. **2012**, *12*, 389.
[97] J. Zhou, K.-H. Kim, W. Lu, *IEEE Trans. Electron Devices* **2014**, *61*, 1369.
[98] Y. Jeong, M. A. Zidan, W. D. Lu, *IEEE Trans. Nanotechnol*. **2018**, *17*, 184.
[99] G. W. Burr, R. S. Shenoy, K. Virwani, P. Narayanan, A. Padilla, B. Kurdi, *J. Vac. Sci. Technol. B: Nanotechnol. Microelectron.: Mater. Process. Meas. Phenom*. **2014**, *32*, 040802.
[100] J. Zhou, F. Cai, Q. Wang, B. Chen, S. Gaba, W. D. Lu, *IEEE Electron Device Lett*. **2018**, *17*, 184.
[101] F. Nardi, S. Balatti, S. Larentis, D. Ielmini, *IEEE Int. Electron Devices Meet.* **2011**, p. 31.
[102] E. Linn, R. Rosezin, C. Kügeler, R. Waser, *Nat. Mater.* **2010**, *9*, 403.
[103] Y. Hayakawa, A. Himeno, R. Yasuhara, W. Boullart, E. Vecchio, T. Vandeweyer, T. Witters, D. Crotti, M. Jurczak, S. Fujii, S. Ito, Y. Kawashima, Y. Ikeda, A. Kawahara, K. Kawai, Z. Wei, S. Muraoka, K. Shimakawa, T. Mikawa, S. Yoneda, *Symp. VLSI Technol.* **2015**.
[104] J.-J. Huang, Y.-M. Tseng, W.-C. Luo, C.-W. Hsu, T.-H. Hou, *IEEE Int. Electron Devices Meet.* **2011**, p. 31.
[105] X. Huang, H. Wu, D. C. Sekar, S. N. Nguyen, K. Wang, H. Qian, *IEEE Int. Mem. Worksh.* **2015**, p. 1.
[106] M. Ueki, K. Takeuchi, T. Yamamoto, A. Tanabe, N. Ikarashi, M. Saitoh, T. Nagumo, H. Sunamura, M. Narihiro, K. Uejima, K. Masuzaki, N. Furutake, S. Saito, Y. Yabe, A. Mitsuiki, K. Takeda, T. Hase, Y. Hayashi, *Symp. VLSI Technol.* **2015**, p. T108.
[107] Z. Wei, T. Takagi, Y. Kanzawa, Y. Katoh, T. Ninomiya, K. Kawai, S. Muraoka, S. Mitani, K. Katayama, S. Fujii, R. Miyanaga, Y. Kawashima, T. Mikawa, K. Shimakawa, K. Aono, *IEEE Int. Electron Devices Meet.* **2011**, p. 31.
[108] X. Y. Xue, W. X. Jian, J. G. Yang, F. J. Xiao, G. Chen, X. L. Xu, Y. F. Xie, Y. Y. Lin, R. Huang, Q. T. Zhou, J. G. Wu, *Symp. VLSI Circuits* **2012**, p. 42.
[109] R. Aluguri, T.-Y. Tseng, *IEEE J. Electron Devices Soc*. **2016**, *4*, 294.
[110] Y. Deng, P. Huang, B. Chen, X. Yang, B. Gao, J. Wang, L. Zeng, G. Du, J. Kang, X. Liu, *IEEE Trans. Electron Devices* **2013**, *60*, 719.
[111] H.-Y. Chen, S. Brivio, C.-C. Chang, J. Frascaroli, T.-H. Hou, B. Hudec, M. Liu, H. Lv, G. Molas, J. Sohn, S. Spiga, V. M. Teja, E. Vianello, H.-S. P. Wong, *J. Electroceram*. **2017**, *39*, 21.
[112] P.-Y. Chen, R. Fang, R. Liu, C. Chakrabarti, Y. Cao, S. Yu, *IEEE Int. Symp on Hardware Oriented Security and Trust* **2015**, p. 26.
[113] M. N. Bojnordi, E. Ipek, *IEEE Int'l Symp. High-Performance Comp. Architecture* **2016**, p. 1.
[114] S. B. Eryilmaz, D. Kuzum, R. Jeyasingh, S. Kim, M. BrightSky, C. Lam, H.-S. P. Wong, *Front. Neurosci*. **2014**, *8*, 205.
[115] F. Alibart, E. Zamanidoost, D. B. Strukov, *Nat. Commun*. **2013**, *4*, 2072.
[116] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, D. B. Strukov, *Nature* **2015**, *521*, 61.
[117] C. Du, F. Cai, M. A. Zidan, W. Ma, S. H. Lee, W. D. Lu, *Nat. Commun*. **2017**, *8*, 2204.
[118] J. Torrejon, M. Riou, F. A. Araujo, S. Tsunegi, G. Khalsa, D. Querlioz, P. Bortolotti, V. Cros, K. Yakushiji, A. Fukushima, H. Kubota, S. Yuasa, M. D. Stiles, J. Grollier, *Nature* **2017**, *547*, 428.
[119] W. Schiffmann, M. Joost, R. Werner, *Optimization of the Back-propagation Algorithm for Training Multilayer Perceptrons*, Technical Report, Institute of Physics, University of Koblenz **1994**.
[120] P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang, W. Zhang, Q. Zhang, N. Deng, L. Shi, H.-S. P. Wong, H. Qian, *Nat. Commun*. **2017**, *8*, 15199.
[121] C. Li, Y. Li, H. Jiang, W. Song, P. Lin, Z. Wang, J. J. Yang, Q. Xia, M. Hu, E. Montgomery, J. Zhang, N. Dávila, C. E. Graves, Z. Li, J. P. Strachan, R. S. Williams, N. Ge, M. Barnell, Q. Wu, *IEEE Int. Symp. Circuits. Syst.* **2018**, p. 1.
[122] F. M. Bayat, M. Prezioso, B. Chakrabarti, H. Nili, I. Kataeva, D. B. Strukov, *Nat. Commun*. **2018**, *9*, 2331.
[123] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang, W. Song, J. P. Strachan, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, Q. Xia, *Nat. Commun*. **2018**, *9*, 2385.
[124] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. L. Gallo, K. Moon, J. Woo, H. Hwang, Y. Leblebici, *Adv. Phys. X* **2017**, *2*, 89.
[125] S. Yu, *Proc. IEEE* **2018**, *2*, 260.
[126] D. Kuzum, S. Yu, H.-SP. Wong, *Nanotechnology* **2013**, *24*, 382001.
[127] S. Yu, P. Y. Chen, Y. Cao, L. Xia, Y. Wang, H. Wu, *Int. Electron Devices Meet.* **2015**, 17-3.
[128] T. Gokmen, Y. Vlasov, *Front. Neurosci*. **2016**, *10*, 333.
[129] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, *Proc. USENIX Symp. Oper. Syst. Des. Implement. (OSDI)* **2016**.
[130] J. Bill, R. Legenstein, *Front. Neurosci*. **2014**, *8*, 412.
[131] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. di Nolfo, S. Sidler, M. Giordano, M. Bodini, N C. P. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, G. W. Burr, *Nature* **2018**, *558*, 60.
[132] L. Gao, P.-Y. Chen, S. Yu, *IEEE Electron Device Lett*. **2016**, *37*, 870.
[133] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, V. Srikumar, *ACM/IEEE 43rd ISCA* **2016**, *44*, 14.
[134] P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, Y. Xie, *ACM SIGARCH Computer Architecture News* **2016**, *44*, 27.
[135] L. Song, X. Qian, H. Li, Y. Chen, *IEEE Conf. on High Performance Computer Architecture (HPCA)* **2017**, p. 541.
[136] M. Cheng, L. Xia, Z. Zhu, Y. Cai, Y. Xie, Y. Wang, H. Yang, *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst*. **2018**. https://doi.org/10.1109/TCAD.2018.2824304
[137] M. Verleysen, B. Sirletti, A. Vandemeulebroecke, P. G. A. Jespers, *IEEE Trans. Circuits Syst*. **1989**, *36*, 762.
[138] J. J. Hopfield, D. W. Tank, *Biol. Cybern*. **1985**, *52*, 141.
[139] S. B. Eryilmaz, D. Kuzum, R. Jeyasingh, S. Kim, M. BrightSky, C. Lam, H.-S. P. Wong, *Front. Neurosci*. **2014**, *8*, 205.
[140] S. G. Hu, Y. Liu, Z. Liu, T. P. Chen, J. J. Wang, Q. Yu, L. J. Deng, Y. Yin, S. Hosaka, *Nat. Commun*. **2015**, *6*, 7522.
[141] V. Milo, D. Ielmini, E. Chicca, *IEEE Int. Electron Devices Meet.* **2017**, pp. 263–266.
[142] P. M. Sheridan, F. Cai, C. Du, W. Ma, Z. Zhang, W. D. Lu, *Nat. Nanotechnol*. **2017**, *12*, 784.
[143] Y. Jeong, J. Lee, J. Moon, J. H. Shin, W. D. Lu, *Nano Lett*. **2018**, *18*, 4447.
[144] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves, Z. Li, J. P. Strachan, P. Lin, Z. Wang, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, Q. Xia, *Nat. Electron*. **2018**, *1*, 52.
[145] J. B. MacQueen, *Proc. 5th Berkeley Symp.* **1967**, p. 281.
[146] M. A. Zidan, A. Chen, G. Indiveri, W. D. Lu, *J. Electroceram*. **2017**, *39*, 4.

[147] E. O. Neftci, B. U. Pedroni, S. Joshi, M. Al-Shedivat, G. Cauwenberghs, *Front. Neurosci*. **2016**, *10*, 241.
[148] S. Yu, P.-Y. Chen, Y. Cao, L. Xia, Y. Wang, H. Wu, *IEEE Int. Electron Devices Meet.* **2015**, p. 17.
[149] M. A. Zidan, Y. Jeong, J. Lee, B. Chen, S. Huang, M. J. Kushner, W. D. Lu, *Nat. Electron*. **2018**, *1*, 411.
[150] H. Simon, T. Zacharia, R. Stevens, *Dep. Energy Tech. Rep.* **2007**.
[151] T. Palmer, *Nature* **2015**, *526*, 32.
[152] N. Aage, E. Andreassen, B. S. Lazarov, O. Sigmund, *Nature* **2017**, *550*, 84.
[153] P. M. Altrock, L. L. Liu, F. Michor, *Nat. Rev. Cancer* **2015**, *15*, 730.
[154] L. Chua, V. Sbitnev, H. Kim, *Int. J. Bifurc. Chaos* **2012**, *22*, 1250098.
[155] C. H. Bailey, M. Giustetto, Y-Y. Huang, R. D. Hawkins, E. R. Kandel, *Nature Rev. Neurosci*. **2000**, *1*, 11.
[156] M. Chistiakova, N. M. Bannon, M. Bazhenov, M. Volgushev, *Neuroscientist* **2014**, *20*, 483.
[157] S. Kumar, J. P. Strachan, R. S. Williams, *Nature* **2017**, *548*, 318.