SCIENCE CHINA
Information Sciences

• REVIEW •

December 2021, Vol. 64 221402:1–221402:46
https://doi.org/10.1007/s11432-021-3327-7

In-memory computing with emerging nonvolatile memory devices

Caidie CHENG[1,2†], Pek Jun TIW[2†], Yimao CAI[2,3], Xiaoqin YAN[1*], Yuchao YANG[2,3,4*] & Ru HUANG[2,3,4*]

[1]*State Key Laboratory for Advanced Metals and Materials, School of Materials Science and Engineering, University of Science and Technology Beijing, Beijing* 100083*, China;*
[2]*Key Laboratory of Microelectronic Devices and Circuits (MOE), Department of Micro/nanoelectronics, Peking University, Beijing* 100871*, China;*
[3]*Center for Brain Inspired Chips, Institute for Artificial Intelligence, Peking University, Beijing* 100871*, China;*
[4]*Center for Brain Inspired Intelligence, Chinese Institute for Brain Research (CIBR), Beijing, Beijing* 102206*, China*

Received 29 April 2021/Revised 30 June 2021/Accepted 24 August 2021/Published online 4 November 2021

Abstract The von Neumann bottleneck and memory wall have posed fundamental limitations in latency and energy consumption of modern computers based on von Neumann architecture. In-memory computing represents a radical shift in the computer architecture that can address such problems by merging computing functions within the memory itself. In this article, we review the emerging nonvolatile memory devices, such as resistance-based and charge-based memory devices, that are explored for in-memory computing applications. We will provide an overview of the materials, mechanisms, and integration of these devices, and discuss the optimizations at the device and array levels that are required to better support in-memory computing. Recent progress in the application of in-memory computing in artificial neural networks, spiking neural networks, digital logic in memory as well as hardware security will also be discussed. Finally, we will discuss the remaining challenges in this field and potential pathways to address them.

Keywords in-memory computing, von Neumann bottleneck, nonvolatile memory, energy efficiency, neural network

Citation Cheng C D, Tiw P J, Cai Y M, et al. In-memory computing with emerging nonvolatile memory devices. Sci China Inf Sci, 2021, 64(12): 221402, https://doi.org/10.1007/s11432-021-3327-7

1 Introduction

With the advent of the era of big data, higher requirements have been put forward for the speed, density, power consumption and cost of high-performance computing. The increasing amounts of calculations cause the appearance of the disadvantages of the original von Neumann-based computing structure. For example, the separation of storage area and calculation area causes the frequent data transported back and forth between them, resulting in a large amount of energy loss and signal delay, which ultimately leads to an increase in chip power consumption and a decrease in computing efficiency (memory wall) [1, 2]. Moreover, under the von Neumann architecture, data and instructions are stored in the same memory. As the result, the fetching of data and instructions cannot be performed at the same time, otherwise it will cause confusion in memory access, and development to the present, the computing speed of the central processing unit (CPU) has far exceeded the access speed of memory, greatly reducing the utilization of the CPU (von Neumann bottleneck) [3, 4]. At the same time, under the guidance of "Moore's Law", microelectronic devices represented by complementary metal-oxide-semiconductor (CMOS) devices are showing a trend of rapid growth in number and integration. However, as device sizes approach physical limits and chip energy consumption increases rapidly, this growth trend has gradually slowed down and is difficult to maintain [5, 6].

* Corresponding author (email: xqyan@mater.ustb.edu.cn, yuchaoyang@pku.edu.cn, ruhuang@pku.edu.cn)
† Cheng C D and Tiw P J have the same contribution to this work.

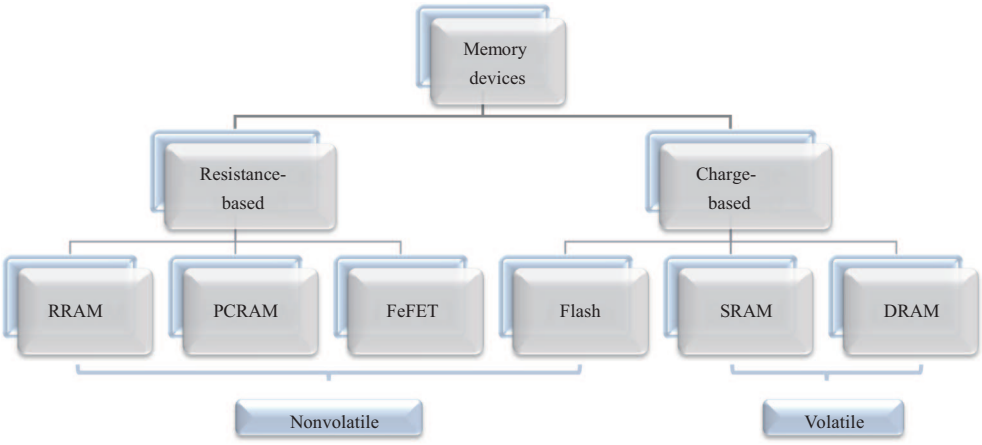


Figure 1 (Color online) Memory devices for in-memory computing.

There are many attempts to solve these problems. The adoption of a memory hierarchy has been proposed to alleviate this problem, in which a series of high-speed caches between the processor and the dynamic storage unit buffer the speed mismatch between memory access and operation. This method reduces the average delay of calculation to a certain extent [7,8]. Furthermore, the graphics processing unit (GPU) is also useful, in processing a large amount of data in parallel, thus greatly improving the efficiency of calculation [9]. Moreover, dedicated acceleration chips such as the tensor processing unit (TPU) are designed to meet the computing power requirements for data streams. This type of acceleration hardware generally has strong parallel processing capabilities and larger data bandwidth [10, 11]. In addition, the memory is prepared as close as possible to the processor (near memory computing) to reduce the distance of data transmission [12, 13]. However, the above improvement methods did not focus on the separation of data storage and data processing. Therefore, although the memory wall problem could be alleviated to a certain extent, the von Neumann bottleneck cannot be eliminated fundamentally.

At present, scholars and institutions have begun to study computing storage, that is, in-memory computing [14, 15]. The main idea of in-memory computing is to integrate computing (processing) functions and storage functions in the same chip. All calculations are implemented inside the storage without data readout and programming, greatly reducing the time and power consumption of data moving back and forth between storage and processor. In this review, we start with the electrical characteristics of the memory device, followed by its resistance transformation physical mechanism, before discussing the array integration method to optimize the performance of the memory device in massively parallel and efficient in-memory computing, such as digital state logic computing and analog brain-like computing. The vector matrix multiplication (VMM) can be applied to accelerate the training and inference in an artificial neural network (ANN). At present, classical machine learning tasks such as information coding, data classification and reinforcement learning have been realized. Chips made based on this principle can complement GPUs to jointly accelerate the processing of massive data in the information age.

2 Nonvolatile memory devices for in-memory computing

Memory device is fundamental in realizing the in-memory computing [16–19]. At present, the mainstream research and development of integrated storage and computing chips are concentrated on traditional volatile memories, including dynamic random access memory (DRAM) [20, 21], static random access memory (SRAM) [22], and non-volatile memories, such as resistance random access memory (RRAM) [23], phase-change random access memory (PCRAM) [24], ferroelectric devices (FeRAM) [25], and flash [26]. As schematically illustrated in Figure 1, there are two basic ways to store information in the memory devices mentioned above, one of which is based on the presence and absence of electric charges [27, 28], while the other is based on changes in resistance caused by the rearrangement of atoms and the reversal of the ferroelectric polarization direction [29, 30]. In this section, we focus on the materials, mechanism, and integration of resistance-based memories, before introducing charge-based memories.

2.1 Resistance-based memories

The conductance of resistance-based memories (memristors) will change with the amount of current passed, and the switching between high resistance (HR) and low resistance (LR) states of the devices is reversible and can be realized by applying an electric. Memristors are considered to be the fourth basic circuit element besides resistors, inductors, and capacitors [29,30], and are regarded as the next generation of non-volatile storage technology, featuring high speed, low power consumption, high integration, and compatibility with CMOS technology [31–33]. Such advantages can meet the performance needs of general electronic devices for high-density information storage and high-performance computing. At the same time, the memristor can carry out non-volatile state logic operations, and in-situ calculation while allowing data storage in a single device. Therefore, memristors can be used to realize in-memory computing, fundamentally solving the problem of separation of storage and calculation [34, 35]. This subsection will summarize the materials, mechanism, and integration of the memristor.

2.1.1 *Materials*

Memristor utilizes the reversible electro-resistive effect of the insulator material in the metal-insulator-metal (MIM) structure to realize the storage function. It is worth noting that the "metal" here can be any electrical conductor [30]. Although there are diverse resistive materials exhibiting resistive switching (RS) phenomenon, some criteria for memory components are still required. From a material perspective, as long as the performance of the device is not affected by resistive materials, CMOS compatibility is one of the main issues for cost-effective and high-density storage. Besides, the RS materials are the core of memristors which will reflect different RS characteristics under the different types of materials. Therefore, in this subsection, we will focus on introducing several common types of resistive materials, namely binary oxide materials [36,37], phase change materials [38,39], ferroelectric materials [40,41], two dimensional (2D) materials [42,43] and polymer materials [44,45].

(1) Binary oxide material. Binary oxide materials have the advantages of simple structure, easy control of material compositions, low cost, good stability, simple preparation process, and compatibility with CMOS processes [31–33]. Thanks to these properties, binary oxide materials form an important material system in the field of microelectronics. With the introduction of RS memory, binary oxides with RS characteristics have become a hot spot for researchers. Among various transition metal oxide materials, TaO_x and HfO_x have been consistently studied as the most promising material systems for emerging memory technologies, and they are among the most mature resistive materials [36, 37, 46, 47]. Therefore, we will focus on binary oxides that have been extensively studied by researchers in academia and industry.

A Ta/Ag-NCs/Ta_2O_5/Pt/Ti memristor based on TaO_x material was fabricated by Li et al. (inset of Figure 2(a)) [48]. The short-term plasticity (STP) and long-term plasticity (LTP) behaviors were realized in both potentiation and depression processes for the first time in this article. The paired-pulse facilitation (PPF) ratio in Figure 2(a) indicated a smaller conductance enhancement under a larger pulse interval. A transition from STP to LTP is observed under the increasing amplitude of the positive pulses (Figures 2(b) and (c)), while a transition from short-term depression (STD) to long-term depression (LTD) is observed as the amplitude of the negative pulses increases (Figures 2(d) and (e)). The electrochemical migration of Ag during the pulses and thermodynamic diffusion during the pulse intervals were responsible for these behaviors. As shown in Figure 2(f), the spike-timing-dependent plasticity (STDP) learning rule is implemented eventually through careful design of pulse pairs, which is an important learning rule based on the inherent mechanism of time in biological systems [49–51]. During the STDP process, the requirement of a higher amplitude of the positive pulses and the appearance of a symmetric pulse pair were owing to a higher Schottky barrier formed on the bottom Ta_2O_5/Pt interface. This work realized various synaptic functions such as the timing-based learning rule in a scalable manner, laying the foundation for building an intelligent neuromorphic system that can encode and process temporal and spatial information.

There are some reports on the research of HfO_2 materials. A highly uniform two-terminal Zr-doped HfO_2 based artificial synapse is reported by Liu et al., which demonstrates excellent performance [52]. From the current-voltage (I-V) characteristics of the TiN/$Hf_{0.5}Zr_{0.5}O_2$(HZO)/Pt devices in Figure 3(a), the analog RS with highly uniformity was realized. More importantly, the devices exhibit LTP and LTD, and demonstrates average cycle-to-cycle variation (CCV) of <4% (Figure 3(b)) and average device-to-device variation (DDV) of <9% in Figure 3(c), reflecting excellent uniformity of the device due to the interface-type resistive switching. Then, the endurance and retention measurements were also carried

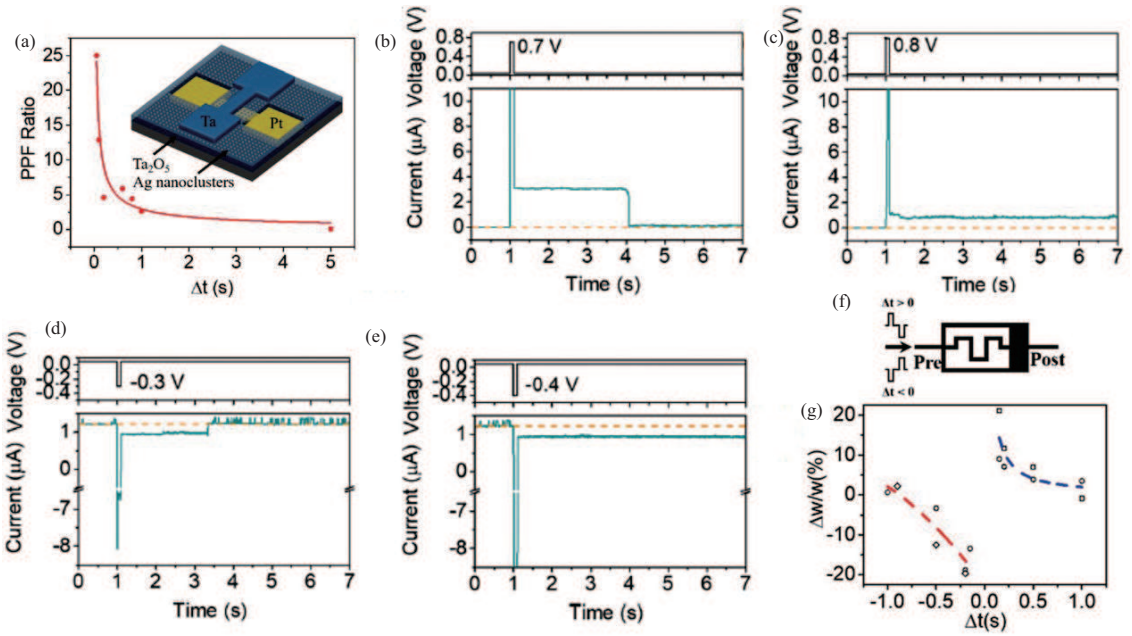


Figure 2 (Color online) (a) PPF ratio of the Ag nanoclusters based devic with the inset showing the schematic of the device. (b) and (c) A transition from STP to LTP. (d) and (e) A transition from STD to LTD. (f) and (g) Implementation of STDP in the device. Ref. [48] @Copyright 2020 The Royal Society of Chemistry.

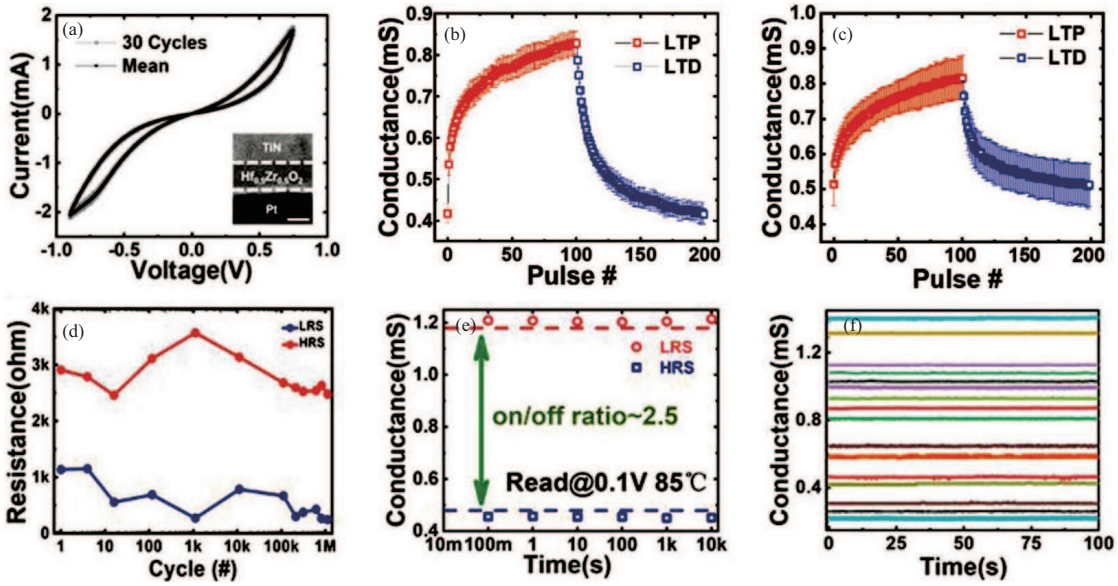


Figure 3 (Color online) Based on the TiN/HZO (5 nm, 600°C/Pt devices). (a) *I-V* curves of the HZO based device with the inset showing the transmission electron microscopy (TEM) image; (b) CCV of the LTP and LTD; (c) DDV of the LTP and LTD; (d) endurance test; (e) retention characteristic measured at 85°C; (f) the measurement of weight states. Ref. [52] @Copyright 2020 John Wiley and Sons Inc.

out to demonstrate more than 10^6 cycles of endurance (Figure 3(d)) and $>10^4$ s of retention at 85°C (Figure 3(e)), as well as 4-bit weights for reliable storage (Figure 3(f)). These characteristics are essential in realizing artificial neural networks. This article greatly improved the performance of the devices and increased its potential for network training.

Besides, a device with yttria stabilized zirconia (YSZ) was fabricated by Cheng et al. [53]. The TEM image of the Ti/YSZ/Pt device with 12 nm YSZ is displayed in Figure 4(a). Furthermore, the typical bipolar RS behavior was demonstrated as the YSZ thickness of the device is 7 nm (Figure 4(b)). A transition from bipolar to unipolar RS was realized as the YSZ thickness film was increased to 12 nm as displayed in Figure 4(c). This can be attributed to the competition between the vertical and radial ion

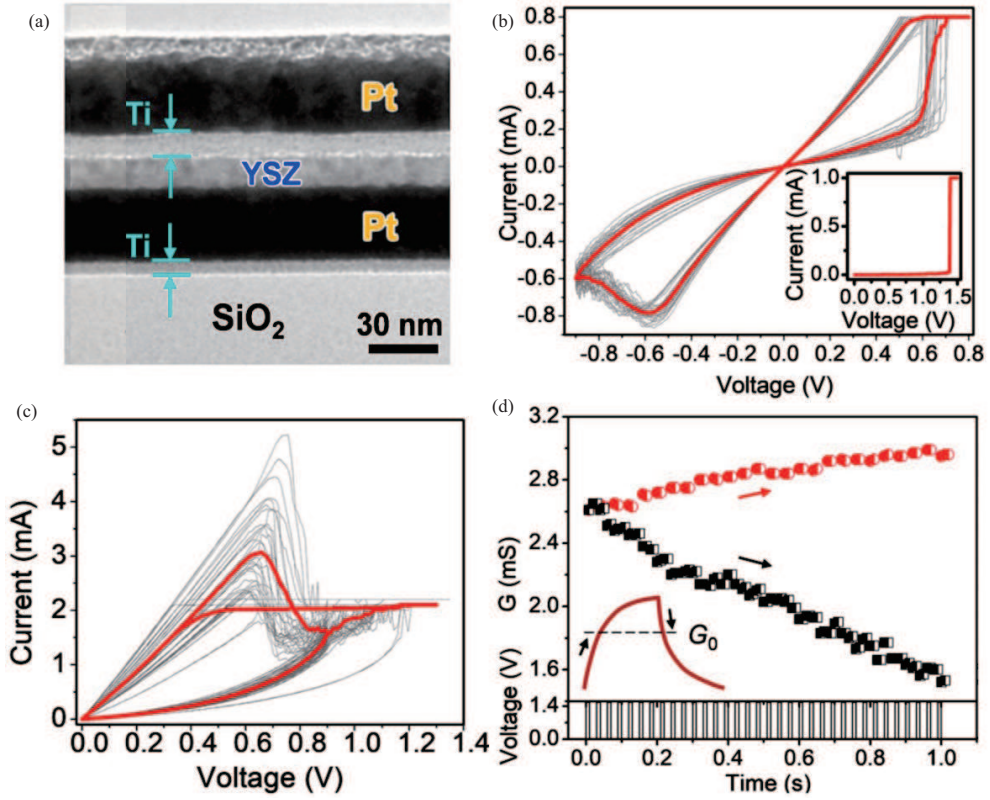


Figure 4 (Color online) (a) TEM image of the device with 12 nm YSZ; (b) bipolar characteristics of the device with the inset showing the forming process; (c) unipolar characteristics of the device; (d) implementation of metaplasticity. Ref. [53] @Copyright 2018 AIP Publishing.

transport dynamics. Metaplasticity, which is a history dependent plasticity [54, 55], was realized due to the unipolar switching characteristic. As shown in Figure 4(d), the device exhibited either potentiation or depression with identical pulse stimulations, which can be ascribed to the different operation histories of the device. This work demonstrated that the manipulation of ionic transport properties can be utilized to construct synaptic elements with rich functionalities.

In addition to the above memristive materials, oxide materials with optoelectronic properties as memristor functional materials have also attracted extensive research attention of researchers, because such photovoltaic or optoelectronic materials can respond to both optical stimuli and electrical stimuli at the same time [56–58]. The introduction of a new physical stimulus such as light can enrich the characteristics, thereby broadening the control methods and application fields of the memristor.

The optoelectronic synapse with a structure of W/MgO/ZnO/Mo was demonstrated by Dang et al. [59]. Figure 5(a) shows the TEM image of the optoelectronic synapse. The multilevel features were obtained by triangular voltage sweeping (Figures 5(b) and (c)). As ZnO is sensitive to short wavelength light [60], the device can be recognized as an optical sensory synapse. Besides, under various light wavelengths and intensity, the LTP and LTD synaptic dynamics were investigated (Figures 5(d)–(f)). It can be explained that light with a shorter wavelength and a higher intensity can generate more carriers in the optoelectronic synapse to increase the conductance [61]. This article shows that the optoelectronic synapse has great potentiality for in-sensor computing.

A vertical 3-terminal device based on TaO_x was demonstrated by Yang et al. [62]. The structure and a high angle annular dark field scanning TEM (HAADF STEM) image of the vertical heterosynaptic devices are schematically illustrated in Figures 6(a) and (b), respectively. Compared with 2-terminal memristors, the structure of the 3-terminal device includes the HfO_2 based sidewalls and the third electrode covering the sidewall. The modulation of RS in TaO_x was realized by applying a modulatory electric field (V_{mod}) on the third electrode. As shown in Figures 6(c) and (d), the LTP and LTD were modulated by different V_{mod} amplitudes. The tunable weight update of the 3-terminal device has a great prospect in neuromorphic computing.

(2) Phase-change materials. When the materials exchange heat with the external environment and

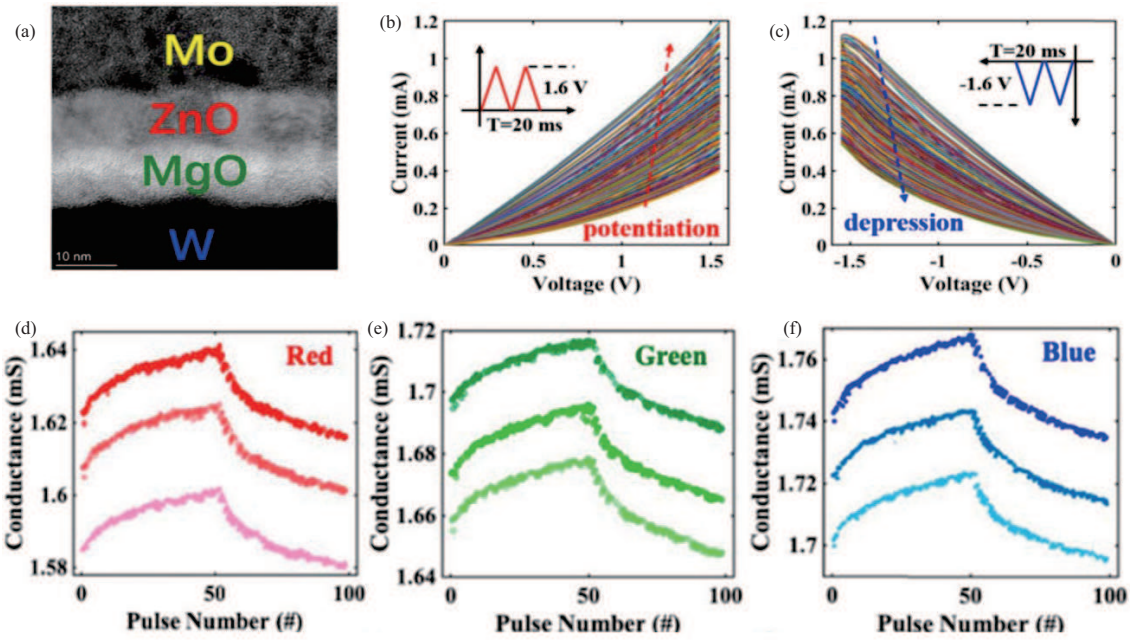


Figure 5 (Color online) (a) TEM image of the optic-neural synapse; (b) and (c) implementation of multilevel characteristics; (d)–(f) LTP and LTD performances under different light stimuli. Ref. [59] @Copyright 2020 IEEE.

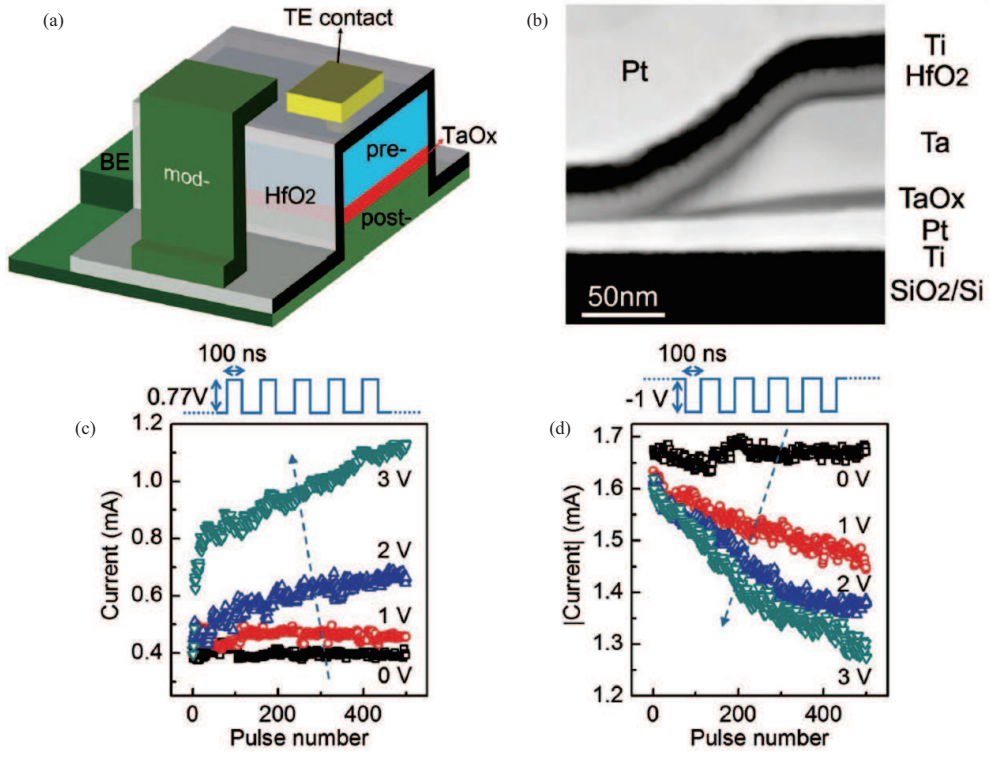


Figure 6 (Color online) (a) The structure of the vertical heterosynaptic devices; (b) HAADF STEM image; (c) conductance facilitation as a function of V_{mod}; (d) conductance depression as a function of V_{mod}. Ref. [62] @Copyright 2017 John Wiley & Sons, Inc.

reach a certain temperature, the physical state of the material will change. This phenomenon is called phase change, and these materials are phase change materials that have the characteristics of chemical stability, rapid conversion between phases, and thermal stability of the amorphous phase [38, 39]. The basic structure of the PCRAM is to add a thin layer of phase change material between the top and bottom electrodes (TE and BE, respectively), and the conductivity difference between the crystalline and

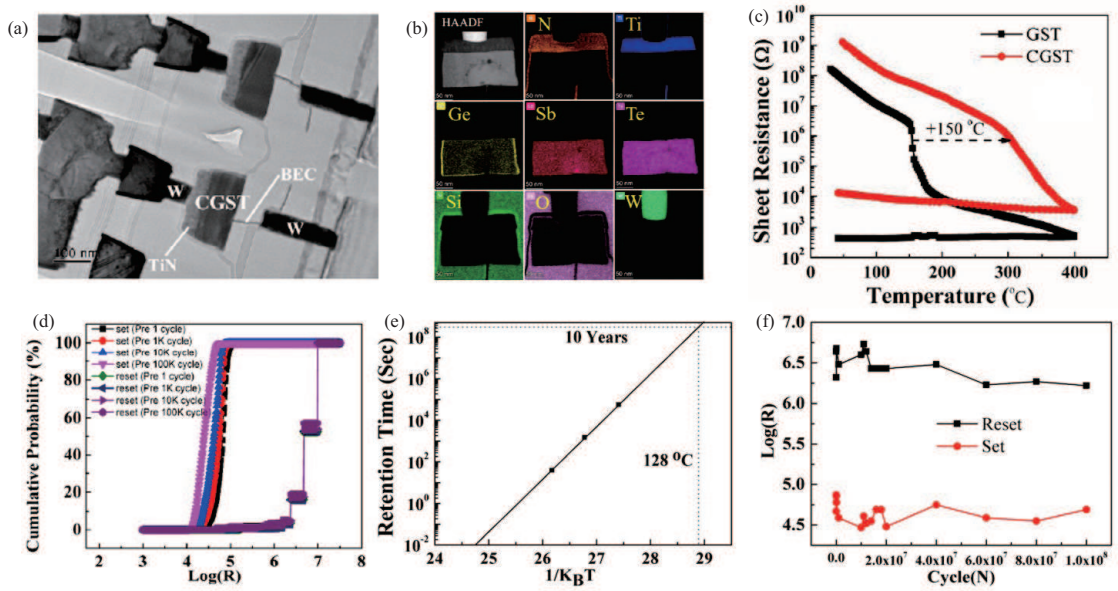


Figure 7 (Color online) (a) TEM image of CGST-based PCRAM in array [68]; (b) element mapping in PCRAM [69] @Copyright 2020 IEEE; (c) temperature-dependent resistance of the PCRAM based on CGST and GST; (d) thermal reliability test of the PCRAM; (e) data retention of the PCRAM; (f) endurance test of the PCRAM [68] @Copyright 2018 IEEE.

amorphous states can be utilized to store information. When an external signal in the form of current injection is applied, the connection points between the electrodes and the phase change materials generate Joule heat causing a phase change. PCRAM is recognized as one of the most promising candidates for non-volatile memory with high response speed, high integration density, low operating voltage, and compatibility with standard CMOS technology [63–65]. As the poor retention will limit its application in embedded systems, it is necessary to improve the performance of PCRAM by optimizing phase change materials [66, 67].

The PCRAM based on Carbon-doped $Ge_2Sb_2Te_5$ (CGST) material with high thermal stability was fabricated by Song et al. [68]. Figure 7(a) shows the TEM of CGST based memory cells in an array with 40 nm node. A lance-shaped CGST film was deposited on the blade BE contact (BEC) area. Furthermore, the element mapping of the PCRAM is displayed in Figure 7(b) [69]. The temperature-dependent resistance of CGST and GST based memory cells indicated that the CGST based PCRAM has a gradual drop in RS compared with the GST based PCRAM (Figure 7(c)). The thermal reliability test of the PCRAM was measured after cycling stress as shown in Figure 7(d) indicating good thermal reliability of the CGST based PCRAM. Figure 7(e) shows the retention time under different temperatures. Extrapolation to a temperature of 128°C shows that the data can be maintained for 10 years. Moreover, the endurance test was performed as shown in Figure 7 reaching 10^8 cycles. The high performance CGST based PCRAM in this article is suitable for the applications in embedded systems [68].

(3) Ferroelectric materials. Ferroelectric material refers to a type of material with ferroelectric effects. Its crystal has spontaneous polarization strengthening, and the direction of the spontaneous polarization can be reoriented to the direction of the applied electric field. Due to the development of new ferroelectric material thin film technology, ferroelectric materials have been used in information storage, image display and holographic pagers, and ferroelectric light valve arrays for holographic storage [40, 70]. FeRAM uses the position of the central atom in the ferroelectric material to store information. Under an applied electric field, the central atom can cross the potential barrier from the original position to another position. Upon removing the electric field, the atom cannot cross the potential barrier, hence the data is retained. FeRAM exhibits low power consumption, fast operation speed, and radiation resistance; therefore it has a wide range of applications in scientific fields such as military and space exploration [71–73]. The current ferroelectric materials for FeRAM are mostly perovskite materials and doped HfO_2. However, there are some problems in integrating classic perovskite materials with modern silicon technology. Compared with classic perovskite materials, the doped HfO_2 is fully compatible with complementary metal oxide semiconductor (CMOS) technology. In addition, the coercive electric of doped HfO_2 is relatively high [73–76].

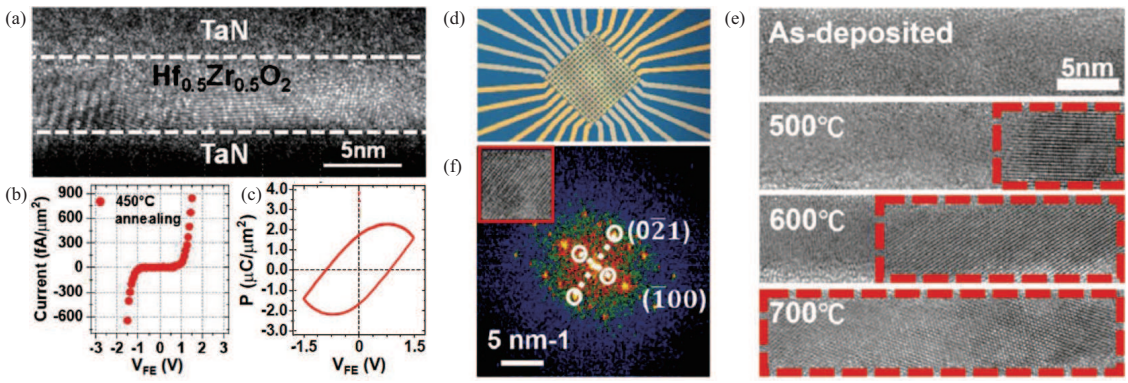


Figure 8 (Color online) (a) TEM image of partially crystallized HZO; (b) measured leakage current of HZO; (c) the *P*-*V* loops of the leaky HZO [77] @Copyright 2019 IEEE; (d) top view of the crossbar array; (e) HRTEM views of HZO with different annealing temperature; (f) FFT on HRTEM images of HZO [52] @Copyright 2020 John Wiley and Sons Inc.

The FE layer designed by Chen et al. [77] demonstrated leaky characteristics. The 5 nm HZO was utilized as the FE layer to obtain a larger depolarization field (Figure 8(a)). Moreover, the partial crystallization was realized by 450°C annealing (Figure 8(b)) to implement a higher leakage current in the HZO layer [78]. The polarization-voltage (*P*-*V*) loop of the leaky HZO was measured as shown in Figure 8(c), confirming the impacts of leakage current in the designed leaky HZO layer. In addition, Liu et al. [52] fabricated a crossbar array based on polycrystalline TiN/HZO/Pt organized in an array with a size of 16×16 as shown in Figure 8(d). The cross-section TEM views of the HZO layer displayed different crystalline states under different rapid thermal process (RTP) conditions (Figure 8(e)). The transition from the amorphous state to the polycrystalline structure of the HZO can be realized when the temperature of RTP is above 500°C and the increase in RTP temperature will increase the crystallinity. In particular, the fast Fourier transformation (FFT) on the crystalline region was in accordance with the formation of the orthorhombic phase as shown in Figure 8(f) [79].

Neuromorphic computing can mimic the architecture of the human brain to deal with complicated compute tasks, in which the realization of the neuronal dynamics is critical [80, 81]. As CMOS neurons have higher hardware costs in implementing advanced biomimetic functions, a leaky-FeFET (L-FeFET) with fast polarization degradation was proposed by Chen et al. [77] to mimic biological neuron behaviors. Figure 9(a) shows the schematic structure of the proposed L-FET based on leaky HZO, and the FE polarization switching was displayed in Figure 9(b). Figure 9(c) displays the current accumulation and dropping effect of the L-FeFET device. The accumulation dynamics were observed upon applying voltage pulses and the fast decay of drain current ($I_{\rm D}$) was realized after the removal of gate voltage ($V_{\rm G}$). Besides, the L-FeFET can implement a leaky integrate-and-fire (LIF) function with only one transistor and one resistor as well as spike-frequency adaption (SFA) function with capacitor [82–84]. With input spikes applied to the gate of the L-FeFET, the device threshold voltage ($V_{\rm th}$) becomes lower due to the accumulation effect. Figure 9(d) demonstrates the self-reset firing function of the L-FeFET. The firing of spikes occurs as the $V_{\rm th}$ is pulled down by the accumulation effect, causing the output voltage to drop. The FE polarization then recovers under the dropping effect resulting in a self-reset process. Furthermore, Figure 9(e) displays the experimentally obtained spike firing pattern for SFA. At first, the time interval increases as the accumulation effect dominates. The increase then ceases subsequently due to the polarization degradation of L-FeFET.

Subsequently, based on the capacitor-less L-FeFET, Luo et al. [85] emulated the influence of excitatory and inhibitory inputs to mimic biomimetic neuronal dynamics. Figure 9(f) displays the $I_{\rm D}$ response under positive or negative input pulses, indicating strong dependences on the voltage of the L-FeFET and mimicking the excitatory and inhibitory accumulation of LIF neuron. In addition, the proposed device can also mimic the desensitization of the neuron when inhibition is blocked resulting in an abnormal firing as shown in Figure 9(g). However, under suitable inhibitory stimuli, the L-FeFET neuron will fire normally (Figure 9(h)). These studies show that the L-FET has great application prospects in neuromorphic computing with low energy consumption.

As the gate leakage of the HfO_2 layer increases, the endurance of FeFET is limited to less than 10^4 as well as the memory window will be degraded. The endurance and retention of Si-doped HfO_2(HSO) based FeFET was improved by the insetting of SiON interface layer as shown in Figure 10(c) and (d) [86].

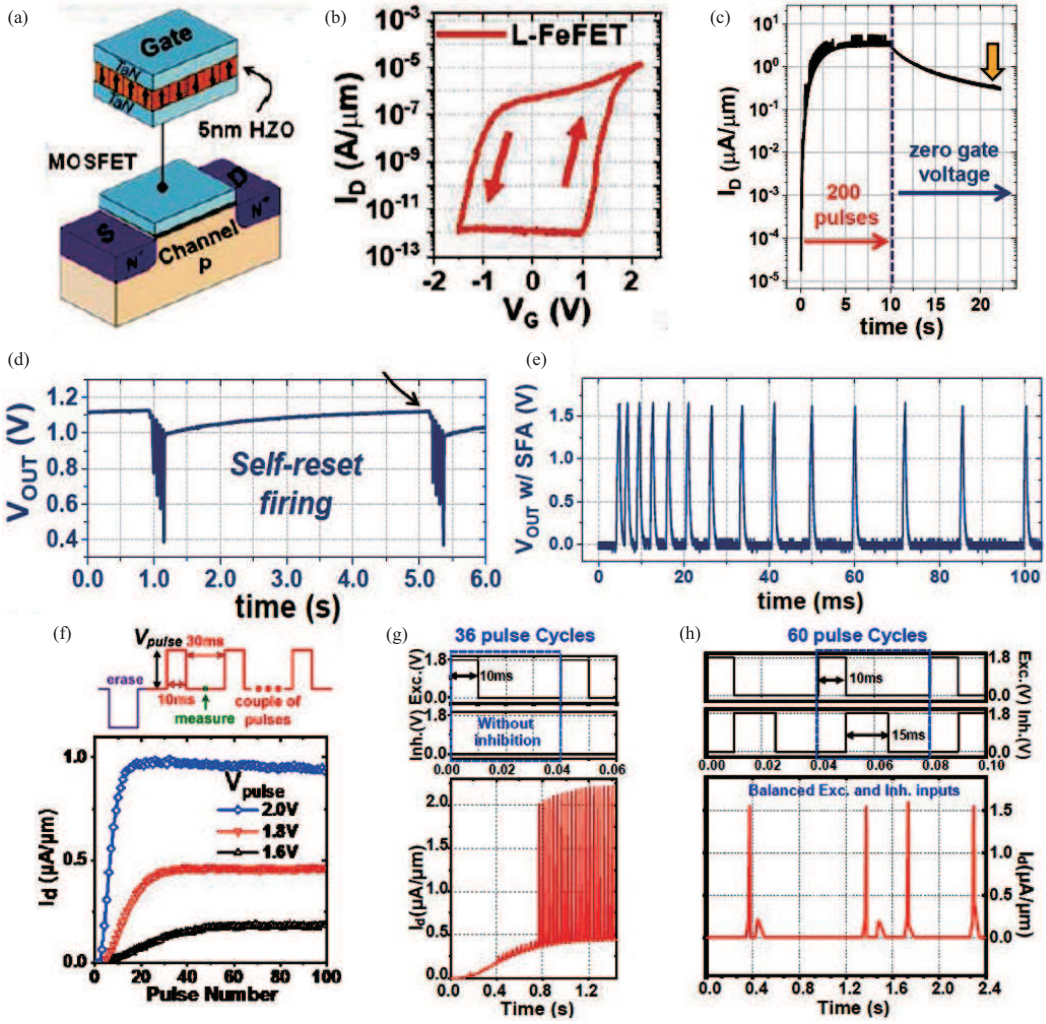


Figure 9 (Color online) (a) Schematic structure of the proposed L-FET; (b) measured I_D-V_G curves of the designed L-FeFET; (c) current accumulation and dropping effect of the device; (d) self-reset firing of the L-FeFET neuron; (e) measured spike firing pattern with SFA [77] @Copyright 2019 IEEE; (f) I_D respond under positive input pulse; (g) and (h) measured spikes without or with moderate inhibitory stimuli, respectively [85] @Copyright 2019 IEEE.

This is caused by the mismatch of the valence band shift between the SiON interface layer and HfO_2 [87] which will lead to more charge tunneling in the HfO_2 rather than trapping.

And the multilevel and high uniformity characteristics were realized based on the FeFET. As shown in Figure 11, the Laminate FeFET based on HSO or HZO was fabricated to realize the 1–3 bit/cell operation [88]. And the Al_2O_3 interlayer was utilized to stabilize the property of FE [89]. Figure 11(a) displayed the different stacks of the laminated FE layer which was utilized to construct the FeFET. And Figure 11(b) shows the I_D-V_G characteristics of the FeFET based on different stack layers, suggesting that the memory window increases accompanied by the stack layer. As the area of the device will cause the variability of the FeFET, the memory window under the different FeFET dimensions was measured. And there is a small variation that was found under the large dimension of the FeFET. It should be noted that less variation is beneficial to realize the multilevel characteristic because the less unstable intermediate state was formed. And under the different threshold voltage, the 1-3 bit/cell operation was realized. The retention of the multilevel based on laminated HSO and HZO FeFET was displayed in Figure 11(d) and (e), respectively.

(4) 2D materials. 2D materials received significant attention because of their atomic thickness, enabling radical length scaling without non-ideal short-channel effects [90, 91]. As the technology of 2D materials matures and with the weak van der Waals (vdWs) heterostructures of the 2D materials, the performance of memristors based on 2D materials is further enhanced, the hardware cost and power consumption are also reduced [92]. 2D materials can be divided into transition metal dichalcogenides (TMDCs) [92, 93]

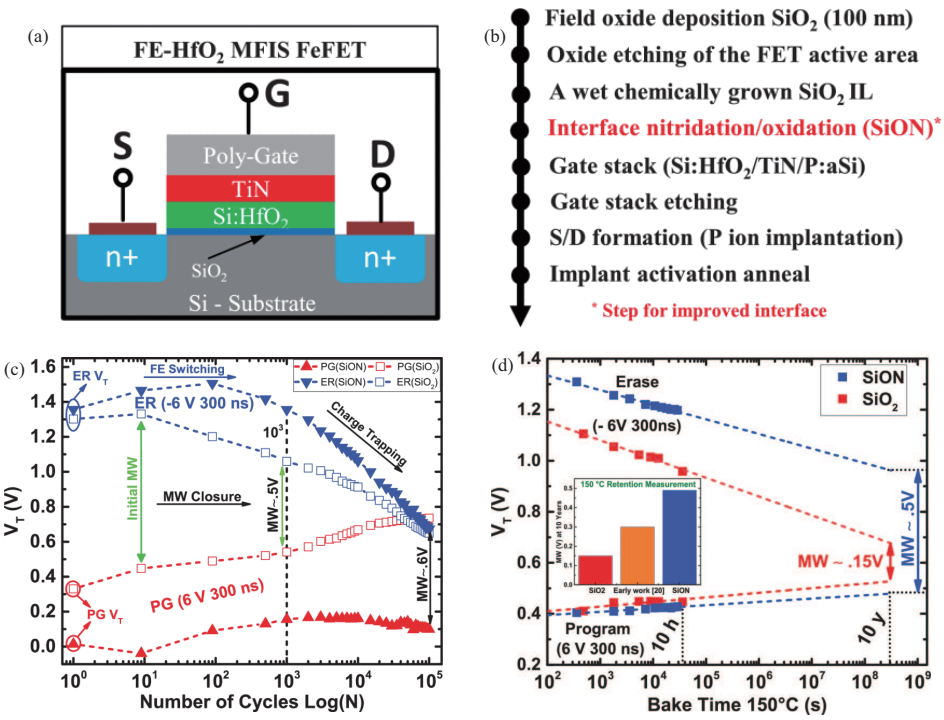


Figure 10 (Color online) (a) The structure schematic of the Si-HfO_2 based FeFET; (b) the fabrication flow of the FeFET; (c) the improvement of endurance with the inserting of SiON interface layer; (d) the improvement of retention with the inserting of SiON interface layer [86] @Copyright 2018 IEEE.

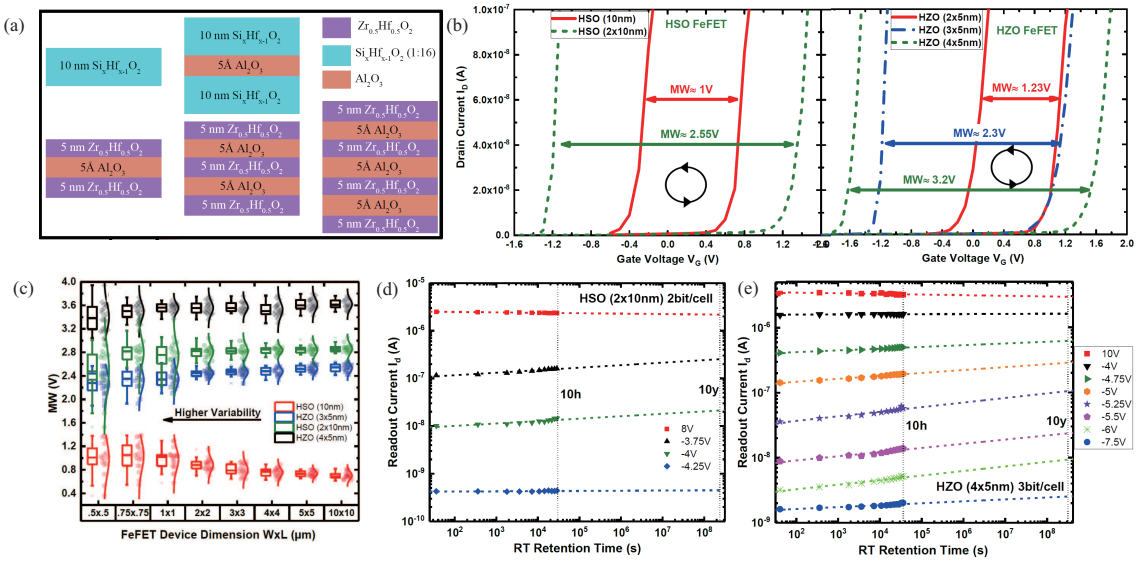


Figure 11 (Color online) (a) The stack of the laminated FE layer; (b) the I_D-V_G characteristic of the FeFET based on different stack layers; (c) the memory window under the different FeFET dimensions; (d) the retention of the 2 bit/cell based on laminated HSO FeFET; (e) the retention of the 3 bit/cell based on laminated HZO FeFET. Ref. [88] @Copyright 2019 IEEE.

and elemental semiconductors such as black phosphorus (BP). BP materials have broader vdWs gaps compared with TMDCs [94, 95]. In 2D materials, the free carriers will move between layers through vdWs gap which will act as a tunnel barrier. There is also significant progress in 2D devices which can be utilized for data storage [96].

Synaptic transistor based on 2D materials such as WSe_2, $NiPS_3$ and $FePSe_3$ with a side gate covered by a polymer electrolyte (PEO: $LiClO_4$ ion gel) had been investigated (Figure 12(a)) [97]. Figure 12(b) schematically illustrates the structure of the WSe_2 with hexagonal symmetry [98]. Furthermore, the characterization of WSe_2 with layer-by-layer stacking is shown in Figures 12(c) and (d). The STP was

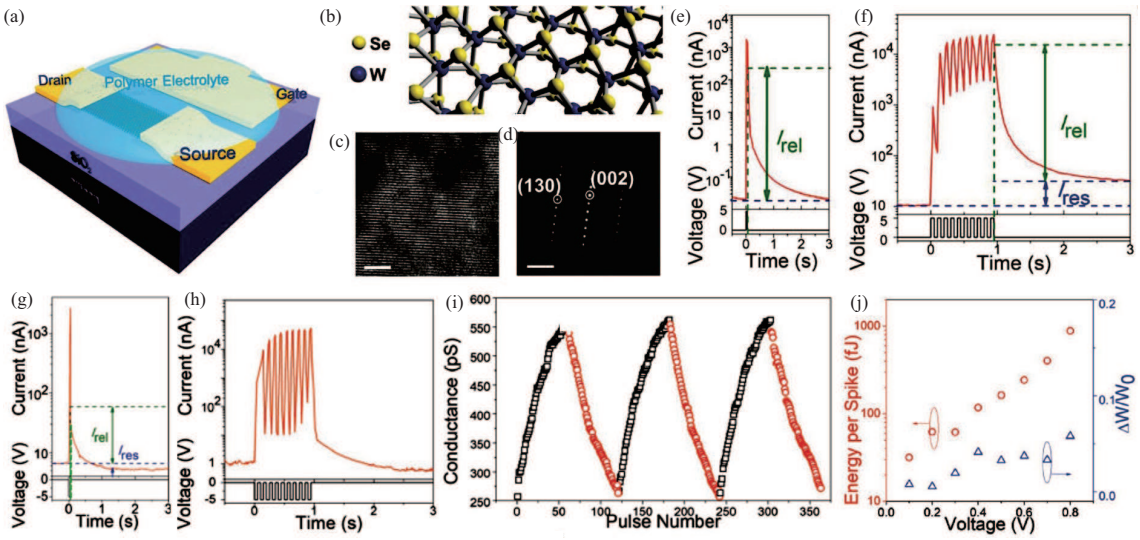


Figure 12 (Color online) (a) The structure of the synaptic transistor; (b) schematic of WSe_2 structure; (c) and (d) HRTEM image and SAED pattern of WSe_2, scale bar: 5 nm; (e) STP measurement under a single positive V_G; (f) STP to LTP measurement under a train of positive V_G; (g) LTD measurement under a single negative V_G; (h) STP measurement under a train of negative V_G; (i) LTP and LTD measurement of the WSe_2; (j) energy consumption and weight change under different pulse magnitudes [97] @Copyright 2018, John Wiley & Sons, Inc.

realized under a 5 V voltage on the gate (Figure 12(e)), while a transition from STP to LTP was observed under a successive train of positive voltage pulses (Figure 12(f)). It can be explained that the adsorption of Li^+ ion on the surface of WSe_2 under the weak stimulus which will backward diffuse into the PEO: $LiClO_4^-$ ion gel after removing the gate pulse, while under the strong stimulus, some of the Li^+ ions will intercalate into WSe_2. Moreover, the ClO_4^- ions can be absorbed on the surface of the WSe_2 under the negative voltage to increase the conductance of the channel. In this case, the LTD characteristic can be realized under a negative gate pulse as the Li^+ ions can be extracted from the channel (Figure 12(g)). Figure 12(h) shows the STP characteristic under a negative voltage pulse train, suggesting the existence of complicated ion dynamics inside the device. In addition, LTP and LTD with remarkable linearity and symmetry were realized as shown in Figure 12(i). Besides, the WSe_2 based synaptic transistors have an ultralow energy consumption of ≈30 fJ per spike as demonstrated in Figure 12(j). It is worth noting that the LTP still occurred under 0.1 V pulse voltage.

Other than WSe_2 material, the BP materials such as $NiPS_3$ and $FePSe_3$ can be utilized as channel material [99, 100]. Figures 13(a) and (b) show the HRTEM of $NiPS_3$ and $FePSe_3$ films, respectively, suggesting the existence of structural defects. The PPF ratio was measured under two consecutive pulses with a varying time interval as shown in Figures 13(c) and (d)), which indicates that the capacity of Li^+ ions in the channel can be altered. In addition to the above 2D materials, MoS_2, another member of TMDCs, also has been extensively studied. Bao et al. [101] fabricated a MoS_2 based dual-gate transistor. The HRTEM in Figure 13(e) displays the multilayer structure of the MoS_2. Similar to Li^+ in WSe_2, Li^+ ions also can be adsorbed onto the surface of MoS_2 under a low stimulus while being intercalated into the channel under a large stimulus. The MoS_2 neuristor can be recognized as an n-type MOSFET, a synapse, or a neuron under different driving signals (Figures 13(f)–(h)). Figure 13(g) shows the STP characteristic under a single pulse while Figure 13(h) displays the integrate-and-fire function with different amplitude of input pulses. These studies demonstrated the great application prospect of 2D materials in enriching the functionalities of neuromorphic computing.

(5) Polymer materials. The chemical and biochemical signals can be transformed into optical, electrical, thermal, and mechanical signals by regulating the transport of ions and molecules in polymer materials, and vice versa [44, 45]. The atoms are connected by covalent bonds without free electrons and ions. Thus while the polymer material has high insulation, elasticity, abrasion resistance, chemical stability, and excellent light transmittance, it also has the characteristics of low dielectric constant and dielectric loss. These properties broaden the application of polymer materials [102, 103]. Polymer RRAM is an emerging device in organic electronic products. It has the dual advantages of electro-electronic materials and resistive memory, including low cost, flexible, transparent, and simple manufacturing process [104–106].

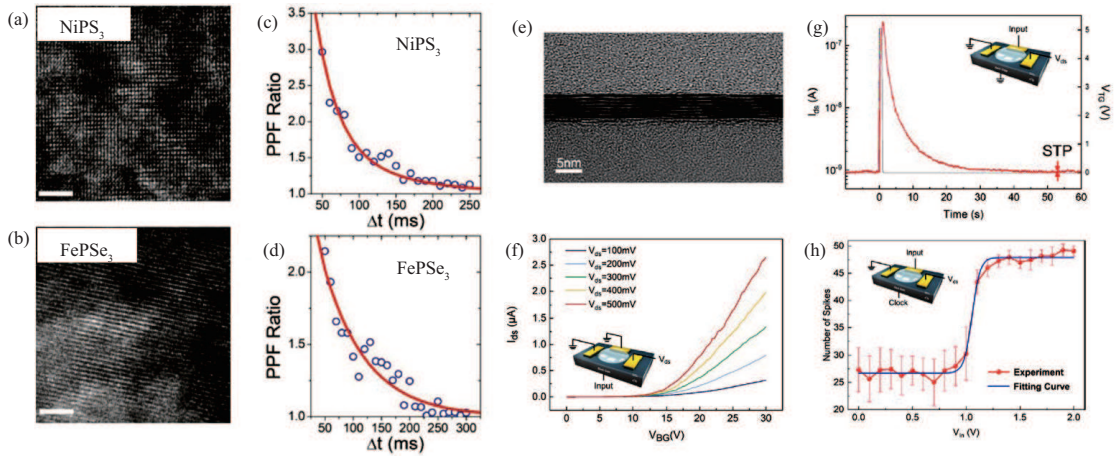


Figure 13 (Color online) (a) and (b) HRTEM of the $NiPS_3$ and $FePSe_3$ films, respectively, scale bar: 5/nm; (c) and (d) PPF ratio of the $NiPS_3$ and $FePSe_3$ based device [97] @Copyright 2018, John Wiley & Sons, Inc.; (e) HRTEM of the MoS_2 nanosheet multilayer structure; (f) I_{ds} response under different V_{ds}; (g) STP characteristic with V_{TG} applied; (h) the integrate-and-fire function under the applied of V_{ds} and V_{TG} [101] @Copyright 2019 American Chemical Society.

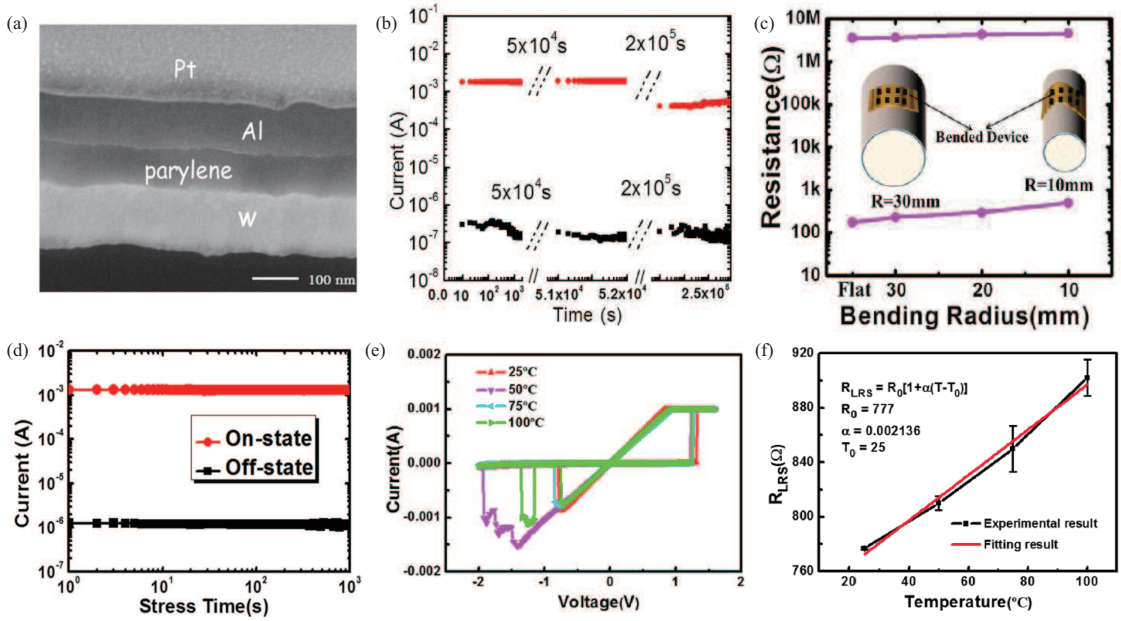


Figure 14 (Color online) (a) Cross-section SEM view of the parylene-based RRAM. (b) retention behavior; (c) bending test under various bending curvature radii; (d) stress test under continuous read pulses [107] @Copyright 2016 IOP Publishing; (e) the *I-V* characteristic under different temperatures; (f) the measurement of the LRS under different temperatures [108] @Copyright 2017 MDPI.

Cai et al. [107, 108] has conducted in-depth research on parylene-C materials based RRAM which is suitable for wearable biomedical applications. Figure 14(a) shows the cross-section SEM of the Al/parylene/W device based on a flexible substrate [107]. The remarkable retention characteristic (storage window $>10^4$), superior bending stability under various bending curvature radii and stable read disturb characteristics were performed to reveal corresponding electrical reliability of the flexible devices (Figures 14(b)–(d)).

Subsequently, Lin et al. [108] demonstrated the temperature sensing functionalities of the parylene-C based RRAM. Figure 14(e) shows the temperature-dependent *I-V* curves. The LR as the function of temperature is displayed in Figure 14(f), in which the linear relationship indicates the metallicity of the conductive filament (CF). These studies manifest the potentiality of the polymer material based RRAM in the application of wearable biomedical products.

In addition to the RS material, the electrode material utilized by the memristor will also have an

important impact on the RS mechanism and performance of the device [109]. The commonly used electrode materials include Ag, Cu, Au, Pd, Pt, Ti, Ta, Hf, TiN, Mo, W. These materials are mainly divided into three categories. Ag and Cu are active electrode materials with high mobility, which can migrate and diffuse in the RS layer under the drive of an electric field. Au, Pd, and Pt are inert electrodes with low field-induced mobility and stable chemical properties at the interface without diffusion and chemical reaction. Ti, Ta, Hf, and TiN are interface active electrodes, which has low mobility in the RS layer, but the chemical activity is strong enough to react with the RS layer to form a sub-state oxide at the interface, which will influence the resistive switching mechanism and performance of the memristor [110–112].

2.1.2 *Mechanisms*

The RS mechanisms of memristors are very complicated due to the coupled electron and ion dynamics which are sensitive to material composition, device structures, environmental conditions, and so on [111–118]. At present, the resistance change mechanisms of memristors can be classified according to several mainstream material categories. The resistance switching of PCRAM is due to the transformation of the phase change material between the crystalline and amorphous state under Joule heating [24]. The FeRAM instead relies on the reversal of the ferroelectric polarization to achieve non-volatile resistance [25]. On the other hand, the RS mechanism of RRAM can be divided into CF type and interface type [23]. The CF type refers to the resistance change phenomenon being modulated by the formation and breakage of the CFs in the RS layer. The LR is independent of the device area as the formation of CFs is localized to a small area. Furthermore, the RS mechanism of the interface-type devices is generally considered to be caused by the change of the contact barrier at the interface between the electrode and the switching material. This RS phenomenon is usually described by the Schottky emission model. As the current path of the interface-type device is distributed throughout the device, the resistance increases as the area decreases [116, 117]. However, current research on the mechanism of interface-type devices is still not mature enough. Below, we will introduce the CF type RRAM in detail.

The RS mechanisms of the memristors have been widely observed in many material systems based on MIM structure [115]. In the filament type RRAM, when the CFs connect to the two electrodes, a conductive channel is formed and the device exhibits LR. When the CFs break, the conductive channel is disconnected and the device exhibits HR. RRAM exhibits different RS characteristics with different materials. As the formation of CFs plays important role in the switching process, CF type memristors can be divided into electrochemical metallization memory (ECM) [113, 119] in which the CFs are formed by metal cations such as Ag and Cu cations, and valence change memory (VCM) [112] which is dominated by oxygen anions or oxygen vacancies.

A typical ECM is composed of an active electrode prone to chemical reactions, an inert electrode, and a sandwiched RS layer which causes initial HRS of RRAM. The active electrode will be partially oxidized into metal cations and then migrate towards the inert electrode under a positive voltage. According to the electrochemical theory, the CFs will form under the reduction of metal ions at the inert electrode, enabling the switch from HRS to LRS which is called the set process. The reverse process, called the reset process, occurs when the CFs rupture as the active metals oxidized into cations under the applied voltage [120].

Through systematic in situ TEM studies, Yang et al. [121] revealed the microscopic origin of the dynamic growth and migration processes of metal nanoclusters in ECM under the driving of electric field. Figure 15(a) shows the schematic of Ag nanoclusters based samples. Figure 15(b) displays an HRTEM of the Ag nanoclusters and the corresponding FFT result is displayed in the inset of Figure 15(b), proving the crystalline face centered cubic (fcc) structure of the Ag nanoclusters. In the applied electric field, the evolution of the Ag nanoclusters can be observed in Figure 15(c)–(f). It can be seen that the cluster gradually dissolved (Figure 15(d) and (e)) before forming another cluster in Figure 15(f), thus realizing the overall movement of Ag clusters and leaving behind a void in their original positions.

As the formation of CFs in ECM involves metal ions, the geometry of CFs depends strongly on kinetic factors such as ion mobility (μ) and the redox rates (Γ_i) of metal. There are four types of filament growth geometries corresponding to the combinations of μ which determines the nucleation site of the ions and Γ_i which influences the supply of ions. Subsequently, devices with different metals were fabricated to investigate the electric fields required to migrate the metals (Figure 16(e)). The results indicated that Ag and Cu are more acceptable for ECM. These studies demonstrated the dynamic growth of CFs and

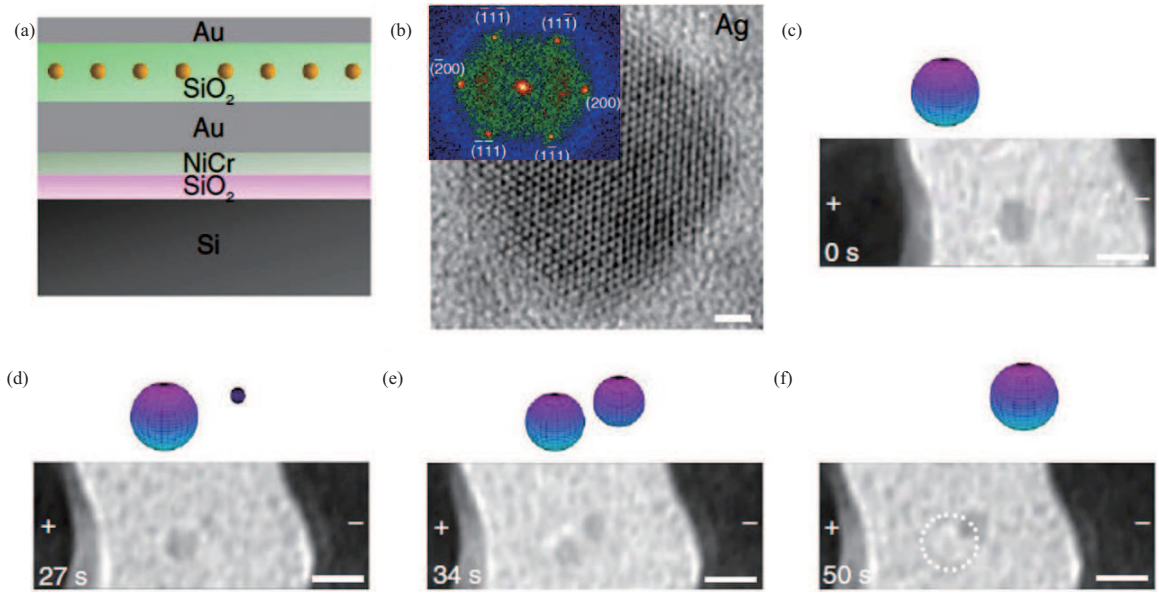


Figure 15 (Color online) (a) Schematic of the Ag nanoclusters based ECM. (b) HRTEM image of the embedded Ag nanoclusters. The inset shows corresponding FFT results of an Ag nanocluster, scale bar: 1 nm. (c)–(f) In situ TEM images after applying an electric field, scale bar: 10 nm. Ref. [121] @Copyright 2014 Springer Nature.

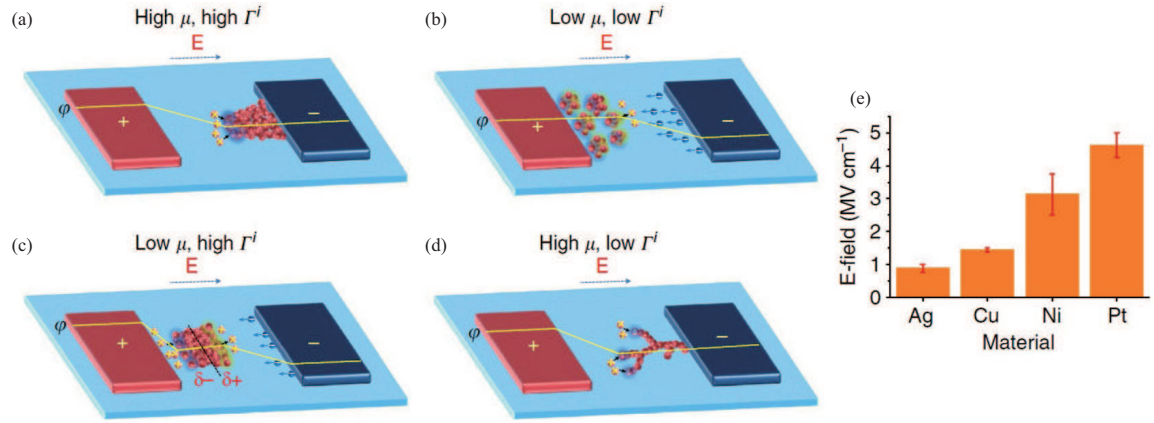


Figure 16 (Color online) (a)–(d) Four types of filament geometries corresponding to the combinations of μ and Γ_i; (e) distribution of the electric field required for the migration of different materials. Ref. [121] @Copyright 2014 Springer Nature.

gave general guidance for the design of novel RRAMs.

Similar to the ECM, the mechanism of VCM is also based on electrochemical reactions and electromigration of ions. However, the VCM mechanism is related to the oxygen-related defects (such as oxygen vacancies or oxygen ions) in the oxide. The RS based on the VCM mechanism is mainly divided into two categories [112,114]. The first type is the formation of sub-oxides under the migration and accumulation of oxygen vacancies at the interface of oxide and metal, thereby changing the Schottky barrier. The other type is the directional movement of oxygen vacancies as the electric field is applied, causing localized oxygen-deficient areas in the oxide material, which reduces the chemical valence of metal elements in the oxide and forms CFs composed of oxygen vacancies. As the movement of oxygen ions is difficult to detect, the mechanism of VCM is still under research.

Therefore, with the help of conductive atomic force microscopy (C-AFM) and electrostatic force microscopy (EFM), Yang et al. [122] gave clear evidence of the migration and accumulation of oxygen ions in HfO_2 (Figures 17(a) and (b)). Figure 17(c) displays the topographic image of the sample under different positive sweeping voltages. Besides, the detailed topographic characteristics, as well as 1ω and 2ω components under different sweeping voltages are displayed in Figures 17(d)–(l). When the sweeping voltage was increased from 6 to 10 V, RS from off to on state was realized, causing the structural damages (Figure 17(g)) and dark spots (Figure 17(i)) due to the formation of oxygen gas at the top interface [118].

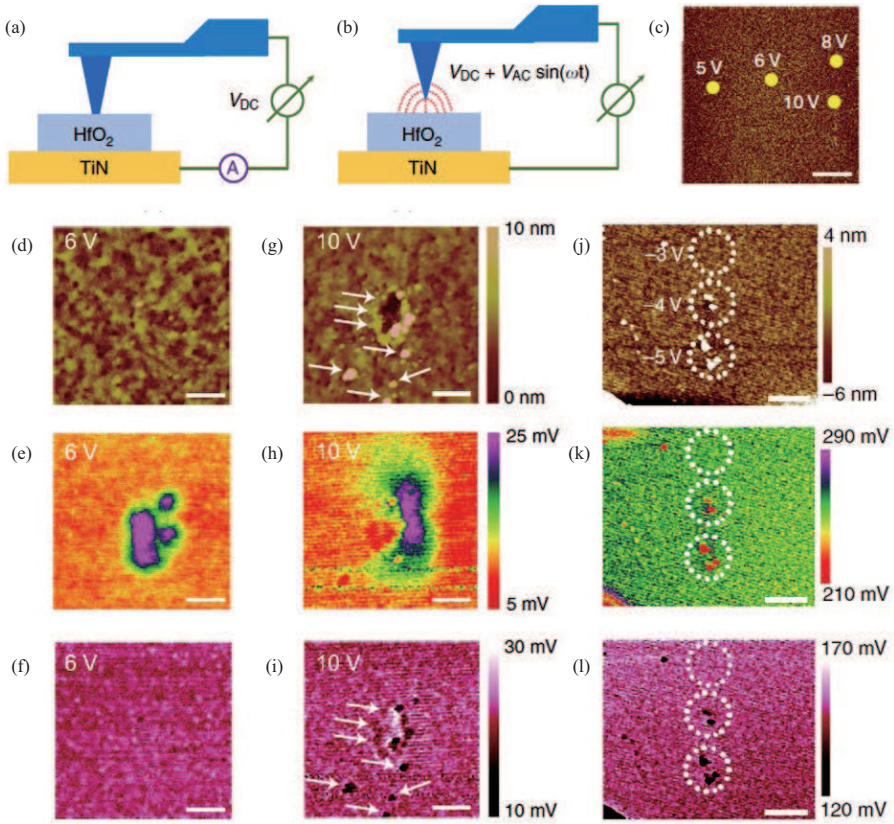


Figure 17 (Color online) (a) and (b) Schematic of the C-AFM and EFM measurements, respectively; (c) topographic image under different positive sweeping voltages, scale bar: 4 mm; (d), (g), (j) topographic; (e), (h), (k) 1ω; (f), (i), (l) 2ω measurements on the region under the stimulation of different sweeping voltages, scale bar: 200 nm. Ref. [122] @Copyright 2017 Springer Nature.

Besides the change in Figure 17(h) indicated the oxidation of the accumulated oxygen ions. Subsequently, the mechanism of VCM was further investigated under a negative sweeping voltage. As the voltage up to −4 V, the structural damages were observed in the 1ω signal, indicating a charge accumulating in HfO_2 (Figures 17(j)–(l)).

The above work explored the migration and accumulation of oxygen ions using EFM characterizations, while Liu et al. [123] explored the redox reaction of the oxygen ions by observing the surface of HfO_2 based VCM fabricated on a thermally oxidized silicon substrate with a symmetric structure. When a positive voltage sweeps across the top electrode (Figures 18(a) and (b)), obvious structural damage with three distinguishable layers can be observed in Figure 18(c). Figure 18(d) further displays the TEM of the bubble structure marked in Figure 18(c). The production of the bubbles can be explained by the formation of oxygen gas through the anodic reaction [78]. Simultaneously, there must be a cathodic reaction due to the charge neutrality [124–127]. Furthermore, the structural damage to the BE can be explained by excessive Joule heating effects [122]. To further verify this point of view, the control experiments were conducted under the applying of negative voltage on the top electrode (Figures 18(e) and (f)). As shown in Figure 18(g), the damage was only found on the top electrode at positions '2' and '3', while the structural damage at position '1' extended into the switching layer, indicating that it is not necessary for the oxidation and reduction reactions to occur at the same destructive sites.

It is worth mentioning that as the thermal effects exist in RS, the thermochemical mechanism (TCM) is one of the important mechanisms of RRAM. In this mechanism, the set process is related to the thermal decomposition of the switching layer, while the thermal melting dominates the RESET process [128,129]. Different from electric effects, the TCM is independent of polarities. And PCRAM is considered the most prominent TCM memory.

In addition to the above-mentioned mechanisms related to ion migration and transport of combined electron/ion, there are also some special RS devices based on pure electronic effects. Some impurities and defects existing in the RS material will affect the electron transport ability in the film. When they are occupied by electrons, the resistance changes. At present, the RS models related to charge traps mainly

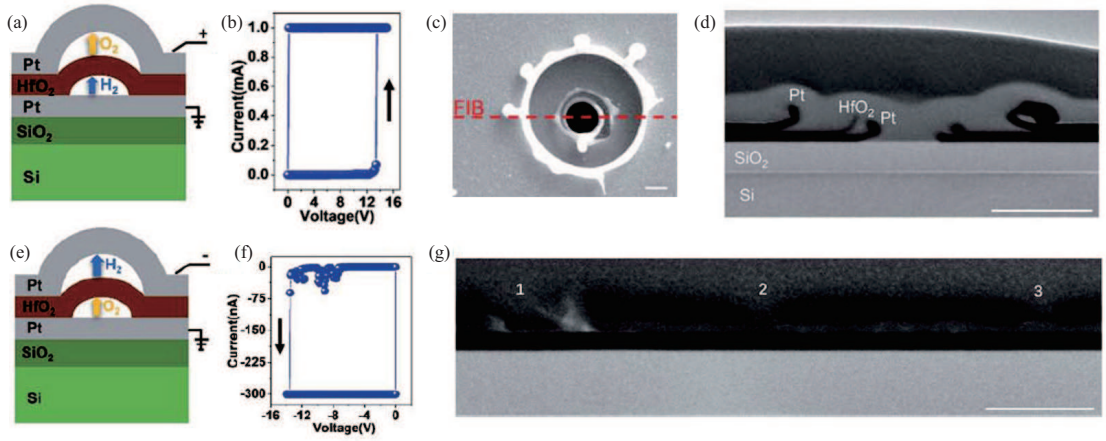


Figure 18 (Color online) (a) Schematic of the positive voltage applied; (b) positive forming characteristics of $Pt/HfO_2/Pt$ device; (c) SEM image after forming process, scale bar: 200 nm; (d) TEM image of the bubble structure, scale bar: 500 nm; (e) schematic of the negative voltage applied; (f) negative forming characteristics of $Pt/HfO_2/Pt$ device; (g) TEM image showing different structural damage, scale bar: 200 nm. Ref. [123] @Copyright 2019 Royal Society of Chemistry.

include the defect-controlled space charge limited current (SCLC) model [130–132] and the Poole-Frenkel (P-F) emission model [133–135]. Memristors based on charge trapping and releasing are usually related to charge traps. When carriers are trapped, changes in the defect energy level will lead to different resistance states of the film, which usually causes the SCLC [131]. The P-F effect is also called the field-assisted thermal ionization effect. The RS material produces large numbers of traps during the growth process or electrical activation process. The Coulomb barrier generated by these traps in the material will severely limit drift and diffusion current. At the same time, the distance between adjacent traps is relatively large, reducing the probability of the tunneling phenomenon. Therefore, the change of conductance can be achieved by controlling the electron concentration in the conduction band through charge trapping or releasing [133]. Understanding the mechanisms of memristors is critical for further performance optimizations.

2.1.3 *Integration*

The integration of memristors can be generally divided into two types, active array constructed by connecting the transistors with memristors in series and passive array formed by the memristors [136–139]. A HfO_x-based memristor cell connected to an NMOS transistor is proposed under a standard CMOS process to realize synaptic function [140] as shown in Figure 19(a). Figure 19(b) shows the LTP and LTD characteristics of the multi-terminal device under continuous positive and negative voltage stimuli, respectively. From the *I-V* characteristics of both devices (Figures 19(c) and (d)), it is clear that the transistor modulates the conductance change of the memristor cell. For the single memristor device, the SET is an abrupt process, while the memristor-based multi-terminal device shows a gradual resistance change which was attributed to the voltage-divider effect of the transistor. During the SET process, the voltage across the voltage multi-terminal device may gradually decrease as its resistance becomes smaller and smaller, which is different from the SET process in a single memristor. Thus, the memristor-based multi-terminal device has good gradual resistance tuning property without compliance current.

As the size of the transistor determines the area of memory to a large extent, which in turn limits the scalability of the integrated structure, the passive array is more dominant in terms of process and integration density. In a passive array, the memory cell is composed of a crossbar structure of mutually perpendicular word lines (WL) and bit lines (BL), which minimizes the size to $4F_2$ [141,142]. At the same time, the passive array does not depend on the front-end preparation of the CMOS process. Therefore, it is beneficial realizing 3D stacking integration to greatly improve the integration density and reducing the effective size of the storage unit to $4F_2/N$ [143, 144]. The 3D RRAM architectures can be classified into horizontal stacking 3D architecture (HRRAM) and vertical 3D architecture (VRRAM). Each layer of the HRRAM structures needs to be prepared separately, thus increasing the storage density per unit area at the expense of the production cost. On the contrary, the fabrication of VRRAM is more cost-effective [145–147]. However, in a large-scale passive array, the sneak path through unselected cells will seriously degrade the performance of the crossbar architecture, thereby decreasing the integration density

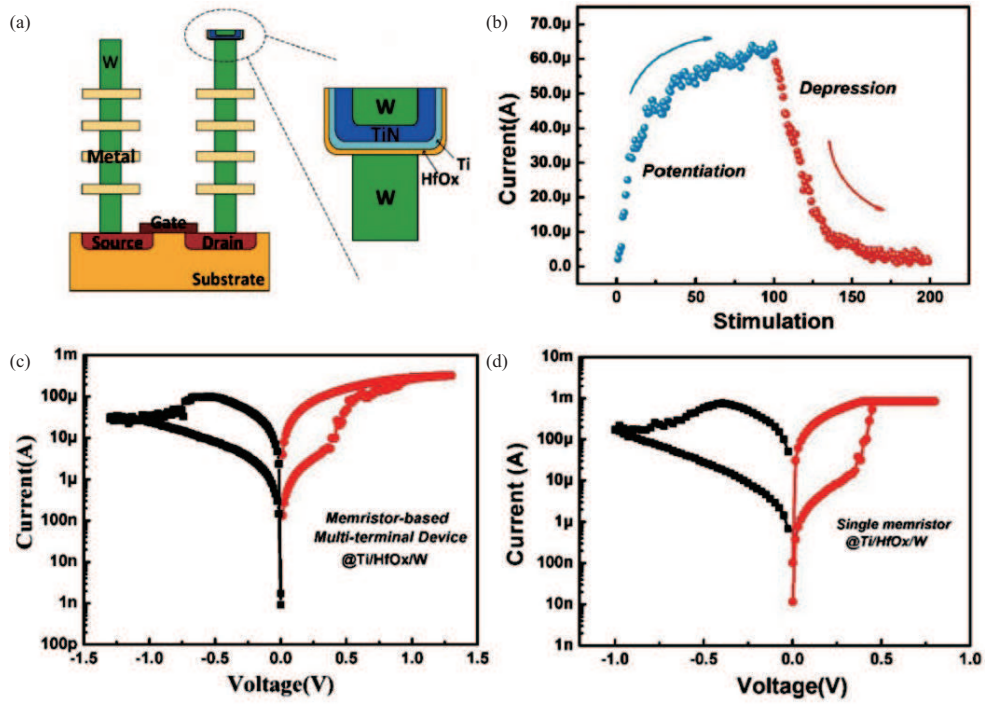
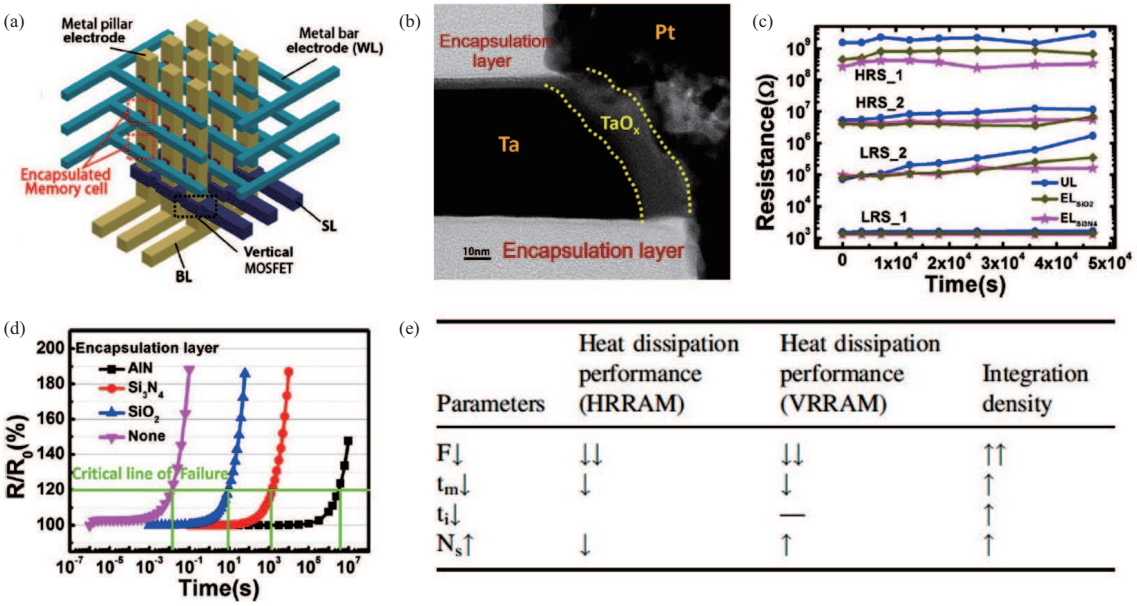


Figure 19 (Color online) (a) The fabricated device consisted of a memristor and an NMOS transistor. (b) LTP and LTD characteristics under continuous voltage stimuli. (c) and (d) *I-V* characteristic of memristor-based multi-terminal device and single memristor. Ref. [140] @Copyright 2016 The Royal Society of Chemistry.



Parameters	Heat dissipation performance (HRRAM)	Heat dissipation performance (VRRAM)	Integration density
F↓	↓↓	↓↓	↑↑
t_m↓	↓	↓	↑
t_i↓		—	↑
N_s↑	↓	↑	↑

Figure 20 (Color online) (a) The structure of encapsulated VRRAM array. (b) TEM image of the encapsulated device. (c) Retention behaviour of four weights at 200°C for the three structures. (d) Resistance degradation of devices with different encapsulated layer. Ref. [148] @Copyright 2016 IOP Publishing. (e) The influence of device parameters on the cooling performance of 3D RRAM. Ref. [151] @Copyright 2019 IOP Publishing.

and complicating the external sensing circuits. These will be discussed in Subsection 3.4.

Although there are extensive researches on 3D RRAM, the reliability issues remain to be resolved. An encapsulated vertical 3D RRAM structure (Figure 20(a)) was proposed to reduce the effect of the thermal disturbance, thus improving reliability [148]. As expected, the cross-sectional TEM image in Figure 20(b) indicates that the TaO_x RS film is wrapped by the surrounding encapsulation layer (EL). Besides, the retention test of four weights at 200°C for the three structures was conducted as shown in Figure 20(c).

Table 1 Comparison between the NOR flash and NAND flash [26, 155, 158–162]

Type	Reading speed	Programming speed	Erasing speed	Endurance	Memory technology driver	Application
NAND flash	Slow	Fast	4 ms	10^6	Needed	Large capacity; data storage
NOR flash	Fast	Slow	5 s	10^5	Not needed	Small capacity; code storage

With the ELs of Si_3N_4 ($EL_{Si_3N_4}$), the resistance has a slight change compared with the unencapsulated layer structure (UL) due to the higher oxygen blocking ability of Si_3N_4, indicating optimized reliability of the encapsulated structure [149, 150]. Figure 20(d) shows the characteristics of resistance degradation based on the 3D RRAM with different encapsulated layers. The calculation demonstrated an improved thermal disturbance immunity under the encapsulation of AlN.

Subsequently, the effect of the device parameters on the integration density of 3D RRAM was investigated by Chen et al. [151]. The finite-element based models were conducted to research the effect of four parameters (feature size (F), the thickness of metal layer (t_m) and isolation layer (t_i), and the number of stacked layers (N_s)) on the heat dissipation performance of VRRAM and HRRAM (Figure 20(e)). From the perspective of thermal effects, the VRRAM is a better choice for high-density RRAM arrays.

2.2 Charge-based memories

The charge-based memories mainly include DRAM, SRAM, and flash. A DRAM memory cell is composed of a transistor and a small capacitor. The quantity of charge stored in the capacitor is used to represent a binary bit "1" or "0". However, the leakage of the capacitor will cause an insufficient potential difference, resulting in memory corruptions. Therefore, the capacitor needs to be charged periodically [21]. The DRAM-based in-memory computing mainly utilizes the charge sharing mechanism between DRAM cells [152, 153]. On the other hand, each SRAM cell requires four to six transistors besides other components. SRAM uses latches in its core and relies on bistable circuits to store information. Therefore, compared with DRAM, the information stored in SRAM will be preserved even if there is no refresh operation, provided the power is uninterrupted [22]. SRAM can be used to implement binary multiply-accumulate (MAC) operation (XNOR accumulation operation) thanks to its binary property. This makes implementing binary neural networks using SRAMs possible [154, 155].

Compared with SRAM and DRAM, flash is non-volatile. It also has the capability of erasing information as a whole and reprogramming by byte. Flash is divided into NOR-flash and NAND-flash according to the array structure and operations [155]. NOR structure usually adopts channel hot electron injection as for main programming method. As shown in Figure 21(a), as the charge-storage layer has a high density of electron traps, the electrons can be trapped in the floating gate (FG) to increase the $V_{\rm th}$ [156]. The TEM images in Figure 21(b) and (c) show the BL and WL cross-sections of the NOR flash cells. As for erasing process, electrons tunnel from the FG to the source region or channel via Fowler-Nordheim (FN) tunneling mechanism [157]. The multi-bit storage of the NOR flash cell was realized by designing the drain voltage and gate voltage (Figure 21(d)). And the error rate of each of the 16 states is displayed in Figure 21(e).

In NAND-flash, information is programmed by the tunneling of electrons from the channel into the FG [158, 159]. Due to the high transmission efficiency of NOR-flash, it is cost-effective in small capacity storage, but the limited programming and erasing speeds greatly affect its performance. NAND-flash has the advantages of large capacity, fast reprogramming speed, etc., which is suitable for storing large amounts of information. And the comparison between the NOR flash and NAND flash is displayed in Table 1 [26, 155, 158–162]. Furthermore, as the FG transistor is an active component, flash memory has an additional advantage over RRAM in the consideration of leakage current [160]. The nanoscale flash memory array has the ability to realize VMM, subsequently improving the computing efficiency in fully connected neural networks [161–164].

A benchmark of the different memories is demonstrated in Table 2 [64, 97, 165–172], and the energy and switching speed in which are the best performance indicators currently reported. It can be seen that PCRAM has the most mature manufacturing process, but there are also some shortcomings such as asymmetric switching, conductance drift, and thermal disturbances. And the flash faces some challenges such as performance, power consumption, and scalability. RRAM is a hot topic of current research due to its excellent performance indicators, but the variation caused by the mechanism based on the filament growth and rupture is a problem we have to confront.

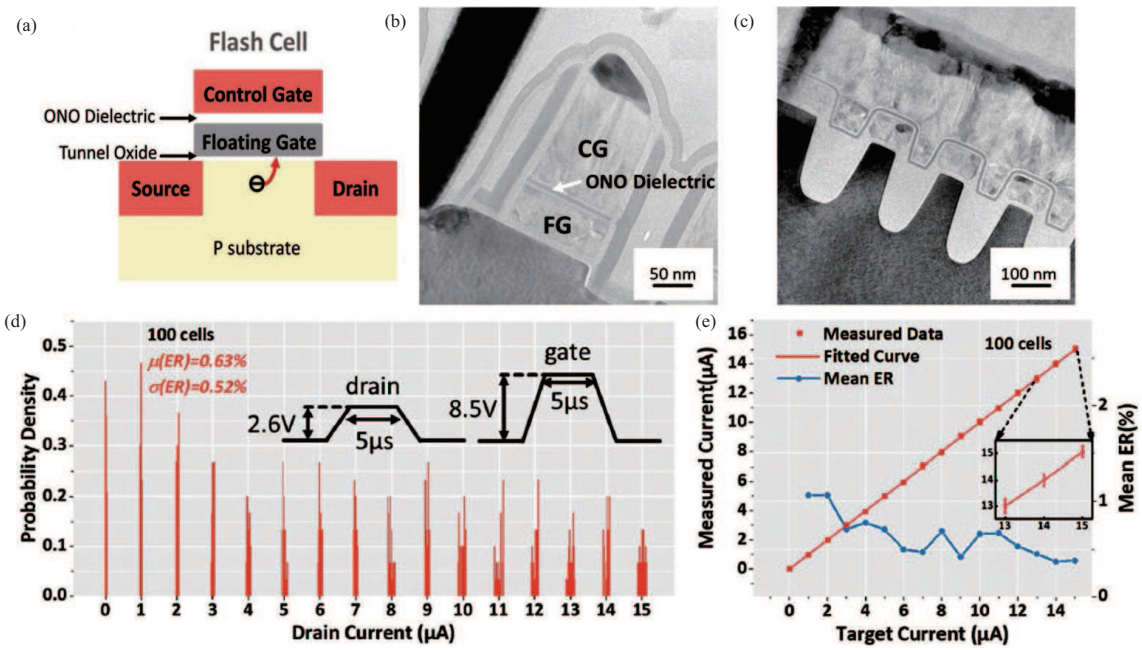


Figure 21 (Color online) (a) Schematic of the nanoscale flash. (b) and (c) BL and WL TEM of the nanoscale flash. (d) I_{ds} distribution of 16 states. (e) The comparison of measured and target I_{ds}. Ref. [156] @Copyright 2019 John Wiley & Sons, Inc.

Table 2 Comparison of different types of resistance-based memories [64, 97, 165–172]

Device	Mechanism	Material	Energy	Switching speed	Cons
Oxide-RRAM	Filament growth and rupture	HfO_x; TaO_x; SiO_x	0.1 pJ	100 ps	Variation stochasticity
PCRAM	Amorphous-crystalline phase change	$Ge_2Sb_2Te_5$ (GST)	0.1 pJ	300 ps	Asymmetric switching; conductance drift; thermal disturbances
FeRAM	Polarization switching	Doped HfO_x; perovskite	1 fJ/bit	10 ns	Hard to realize multilevel
2D-RRAM	Defects/ions migration	MoS_2; graphene; WSe_2	30 fJ	5 ns	Tough to infer the scalability
Polymer-RRAM	Filament growth and rupture	parylene-C; PMMA; PEO	1.23 fJ	15 ns	Variation stochasticity
Flash	Hot electron injection FN tunneling	SiO_2; SiN_x	0.9 PJ/bit	100 μs	Limited endurance; low storage density

3 Device and array optimizations for in-memory computing

The high-performance memristor is a basic element for in-memory computing in high-density arrays. Compared with memristors of other material systems, binary metal oxides have received more attention because of their simple preparation process and compatibility with traditional CMOS processes. In the application of in-memory computing, the memristors with high weight precision are required for deducing, as the uniformity and linearity are the critical indexes for training. And as mentioned above, the sneak path is an obstacle to high-density integration. In addition, the operation speed, power consumption, endurance and retention, and so on are important for in-memory computing. In the current research, on the filament-type memristor, it is generally believed that the randomness of the generation and fracture of CFs in the resistive function layer are the source of the discreteness of the resistance transition parameters [173]. Therefore, suppressing the random generation of CFs can effectively improve the performance of the device. This subsection mainly introduces the device and array optimizations for in-memory computing.

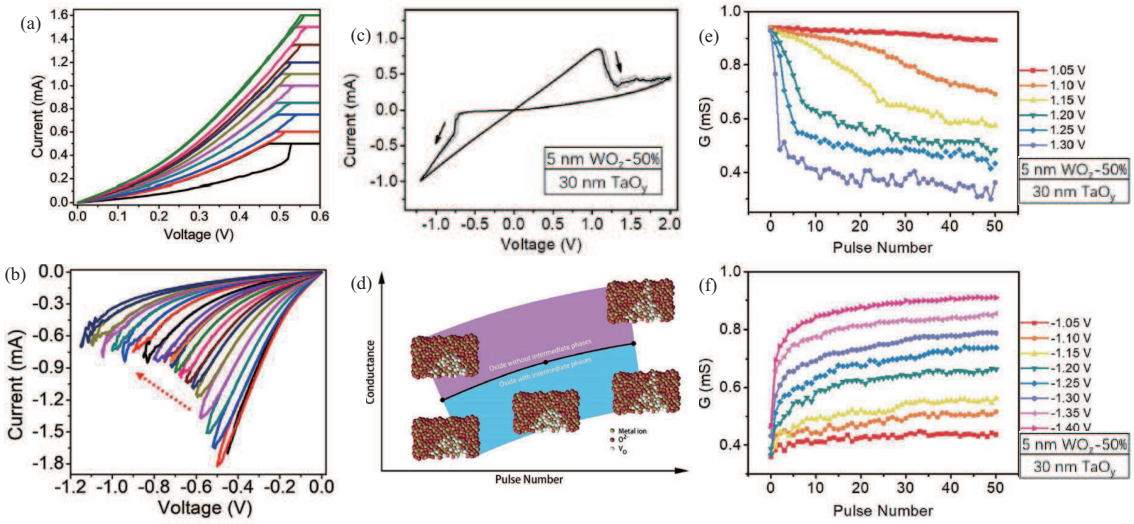


Figure 22 (Color online) (a) Gradual potentiation under a series of positive voltage sweeps. (b) Gradual depression under a series of negative voltage sweeps. Ref. [53] @Copyright 2018 AIP Publishing. (c) *I-V* sweeps of device based on Pt/Ti/WO_z-50%/TaO_y/Pt. (d) Different geometry of filaments in different switching layers. (e) and (f) LTD and LTP characteristics of the device. Ref. [177] @Copyright 2017 The Royal Society of Chemistry.

3.1 Weight precision

Having more weight states per cell is a highly desired characteristic in the application of in-memory computing. It directly increases the storage capacity of the cell without sacrificing additional layout area. RRAM has the ability to demonstrate multilevel conductance (MLC) by altering the compliance current of the set process and the stop voltage of the reset process, as shown in Figures 22(a) and (b) [53, 174]. The above-mentioned work by Liu et al. [52] realized a multilevel (4-bits) device by doping Zr into HfO_2. Moreover, the abundant sub-oxide phases in the switching layer material will be beneficial in realizing the MLC. This is because the filaments with different geometries will be formed during the programming process [175, 176]. Based on the theory, Li et al. [177] fabricated a WO_x based RRAM with rich intermediate phases to realize multi-states of conductance. Figure 22(a) exhibits typical bipolar switching in Pt/Ti/WO_z-50%/TaO_y/Pt devices, which depends on the gradient of oxygen concentration in the bilayer oxides. Figures 22(e) and (f) display the potentiation and depression processes in the devices, showing the MLC of the device. The different geometry of filaments in the switching layers with and without intermediate phases are displayed in Figure 22(d), which indicates that the MLC is achieved through stable changes in filament geometry.

3.2 Uniformity

Variability is the main challenge for RRAM mass production [178]. The variability is mainly reflected in the wide distribution of switching voltage and resistance states. It can be divided into CCV and DDV according to time and space fluctuations. These random effects will erroneously program/erase the memory cell, or misidentify the memory state, thereby posing challenges on peripheral sensing and programming circuits. Since the poor uniformity is caused by the random formation and dissolution of the CFs, the variability of the RRAM can be alleviated by reducing the inherent stochasticity of the filament size and the effective switching position.

The work of Yu et al. [148] mentioned above improved the reliability of the encapsulated 3D RRAM by suppressing unnecessary V_o migration. Moreover, Liu et al. [52] fabricated an interface-type RRAM based on Zr-doped HfO_2 to exhibit excellent uniformity of the device. In addition, the uniformity of the HfO_x based RRAM was improved by Fang et al. [179] with the insertion of a TiN buffer layer. As shown in Figure 23(a), two types of samples with or without a 2 nm-TiN were fabricated in the same processes. The box plots in Figure 23(b) present the distribution of initial resistances and forming voltages. Furthermore, the pulse cumulative curves of HRS and LRS are displayed in Figure 23(c), which demonstrate that the RRAM with TiN buffer layer has a more stable HRS. In addition, the CCV and DDV were characterized by normalized quantities calculated using two different sets of methods, and

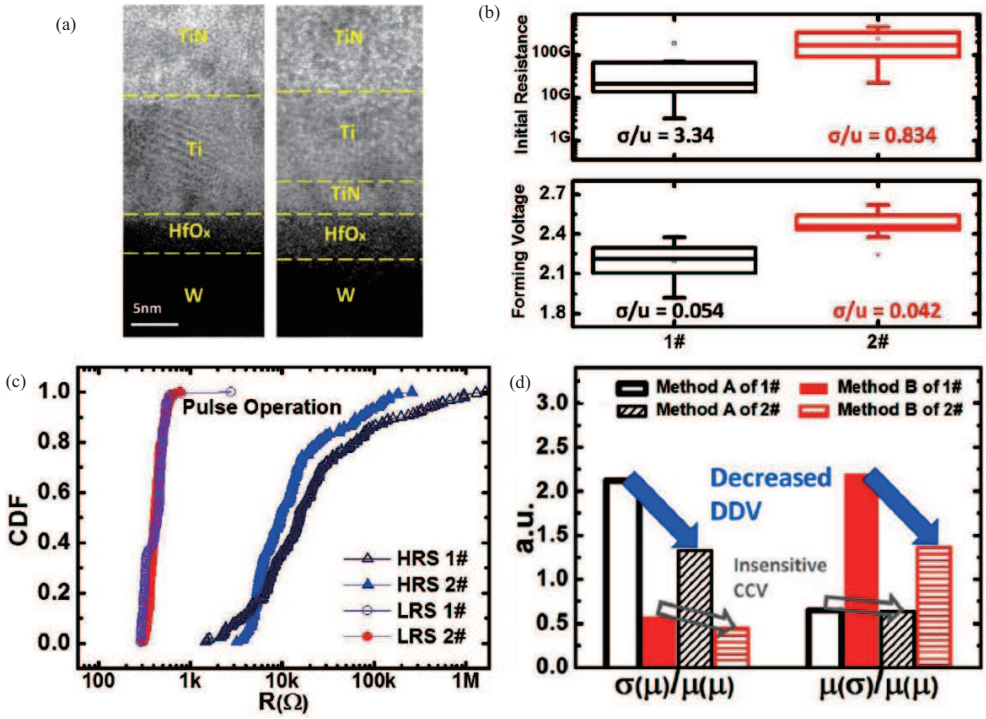


Figure 23 (Color online) (a) HRTEM images of samples; (b) the box plots of initial resistance and forming voltage; (c) the distribution of HRS and LRS; (d) the bar plots of the DDV and CCV on HRS. Ref. [179] @Copyright 2018 IEEE.

the results are as shown in Figure 23(d) [180], in which the improvement of DDV is more obvious than CCV. These optimizations imply that the TiN not only reduces the side effects of the Ti layer but also maintains its deoxidizing ability.

3.3 Linearity

The linearity of the conductance in a memristor refers to the uniformity of the conductance change under the same electrical pulse [180, 181]. Linear conductance modulation is important in the application of in-memory computing, as it simplifies the weight adjustment process and peripheral circuit to a great extent [32].

The linearity of TaO_x based memristors is optimized via the introduction of an ion diffusion limiting layer (DLL) by Wang et al. [182]. A typical nonlinear LTP and LTD are displayed in Figure 24(b), both of which can be divided into two regimes including the initial abrupt increase (decrease) regime and the following gradual modulation regime. It can be explained as follows. Initially, the growth of CF is caused only by the movement of oxygen vacancies around it. And after these adjacent ions were exhausted, the vacancies originally far away from CF will then participate. The stable growth/dissolution of CF in regime II is achieved through a longer travel distance and a lower temperature gradient. The nonlinearity can be improved by inserting a 1 nm SiO_2 DLL to limit the ion diffusion speed (Figure 24(c)). The STDP behavior was realized through carefully designed spikes (Figure 24(d)) [49–51].

3.4 Sneak path

The sneak current issue in passive crossbar arrays will limit the array density and hinder the application of the crossbar structure in in-memory computing, as it will interfere with the read operation and cause incorrect programming [183, 184]. Many methods have been adopted to solve this problem. For example, additional selective devices can be connected in series with the RRAM, and RRAM cells with self-selection functions can be chosen to form the array. As the selective device will reduce the integration density and complicate the fabrication process, self-selected RRAM is highly favorable [185, 186].

An RS device with inherent nonlinear characteristics can effectively suppress sneak current in a passive array [187]. A TaO_x-based bipolar RS device with an ultra-thin SiO_{2-x} interfacial layer was fabricated by Wang et al. [188] as shown in Figure 25(a). Figure 25(b) shows the I-V characteristics of the device after

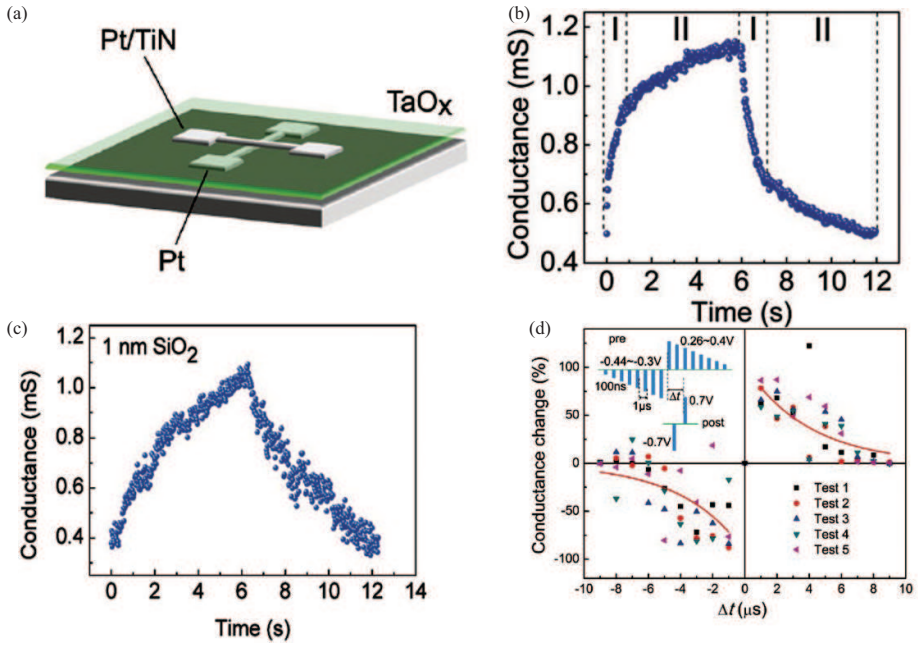


Figure 24 (Color online) (a) The structure of the device; (b) a typical nonlinear LTP and LTD; (c) linear LTP and LTD with 1 nm SiO_2 inserted; (d) STDP characteristic with the inset showing the applied spikes. Ref. [182] @Copyright 2016 The Royal Society of Chemistry.

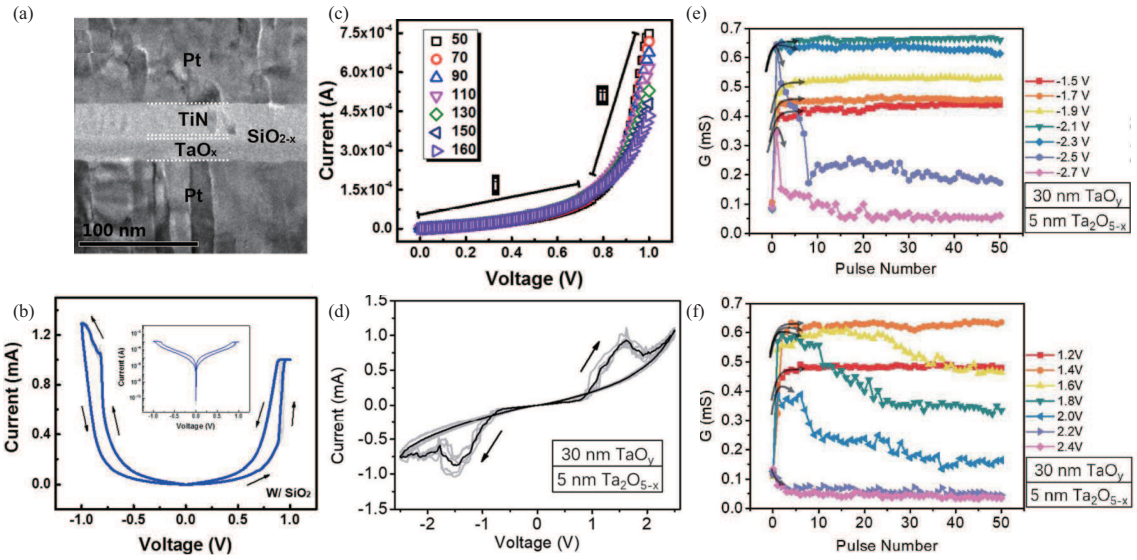


Figure 25 (Color online) (a) SEM image of the SiO_{2-x} inserted devices. (b) *I-V* characteristics of the interface-engineered device. (c) Temperature dependence of the device. Ref. [188] @Copyright 2016 IOP Publishing. (d) CRS characteristics of the bilayer device. (e) and (f) LTD and LTD of the CRS devices. Ref. [177] @Copyright 2017 The Royal Society of Chemistry.

the forming operation and the inset shows the switching behavior in a logarithmic scale, which displays a nearly symmetrical nonlinearity of ≈12 under the adoption of the V/3 method [187]. Figure 25(c) displays the evolution of LR as a function of temperature which also demonstrated the nonlinearity of the LR. Complementary resistive switches (CRS) composed of two anti-serially connected bipolar memory cells also can solve the sneak path problem [189]. Li et al. [177] fabricated a TaO_y/Ta_2O_{5-x} based device which displayed CRS behavior (Figure 25(d)). As the oxygen-rich Ta_2O_{5-x} film was deposited before the oxygen-deficient TaO_y film without contacting Ti, therefore, the V_Os is possible to get depleted, leading to the observed CRS behavior. Furthermore, Figures 25(e) and (f) display the analog conductance modulation of the CRS device, in which the conductance increases before decreases. This is consistent with the characteristic of CRS.

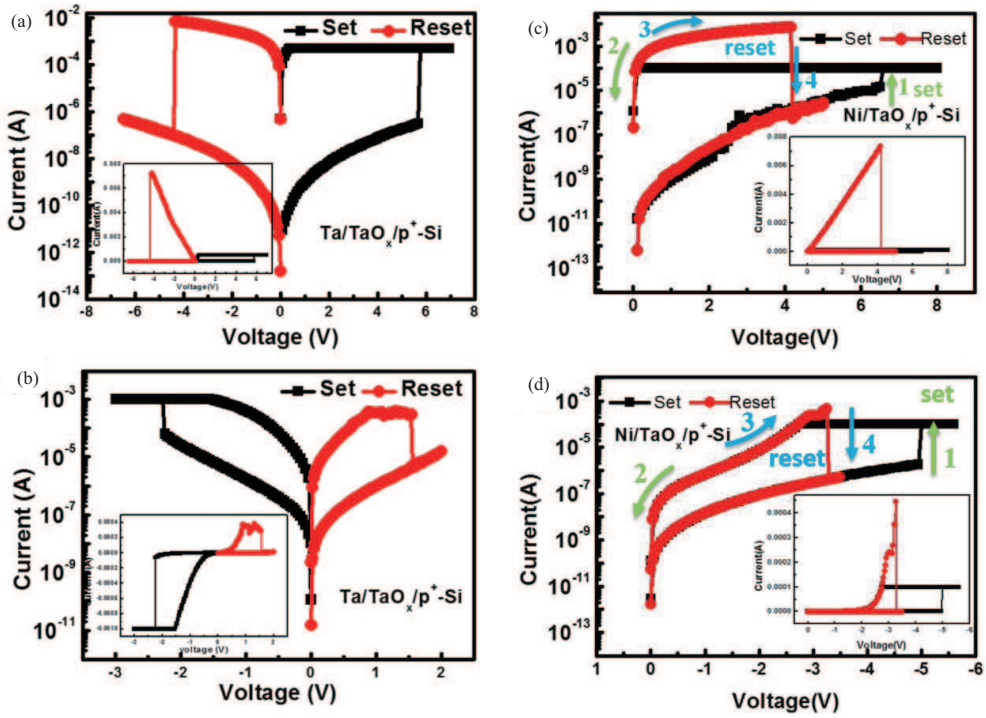


Figure 26 (Color online) (a) and (b) Bipolar characteristics of Ta electrode based device under different set voltages. (c) and (d) Unipolar characteristics of Ni electrode based device under different set voltages. Ref. [191] @Copyright 2016 AIP Publishing.

RRAM cell with self-selection function is also a solution for the sneak path problem [190]. Yu et al. [191] proposed a TaO_x/p^+-Si RRAM device with a self-selection capability. There were two kinds of devices fabricated, one with a Ta-TE while the other with a Ni-TE. The devices presented different characteristics under different operating voltages. The bipolar performances of the Ta-TE device are displayed in Figures 26(a) and (b), and the non-linearity of the device achieved 10^3 under the applying of negative switching bias. On the other hand, the Ni-TE device displayed unipolar characteristics instead (Figures 26(c) and (d)) in accordance with the oxygen-affinity comparison. Similar to the Ta-TE device, the Ni-TE device exhibited non-linear RS under a negative bias.

Furthermore, a novel resistive device based on $TiN/VO_2/HfO_2/TiN$ can switch between volatile threshold switching (VTS) and non-volatile RS (NVRS) [192]. As VO_2 exhibits the insulator-metal-transition (IMT) under thermal or electrical stimulation, while HfO_2 has excellent non-volatile characteristics [193], the dual functional layers VO_2/HfO_2 can demonstrate multi-functional characteristics under a proper voltage as shown in Figures 27(a) and (b). Figure 27(c) shows a tight distribution of $V_{\rm th}$, $V_{\rm hold}$, $V_{\rm reset}$ and $V_{\rm set}$ extracted from 50 consecutive cycles. The read margin calculated according to the array size is displayed in Figure 27(d). Besides, when the device shrinks to 40 nm node, the integration level of the array can reach $10^3 \times 10^3$.

3.5 Optimization on other aspects

In addition to the optimizations of linearity, multilevel, uniformity, and suppression of the sneak path mentioned above, the power consumption, on-off ratio, retention, and endurance of the device are also important for in-memory calculations. Wang et al. [192] also investigated the response speed of the VO_2/HfO_2 based device in VTS mode. And the switching/recovery speed can reach 30 ns, which is faster than the Ag/Cu-based VTS devices [194]. Moreover, Chen et al. [195] fabricated a graphene integrated parylene-based flexible memristor to attain low power consumption. Figures 28(c) and (d) display the comparison of reset current and power consumption between memristor with graphene (G-memristor) and memristor without graphene (NG-memristor), which indicates that the power consumption of G-memristors was nearly 14 times lower than that of NG-memristors. Furthermore, the ion gate device prepared by Zhu et al. [97] mentioned above greatly reduces the power consumption of the device by using a coupling of 2D materials and ion glue.

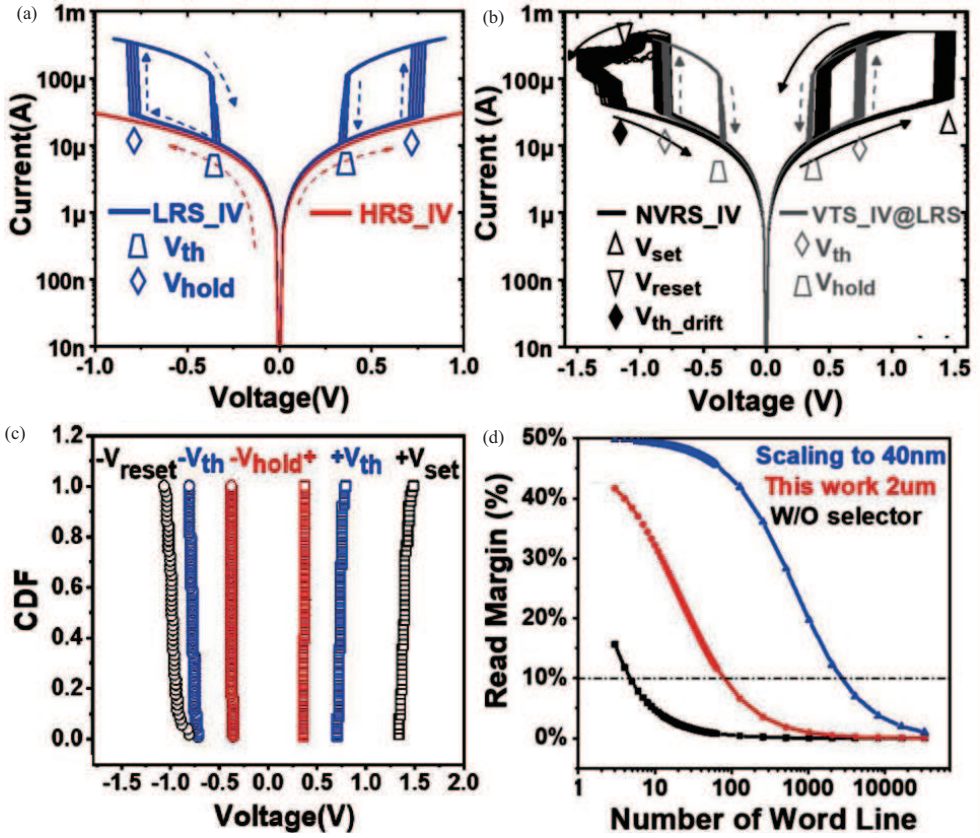


Figure 27 (Color online) (a) and (b) Typical *I*-*V* curve of VTS and NVRS; (c) the distribution of $V_{\rm th}$, $V_{\rm hold}$, $V_{\rm reset}$ and $V_{\rm set}$; (d) the read margin calculated as a function of array size. Ref. [192] @Copyright 2020 IEEE.

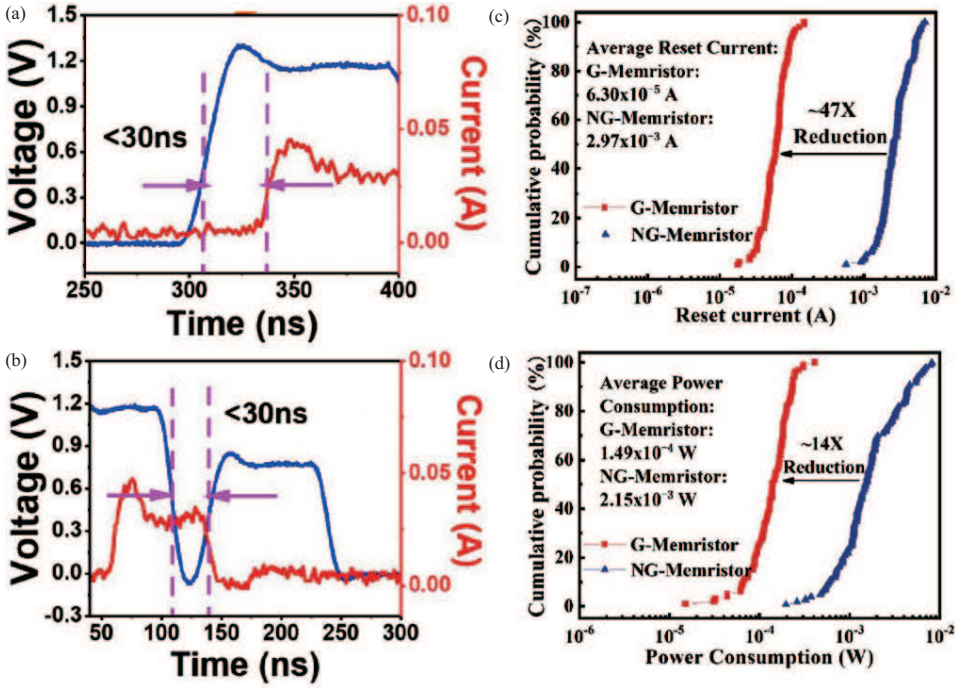


Figure 28 (Color online) (a) and (b) The measurement of switching speed and recovery speed in VTS mode. Ref. [192] @Copyright 2020 IEEE. (c) and (d) Distribution of the reset current and the power consumption. Ref. [195] @Copyright 2019 John Wiley & Sons, Inc.

In fact, the characteristics of on-off ratio, retention, and endurance have also improved in the optimizations mentioned above. This is because device optimizations tend to enhance overall performance instead of improving only certain aspects.

4 Applications based on emerging nonvolatile memory devices

The chip technology of in-memory computing aims to transform the traditional computing-centric architecture into a data-centric architecture. It directly uses memory for data processing to completely eliminate the von Neumann bottleneck. This is especially suitable for large-scale parallel computing tasks such as deep learning neural networks. In 2016, the RRAM was proposed to build a deep learning neural network given the in-memory computing architecture, which has received extensive attention [196]. Test results demonstrate that the network can lower the power consumption by about 20 times and increase speed by about 50 times compared to traditional solutions based on the von Neumann architecture. Memristor has multi-bit characteristics and can realize vector-matrix multiplication based on similar principles. The in-memory computing based on memristors can be divided into several aspects: logic operations using binary memristors, analog computing using analog memristors and other types of in-memory computing [14, 15, 197]. This chapter mainly introduces the application of in-memory computing in four aspects, including artificial neural network (ANN), spiking neural network (SNN), logic in memory and hardware security.

4.1 Artificial neural network

ANN is utilized to simulate the structure and function of a biological neural network and has been applied successfully in a broad range of signal-processing problems [198–200]. Since the ANNs can process information in parallel, the tasks such as classification, pattern recognition, and decision-making can be carried out [201–203]. The ANNs can be divided into two basic classes based on the types of interconnections: feed-forward ANNs where information flows one-way from the input to output, and recurrent neural networks (RNNs) where part of the information flows back. Perceptron and convolutional neural network (CNN) belong to the feed-forward ANNs [204, 205]. Next, we will introduce the fully connected network, CNN and RNN in detail.

4.1.1 *Fully connected network*

The fully connected network is the most fundamental feed-forward neural network forming the basis of many ANNs and, in its vanilla form, constitutes the classification layers in other networks. Below, we summarize a few studies on perceptron implemented using a variety of devices [201].

Using the HZO interface-type RS synapse, Lu et al. [52] simulated a 3-layer perceptron in recognizing MNIST handwritten digits (Figure 29(a)). This network was trained using the sparsified back-propagation (SBP) algorithm, which, due to its selective weight update strategy, improved the speed-energy efficiency (Figure 29(b)). Coupled with the exceptional uniformity characteristic of the non-filamentary switching device, this network maintained a high recognition accuracy. The algorithm also helped in distributing high updating rates throughout the crossbar (Figure 29(c)), subsequently preventing excessive degradation of the devices. Yang et al. [62] utilized 3-terminal TaO_x synapses in a single-layer perceptron simulation. Since the third terminal was able to regulate the growth dynamics of the conducting filament, it provided control over the learning rate of the entire neural network (Figure 29(d)). This feature augmented the functionalities of the synapse in addition to the usual multiply-accumulate (MAC) acceleration of crossbar arrays. The ability to control the learning rate is crucial as a gradually decreasing learning rate was shown to improve network performance (Figure 29(e)), in this case, the MNIST recognition task.

Yao et al. [206] presented an optimized synapse based on the $TaO_x/HfAl_yO_x$ material stack, in which the $HfAl_yO_x$ switching layer was essentially a HfO_x/AlO_y laminate structure. The 1T1R synapse exhibited improved analog switching behavior which was a huge step towards realizing neural networks. The authors implemented a single layer perceptron in classifying faces of 3 persons from the Yale Face Database, trained using batch learning rules (Figure 29(f)). Compared to the non-write-verify scheme during training, the write-verify scheme offered better converging speed, energy efficiency, and accuracy

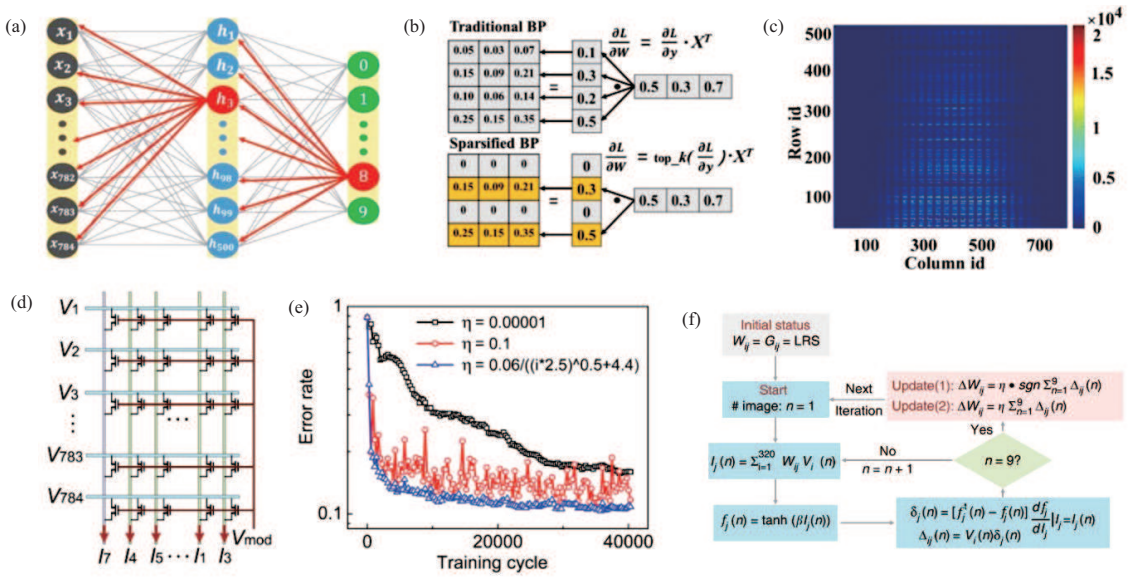


Figure 29 (Color online) (a) A schematic of the 3-layer fully-connected perceptron network. (b) An example of the SBP algorithm versus the traditional BP algorithm. (c) The distribution of the weight updating rate. Ref. [52] @Copyright 2020 John Wiley and Sons Inc. (d) A schematic of the crossbar with a modulatory terminal. (e) The error rate in MNIST classification with 3 different learning rates. Ref. [62] @Copyright 2017 John Wiley & Sons, Inc. (f) The training flowchart for the face classification of 3 persons (a total of 9 input faces) under either the Delta rule or the Manhattan rule. Ref. [206] @Copyright 2017 Springer Nature.

at the expense of a more complex control circuit. To the best of our knowledge, this hardware-based system was one of the earliest studies in physically realizing a neural network.

Li et al. [207] explored in situ training in their HfO_2 memristor crossbar array by stochastic gradient descent (SGD) instead. Their approach involved reading out the current of each column in the crossbar after a multiply-accumulate operation and then performing signal conversions, activations, and the error backpropagation in software based on the measurements (Figure 30(a)). The synaptic weights were then physically updated in the crossbar. This training paradigm allows the network to adapt to the non-ideal hardware, crucial for improved performance (Figure 30(b)). The authors demonstrated MNIST digit classification using a 2-layer fully connected network and the proposed training paradigm. Despite achieving an accuracy of only 91.71%, simulations suggested substantial improvement using a larger network. This paved the way for realizing large memristor-based neural networks.

To improve classification accuracies under in situ training, Ambrogio et al. [208] proposed combining non-volatile PCM with volatile capacitive storage within a synaptic cell (Figure 30(c)). By applying different scale factors on the read current of the cell, the PCM pair was assigned a higher significance and the capacitive storage otherwise, thus increasing the dynamic range. Only the lower significance pair was updated during training. Transferring of weights from the lower significance pair to the higher significance pair was done occasionally and periodically. Overall, this method improved the linearity and symmetry of weight updates. Besides, the authors tackled variations arising from fixed device asymmetry using polarity inversion, that is by alternating between charging (discharging) and discharging (charging) of the capacitor to represent weight increase (decrease) between transfers. They further demonstrated MNIST and MNIST-backrand classification tasks on 3-layer fully connected networks (or 4-layer, depending on the definition of a layer). They also performed CIFAR-10 and CIFAR-100 classification tasks using transfer learning, retraining only the fully connected layer of a pre-trained CNN. The accuracies were similar to software-based networks. All operations were simulated, except for the reading and tuning of each PCM conductance. Nevertheless, the authors envisioned a full hardware implementation in the future.

A rectified memristor equipped with both VTS and multilevel non-volatile switching capability can act as a filter similar to the ReLU activation function. Wang et al. [209] simulated a 3-layer self-activation neural network (SANN) wherein the crossbar array simultaneously performed the activation operation and the MAC operation (Figure 30(d)), further reducing the area and energy consumption of the CMOS activation circuits. The training process needed modification such as adding a bias for the activation function because the order between the MAC operation and the activation operation was effectively

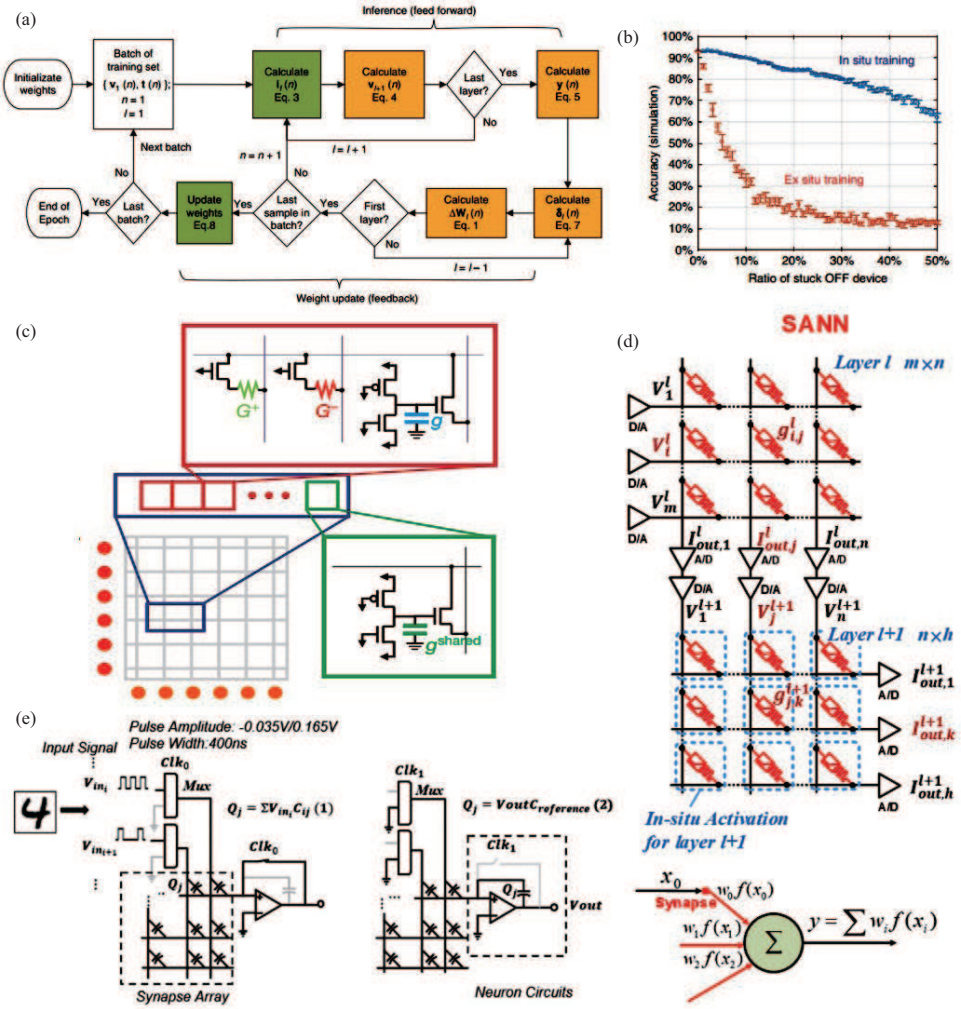


Figure 30 (Color online) (a) The in situ training flowchart. Green boxes were done in the crossbar array, while the orange boxes were done in software but could be realized in circuits integrated together with the crossbar according to [207]. (b) In situ training was proven to be more defect-tolerant than ex situ training. Ref. [207] @Copyright 2018 Springer Nature. (c) A schematic of the synapse combining PCM (G^+, G^-) and capacitive storage (g, g^{shared}), in which g^{shared} is shared between multiple synapses in the same row. Ref. [208] @Copyright 2018 Springer Nature. (d) A schematic of the SANN with an illustration of the working of a single neuron. Ref. [209] @Copyright 2020 IEEE. (e) The schematic of the ferroelectric capacitors based network in the sampling phase and the computing phase. Ref. [210] @Copyright 2019 IEEE.

reversed. Nevertheless, this architecture attained an accuracy comparable to the conventional 1T1R crossbar with CMOS activation modules on the MNIST database.

Apart from RS devices, Zheng et al. [210] simulated a single-layer perceptron based on ferroelectric capacitor synapses in a switched capacitor circuit (Figure 30(e)). Despite achieving only 78.28% accuracy on the MNIST recognition task due to the lack of optimizations, the minimal static power consumption of capacitor-based circuits is very desirable for energy-efficient systems.

It is worth noting that several non-idealities in defect-based (vacancies) RRAM devices such as time-dependent variability (TDV) [211] and early-stage resistance fluctuations (ERF) [212], and their impact on pattern recognition of the RRAM-based neural networks has been discussed. Results showed that network performance can degrade seriously; hence future network implementations based on these devices, for instance, in [62, 209] should consider the proposed mitigation strategies in [211, 212]. Although these non-idealities were studied using multi-layer perceptron, they are intrinsic to the devices themselves and thus should be taken into account in other networks.

4.1.2 *Convolutional neural network*

CNNs have found a niche in image processing for computer vision. However, some fields, such as the self-driving car industry, demand stringent requirements on the efficiency of computer vision systems.

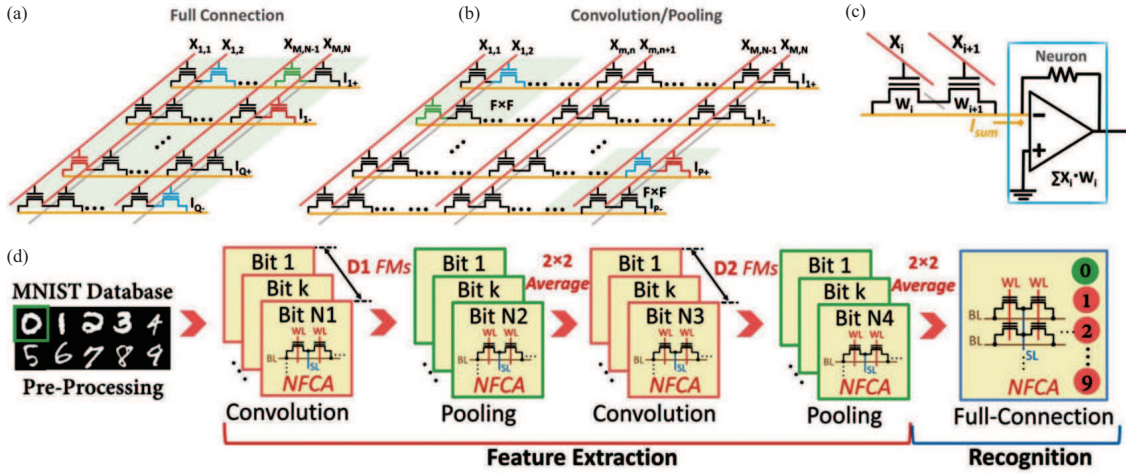


Figure 31 (Color online) (a) and (b) Schematics of the fully-connected layer and the convolution/pooling layer, both having a same array structure. (c) A neuron implemented using a transimpedance amplifier. (d) The NFCA-CNN for MNIST classification. Ref. [156] @Copyright 2019 John Wiley & Sons, Inc.

Hence, implementing CNNs using the in-memory computing paradigm is extremely important [202, 203].

Xiang et al. [156] proposed a universal and reconfigurable in-memory computing paradigm based on NOR flash computing arrays (NFCA). Given the fact that convolution, pooling, and fully-connected operations are essentially based on the same VMM [213], the same array structure can be used for each operation (Figures 31(a) and (b)), with convolution and pooling requiring only the shift-mapping of kernels. Therefore, by cascading the same array structure (Figure 31(d)), the authors demonstrated a 5-layer CNN using SPICE simulation and achieved 97.8% accuracy on the MNIST database, which is comparable to the CPU-trained counterpart.

Wang et al. [214] utilized a crossbar with TaO_x memristors in the 1T1R structure to construct their CNN. In their approach, convolution kernels were unrolled into respective 1D vectors and positioned side-by-side in the crossbar array, allowing for parallel multiply-accumulate operations between a single input and multiple kernels. The shifting of kernels over the input was done simply by changing the input vectors being fed into the crossbar array. The training process was carried out in situ to accommodate for hardware non-idealities. The authors realized a 5-layer CNN for recognizing MNIST digits. In addition, they demonstrated a convolutional long short-term memory (LSTM) network for classifying MNIST digit sequences, showcasing the possibility of sharing weights spatially and temporally.

Using multiple 2048-cell crossbars of material stack TaO_x/HfO_x memristors and a similar method of flattening and arranging convolution kernels (Figure 32(a)), Yao et al. [215] demonstrated MNIST digit recognition in a 5-layer CNN and achieved an accuracy of more than 96%. They proposed a different training scheme to overcome hardware non-idealities (Figures 32(b) and (c)). The training was performed ex-situ in software before mapping the weights into the crossbar. Subsequently, only the fully connected layers were retrained in situ to adapt to hardware imperfections in all layers. To address the speed mismatch between the slower convolution layers and the faster fully-connected layers, the authors ran the CNN again but with 3 replicated parallel convolvers. Most importantly, the various techniques presented apply to other networks for overall performance gain.

The CGST PCRAM chip mentioned above was utilized to implement VGG-16 and LeNet-5 CNNs, which achieved 90% accuracy on CIFAR-10 and 98% accuracy on MNIST [69]. The convolution layers were trained by transfer learning, while the fully-connected layers were trained on the PCRAM chip using an optimized direct feedback algorithm (DFA). The algorithm was modified by merging independent error-propagating random feedback matrices into one (Figure 33(a)). This feedback matrix was generated using the intrinsic stochasticity of the PCRAM cells via G-B mapping and was stored in the PCRAM array itself, thereby eliminating the need for random number generators and external memory (Figures 33(b) and (c)). Moreover, the authors proposed either to use a differential pair of PCRAM cells for each weight or to periodically update the averaged conductance in G-B mapping to lessen the impact of conductance drift on the network performance.

Incorporating optical sensing features at the input of CNNs can further enhance the speed and energy-efficiency of machine vision. Seeking to mimic color vision, Dang et al. [59] demonstrated an optical

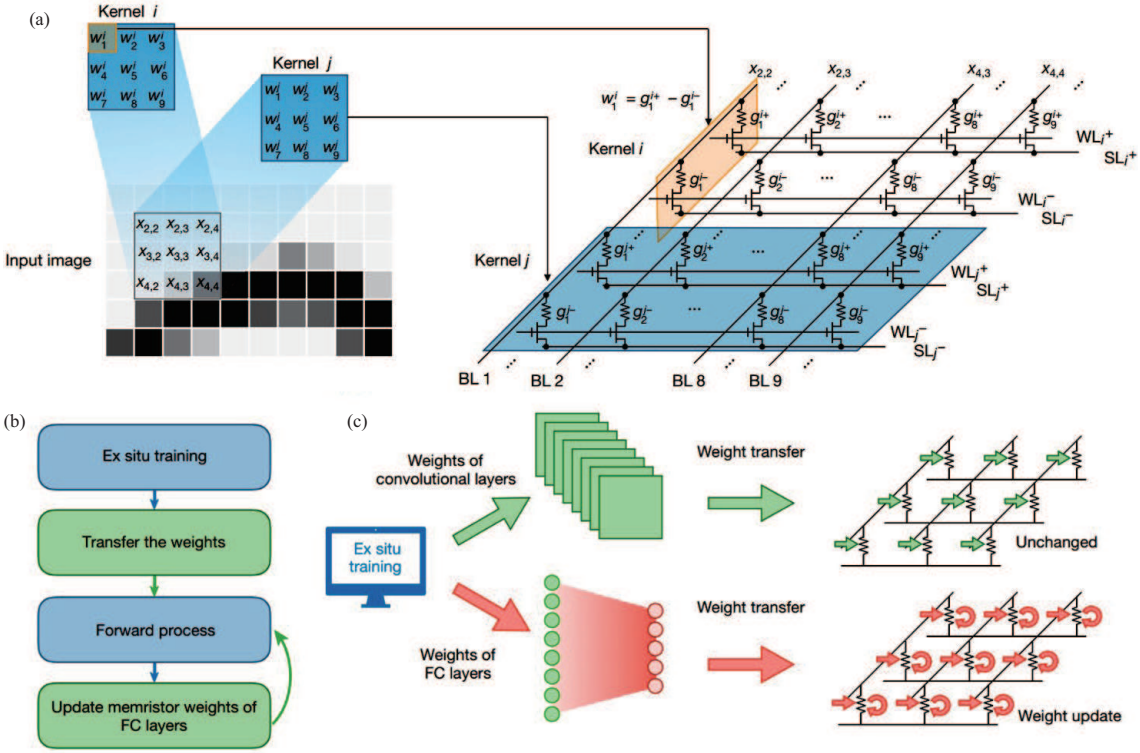


Figure 32 (Color online) (a) An illustration of the flattening of kernels and the side-by-side arrangement in a crossbar. Each kernel receives a shared input during operation. (b) The simplified hybrid training flowchart. (c) An illustration of the hardware-based CNN under this training approach. Ref. [215] @Copyright 2020 Springer Nature.

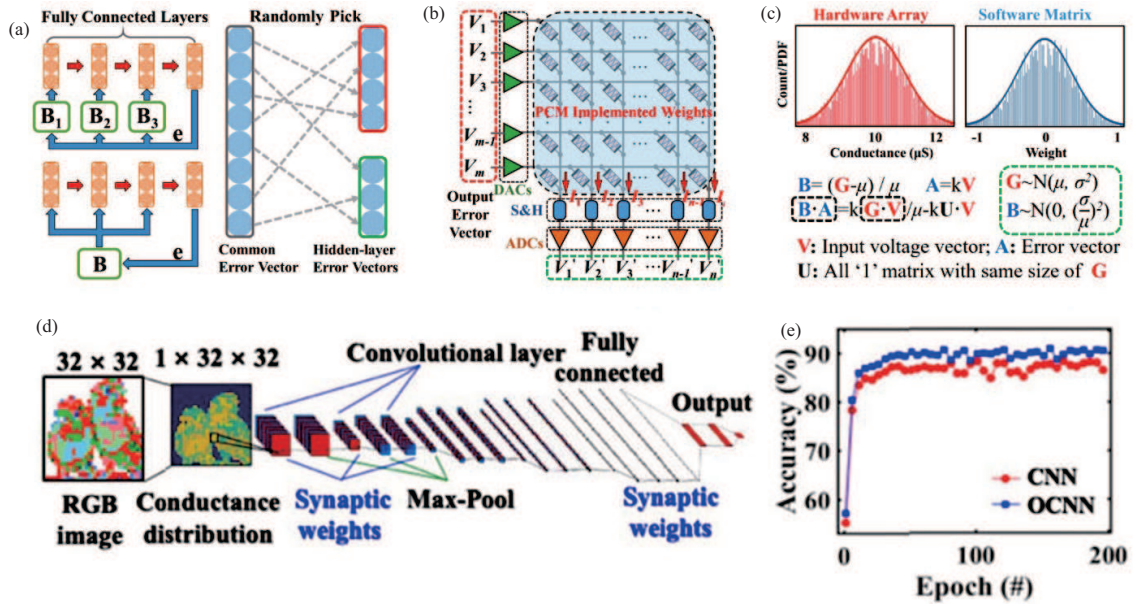


Figure 33 (Color online) (a) An illustration of the conventional DFA versus the optimized DFA. (b) The schematic of in-PCRAM error-computing. (c) The G-B mapping procedure: the mapping relationship between the conductance G and the feedback matrix B. Ref. [69] @Copyright 2020 IEEE. (d) The architecture of OCNN. (e) The accuracies of conventional CNN and OCNN in colored-image recognition. Ref. [59] @Copyright 2020 IEEE.

CNN (OCNN) via simulation based on the previously mentioned W/MgO/ZnO/Mo optic-neural synapse (Figure 33(d)). The network slightly outperformed conventional CNN in colored image recognition (Figure 33(e)).

4.1.3 *Recurrent neural network*

In this subsection, the memristive implementations of three types of recurrent neural networks will be discussed: the LSTM network, the continuous attractor network and the discrete attractor network (also known as the Hopfield network) [204, 205].

Li et al. [216] utilized Ta/HfO_2 synapses in a 1T1R crossbar structure to store synaptic weights shared between different time steps in an LSTM network to accelerate matrix multiplications during both forward and backward passes (Figure 34(a)). The synapses were trained within the crossbar to capture and compensate for hardware imperfections. In the process, the weight updates were calculated and then fed to the crossbar using off-chip electronics. Besides, the gates of the LSTM and the nonlinear activations were software-based. The authors demonstrated airline passenger prediction (regression task) and human identification based on gait (classification task) using 2-layer RNNs (Figures 34(b) and (c)), a proof of concept for realizing networks with different structures and configurations in a crossbar array to reduce latency and power consumption.

Wang et al. [217] investigated a memristor-based implementation of the discrete attractor network (Figure 34(d)). Taking an online and unsupervised approach, the network was trained following the Oja rule. In contrast to training with the Hebbian rule, the memory capacity and the anti-noise ability of the network improved considerably. Furthermore, the authors did pioneering work on implementing a memristor-based continuous attractor network via simulation. Moreover, the impact of noise and device non-idealities on these networks was also studied. The intensive vector-matrix multiplication operations in these networks were computed directly by the crossbar, while the transpose of the weight matrix required by the learning rule was done simply by inverting the input and output ports, both of which underscored the advantage of using memristive in-memory computing.

On the other hand, Yang et al. [218] experimentally demonstrated a memristive optimizer based on the Hopfield network using TaO_x memristors (Figure 34(e)). The optimization process started with a chaotic searching to help the system escape local optima before gradually stabilizing and converging towards the global optimum (Figure 34(f)). The transient chaos was realized by setting a non-zero self-feedback weight to the devices on the diagonal. In subsequent iterations, these weights were ramped down to facilitate convergence (Figure 34(f)). Instead of using carefully designed voltage pulses to update these weights, the authors exploited the nonlinear depression curve of the memristors (Figure 34(g)) and opted for simple identical voltage pulses. This optimizer was applied to find the minimum of the sphere and Matyas function, and to search for the best solution to the Max-cut issue and the 10-city traveling salesman problem.

Meanwhile, Cai et al. [219] utilized 1T1R TaO_x memristors in their Hopfield network to solve binary max-cut problems of various sizes. The authors highlighted the feasibility of leveraging intrinsic noise such as dynamical noise due to conductance fluctuations of the memristors and read noise due to non-ideal crossbar arrays. Hysteretic threshold function was chosen to mimic simulated annealing in modulating the intrinsic noise. Interestingly, this was implemented using a simple comparator-based circuit while achieving performance similar to using the costly simulated annealing. Furthermore, the authors also showed the existence of an optimal noise level for attaining the maximum probability of arriving at the optimal solution (Figures 35(a) and (b)) and the fact that intrinsic noise is less than the optimal noise in arrays of sizes up to over 1000. Thus, they envisioned solving very large max-cut problems with the proposed technique.

Using similar 1T1R TaO_x devices, Lu et al. [220] also studied the effect of intrinsic read noise in memristive Hopfield network while tackling the traveling salesman problem. They included the measured statistics of read noise in the crossbar array in their simulation study and found a trend similar to that in the previous work. The success probability peaked at an optimal noise level. Although larger read noise increases the diversity of solutions and, consequently, the chances of arriving at better solutions, the optimal noise level proved to be the best in terms of time and energy required (Figures 35(c) and (d)).

4.2 Spiking neural network

Although ANNs have proven to be successful in tasks that are otherwise hard to solve using the conventional computing paradigm, they encode information using continuous real values, which differs from our efficient nervous system. On the contrary, SNNs use spike-based coding inspired by their biological counterpart. Therefore, they have far superior energy efficiency as well as the ability to code a higher

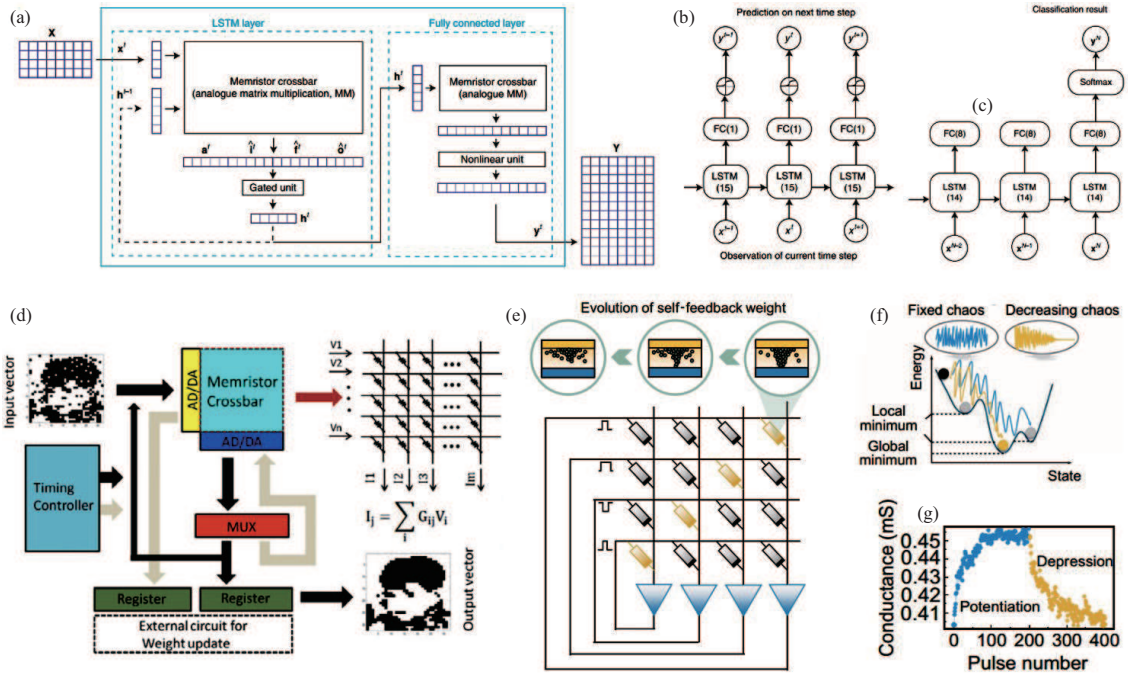


Figure 34 (Color online) (a) A data flow diagram of the memristor-based LSTM network. (b) 2-layer LSTM network with 15 LSTM units and 1 output unit for regression. (c) 2-layer LSTM network with 14 LSTM units and 8 FC output units for classification. Ref. [216] @Copyright 2019 Springer Nature. (d) A schematic of the discrete attractor network implemented using memristors. Ref. [217] @Copyright 2020 John Wiley & Sons, Inc. (e) A schematic of the Hopfield network with self-feedback weights (diagonal devices) and the evolution of these weights in terms of memristor conductance. (f) An illustrative comparison between decreasing chaos and fixed chaos, in which the former facilitates the search for the global minimum (yellow) while the latter fails to converge in the end (blue). (g) The LTP and LTD of the TaO_x based memristor. Ref. [218] @Copyright 2021 The American Association for the Advancement of Science.

density of information within a spike train [221, 222]. To get the most out of SNNs, researchers have explored various hardware implementations of these networks. Herein, we review several notable works.

Xiang et al. [223] simulated a spiking LeNet-5 based on a NFCA tasked with MNIST handwritten digit recognition (Figure 36(a)). The weights were trained on a non-spiking CNN by back-propagation before being mapped onto the spiking network. Each neuron is comprised of an integrated circuit, a latch comparator, and a D flip-flop. The absence of ADCs substantially reduced the chip area and the energy consumption, while the added neuron circuitry contributed very little to these aspects (Figure 36(b)). Although this network was much slower than the non-spiking variant, further optimization will alleviate this problem. These results, coupled with the high classification accuracy, proved that this hardware approach is plausible.

A 2-layer unsupervised SNN based on the L-FeFET neuron was demonstrated by Luo et al. (Figure 36(c)) [85]. The excitatory and inhibitory inputs of the neuron endowed the network with local excitatory and lateral inhibitory features without the need for an extra inhibitory neuron layer. Using these features in the output layer of the SNN, clustering by self-organizing map learning was simulated (Figures 36(d) and (f)), thus mimicking the representation of features in the primate cortex. Moreover, using only lateral inhibition, the network attained high accuracy in an inference task with winner-take-all learning (Figure 36(e)). Due to the capacitor-less nature of the neuron and the aforementioned excitatory and inhibitory connections, this SNN solution is highly scalable.

The information encoding strategy in spike trains determines the capability of an SNN in solving complex tasks. Bao et al. [224] simulated a spiking correlated neural network (SCNN) comprised of three functional layers (Figure 37(a)), wherein the spatial configuration layer (SCL) encodes spatial information, the temporal gating layer (TGL) correlates the SCL outputs via spatiotemporal integration with a clock signal, while the selective output layer decodes the information. This network was capable of recognizing rotated patterns (Figure 37(b)) and identifying their rotational angles simultaneously. The pattern types were represented by neural correlation between output spike trains, upon cross-correlation analysis, which reveals the encoded information regarding the pattern. On the other hand, the rotational angles were encoded as the spiking frequency of one of the output neurons. All in all, this network was

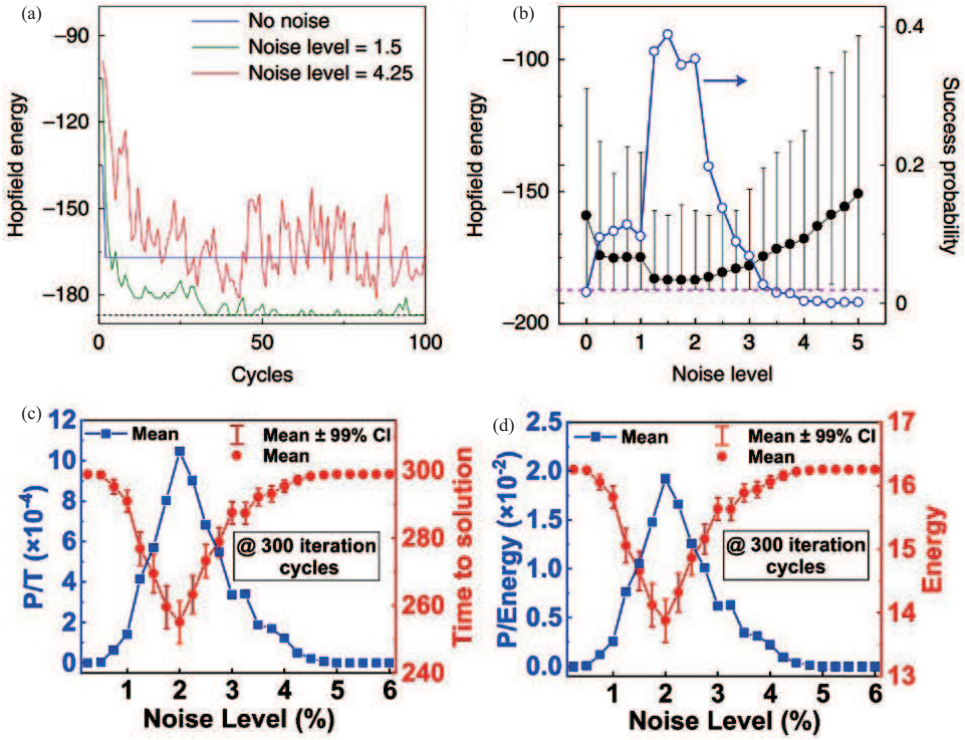


Figure 35 (Color online) (a) An example of the effect of different noise levels in the search for the lowest energy in a Hopfield network. (b) The effect of noise level on the energy result statistics and the probability of reaching the optimal solution. Ref. [219] @Copyright 2020 Springer Nature. The effect of read noise level on the time (c) and energy (d) required to reach a solution. P/T and P/Energy represent the success probability per unit time and per unit energy, respectively. Ref. [220] @Copyright 2020 IEEE.

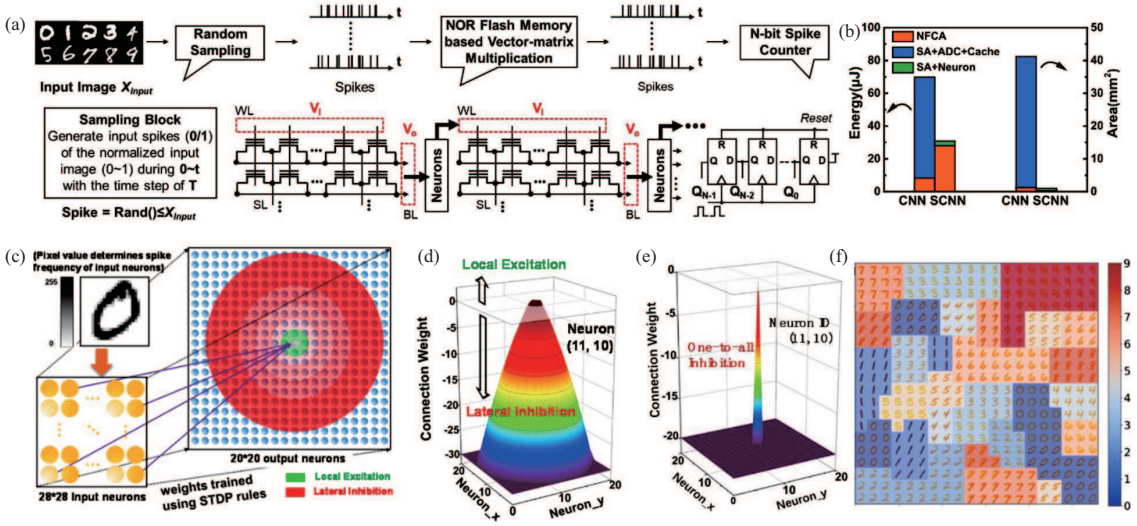


Figure 36 (Color online) (a) A schematic of the NFCA-CNN. (b) The energy and area overhead of the spiking and non-spiking NFCA-CNN. Ref. [223] @Copyright 2020 IEEE. (c) The architecture of the leaky-FeFET SNN. (d) The connection weight topology of the center neuron illustrating local excitation and lateral inhibition. (e) One-to-all inhibition (winner-take-all). (f) The clustering and spatial representation of features. Ref. [85] @Copyright 2019 IEEE.

made possible by the network topology employed and the MoS_2 neuristor in the TGL layer, which was capable of integrating SCL output signals and a clock signal applied to its gates (Figure 37(c)).

Apart from implementing SNNs using 3-terminal devices as discussed above, 2-terminal devices can also be used (Figure 37(d)). Wang et al. [225] constructed a fully memristive SNN using Ag-nanoparticle-based diffusive memristors as neurons and HfO_2 non-volatile memristors as synapses. Apart from demonstrating a simple forward convolution process with pre-programmed synaptic weights (Figure 37(e)) and the

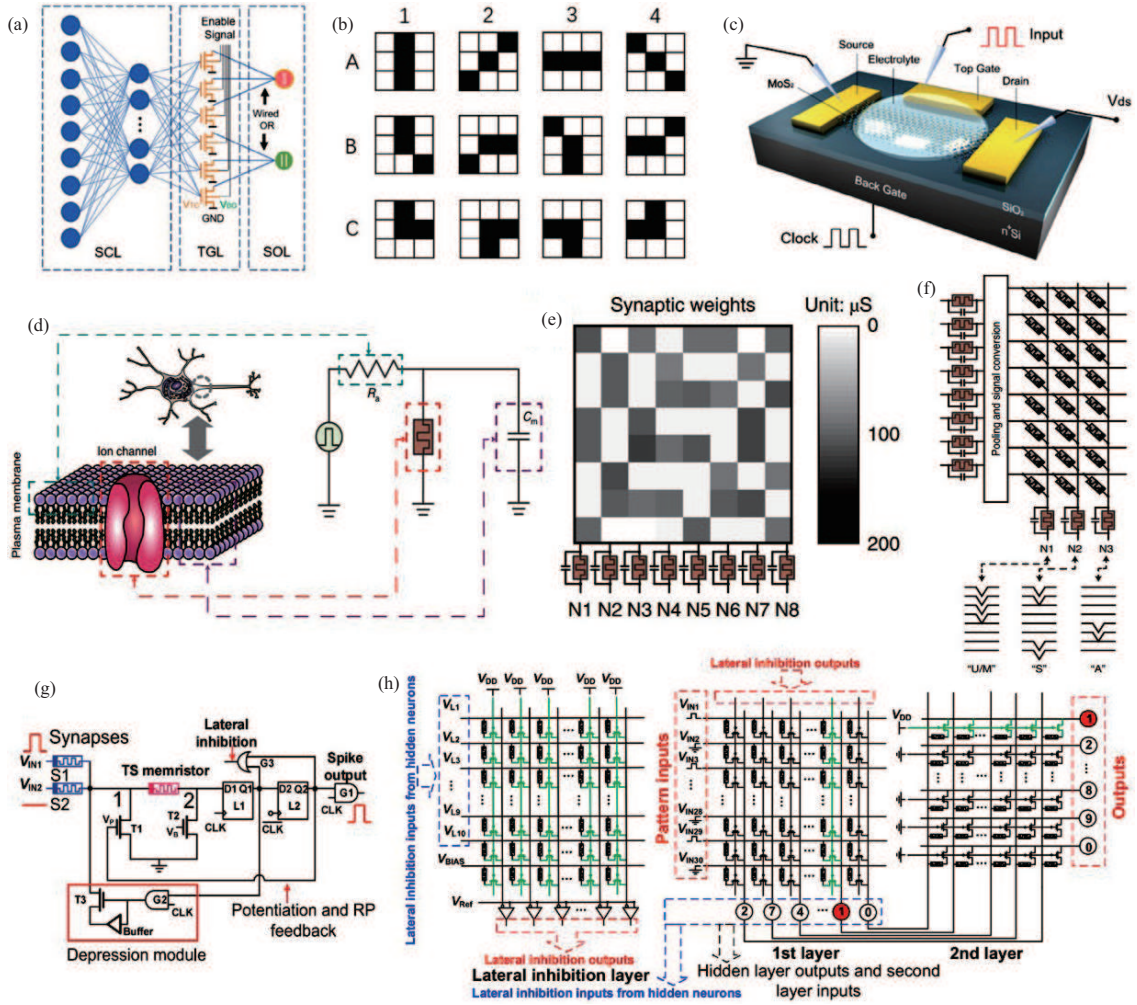


Figure 37 (Color online) (a) The architecture of the SCNN is comprised of a spatial configuration layer, a temporal gating layer, and a selective output layer. (b) The input patterns (A, B, C) with 4 rotations each. (c) The MoS_2 device with ionic top gate and electronic bottom gate. Ref. [224] @Copyright 2019 John Wiley & Sons, Inc. (d) A schematic of a LIF neuron using a 2-terminal diffusive memristor. The artificial neuron is analogous to the membrane of a biological neuron. (e) A synaptic array connected to 8 neurons. Each column is a pre-programmed convolution kernel. (f) A fully connected network receiving software-processed inputs from the neurons in (e) for the classification of 3 different patterns. Ref. [225] @Copyright 2018 Springer Nature. (g) A schematic of the hybrid spiking neuron. (h) A schematic of the 2-layer SNN with lateral inhibition in the first layer. Ref. [226] @Copyright 2021 Elsevier.

inherent ReLU activation of the neurons, the authors trained a simple fully connected network (Figure 37(f)) using the STDP rule with an unsupervised approach. Thanks to the design of the neuron, the potentiation of the synapses with an input happens automatically during a firing event without the need for external circuitries. Nevertheless, the simple circuit requires software-based pooling operation and signal conversion in addition to lateral inhibition and depression of synapses by a microcontroller.

By introducing digital components into the basic memristive neuron circuit, Zhang et al. [226] realized a hybrid spiking neuron featuring active spike output in addition to in situ potentiation and depression of synapses (Figure 37(g)). Besides, this neuron conveniently enables lateral inhibition in SNNs using a pre-programmed array instead of microcontroller units. The authors built a 2-layer SNN from Ag-nanoparticle neurons and HfO_2 synapses (Figure 37(h)). The network performed simple digit recognition in which the first layer was trained in an unsupervised fashion with a winner-take-all rule and a lateral inhibition process. On the other hand, supervised learning simply by controlling the shared gates of the synapses was employed in the final layer. This full-featured neuron represents a possible path towards realizing larger and deeper fully hardware-implemented SNNs.

Other than the Ag diffusive memristor, neurons featuring the NbO_x Mott device have also been utilized in SNNs. A fully memristive SNN utilizing only 2-terminal devices which were NbO_x volatile

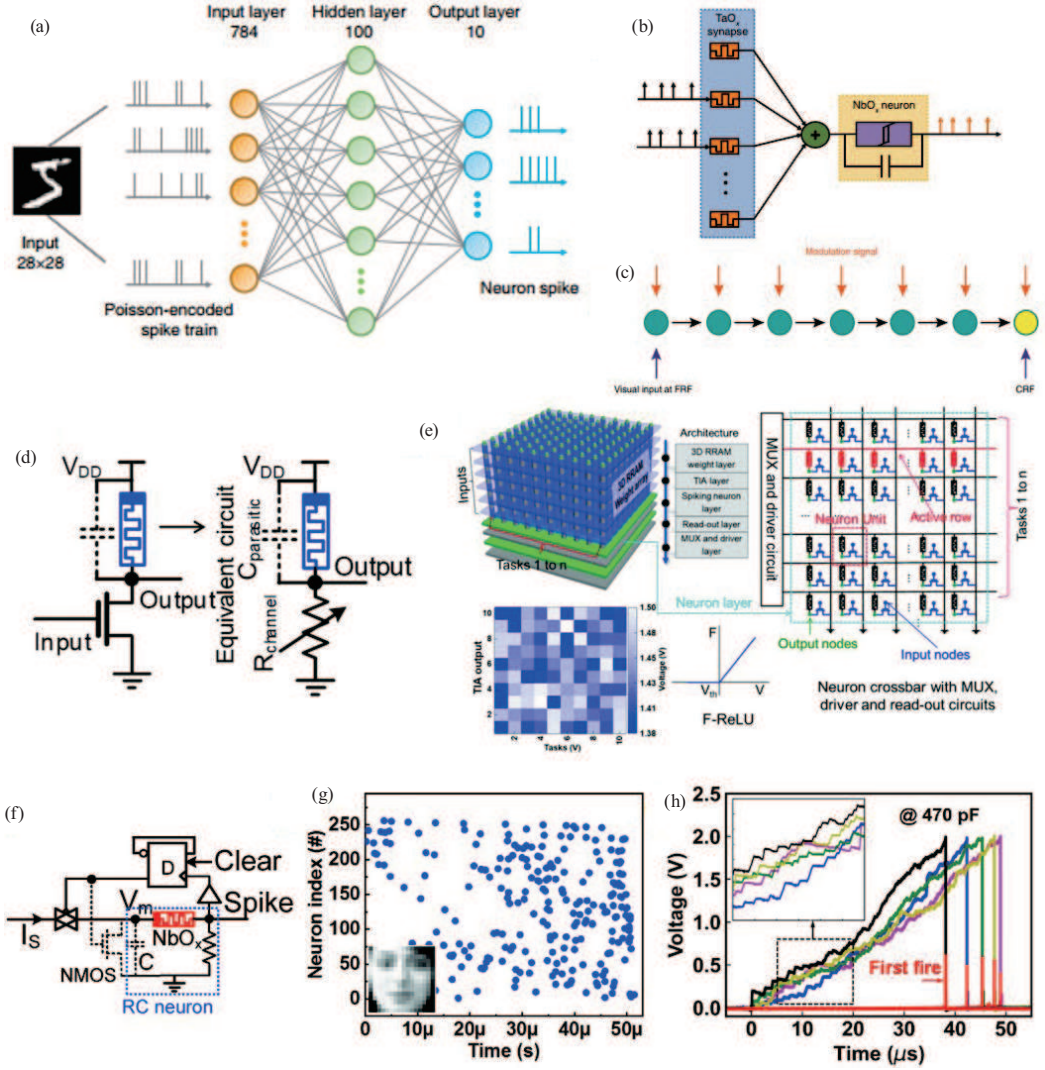


Figure 38 (Color online) (a) A schematic of the 3-layer SNN for image classification. (b) Coincidence detection. (c) The 1D unidirectional network for receptive field remapping. Ref. [227] @Copyright 2020 Springer Nature. (d) A schematic of the 1T1R neuron. (e) An illustration of a neuron crossbar with integrated 3D synaptic array, the outputs of 10 simulated parallel tasks and the voltage-to-frequency ReLU function of the neuron in (d). Ref. [228] @Copyright 2019 IEEE. (f) A schematic of the single-spiking neuron for temporal coding. (g) Every neuron fires only once. The firing delay encodes the corresponding pixel value. (h) The response of each output neuron. Ref. [229] @Copyright 2020 IEEE.

memristors as spiking neurons and TaO_x non-volatile memristors as synapses was presented by Duan et al. [227]. The devices were monolithically integrated to form an SNN. Offline learning, as well as online learning using a simplified δ-rule, was experimentally demonstrated on the network to recognize simple patterns. Besides, the authors simulated a 3-layer SNN in classifying MNIST handwritten digits, trained using back-propagation (Figure 38(a)). Furthermore, a large-scale coincidence detection application was also simulated, in which the network fired spikes upon detecting synchronous input spiking events (Figure 38(b)). The authors also simulated receptive field remapping via multiplicative gain modulation in a one-dimensional network (Figure 38(c)). This is crucial for a more stable artificial visual system.

Zhang et al. [228] reported a conversion-based rate-coded single layer SNN using NbO_x-based neurons (Figure 38(d)). The training was done in software by incorporating the extracted ReLU-like statistical spiking rate versus input gate voltage relationship in addition to the membrane potential of the LIF neuron. High conversion accuracy from software to hardware was attained. In addition, the authors also put forward the possibility of a neuron crossbar owing to the 1T1R structure. By integrating it with a 3D synaptic array, parallel multitasking can be achieved (Figure 38(e)).

Recently, Zhang et al. [229] also introduced a single-spiking NbO_x-based LIF neuron taking advantage of fast and energy-efficient temporal coding (Figure 38(f)). In their approach, information is encoded in

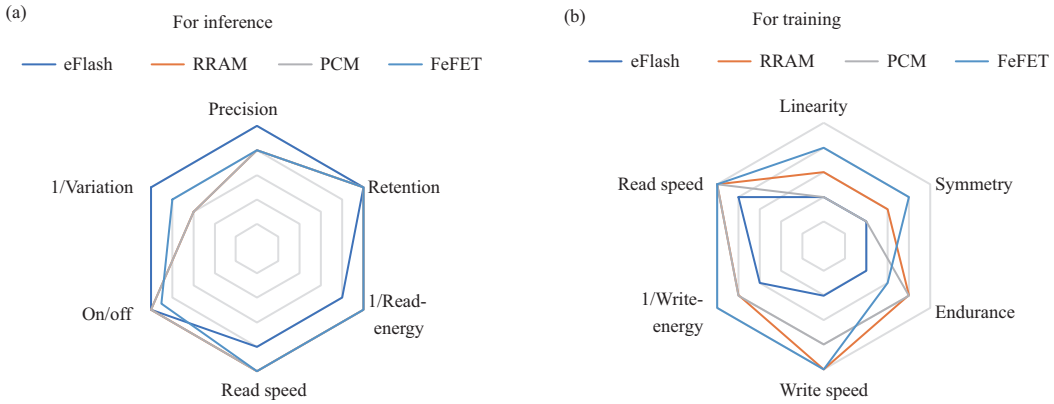


Figure 39 (Color online) Comparison of in-memory computing performance based on emerging nonvolatile memory devices (a) for inference and (b) for training.

the spiking latency, meaning that the input spike delay is dependent on the input value and the output neuron that fires first defines the result (Figures 38(g) and (h)). To demonstrate the feasibility of such a neuron, a single layer SNN was trained to recognize Olivetti face patterns using a supervised temporal backpropagation algorithm, after which the trained weights were mapped into a physical synaptic array. It is also worth mentioning that this neuron features a wider input current range and a longer lifetime compared to its rate-coding counterpart.

Lastly, Cu-Ta/IGZO/TiN Schottky device with unipolar threshold switching exhibits intrinsic stochastic dynamics, which is desirable for certain applications. Dang et al. [230] took advantage of this property to compute the global minimum of a numerical function. The optimization problem was solved iteratively using a random walk algorithm. Thus, this work showcased the viability of incorporating stochastic neurons in computing systems to enrich their processing capabilities.

For the applications of neural networks, inference and training have different requirements for device performance. And the best performance indications that can be achieved by the different kinds of devices for inference and training are displayed in (Figures 39(a) and (b)), respectively.

There are some challenges for the application of the neural network. The non-idealities in RRAM devices such as time-dependent variability and early-stage resistance fluctuations will impact the accuracy of pattern recognition. And the endurance of the devices will bug the implementation of in-memory neural networks, especially during frequent weight updates when training. Moreover, devices that support high weight precision are also required to implement certain networks, which poses a challenge for memristors. Besides, while controlled noise in crossbar arrays is beneficial for various optimization problems, attenuating it in working memory applications might improve memory retention.

And although the simulation of NOR flash shows promising results, the scaling down conundrum of transistors does not benefit the scaling up of CNNs and ANN as a whole. Hence, emerging neuromorphic devices is in the spotlight. While the size of simulated in-memory implementations of CNNs can reach about 20 layers deep, physical demonstrations are still limited to very few layers due to the sensitivity of large arrays or multiple arrays to hardware imperfections. Except this, the hardware optimizations such as reducing device-to-device variations and IR drops due to interconnects are still required to improve training efficiency and network performance.

As can be seen, the quest for better hardware is a universal goal towards plausible hardware-based neural networks. It is worth mentioning that the potential of combining CNNs and recurrent connections for more advanced applications like video processing will further tighten these requirements. And several strategies on the algorithmic level such as the sparsified backpropagation and the 'example triage' selective training method can get around this problem.

As for the application of SNN, although simulations of slightly larger networks exist, the physical implementation is currently limited to shallow depths. Besides, only layer-by-layer basis training and simple tasks had been demonstrated. These are due to the lack of better algorithms. Furthermore, a neuron requires either peripheral circuitries or additional components in its circuit design to drive the synapses in subsequent layers, thus increasing area overhead and undermining the promised benefits of in-memory computing. This might limit the scalability of SNNs in the future. Generally, the search for new algorithms and improved circuit design are desperately needed to implement SNNs in hardware

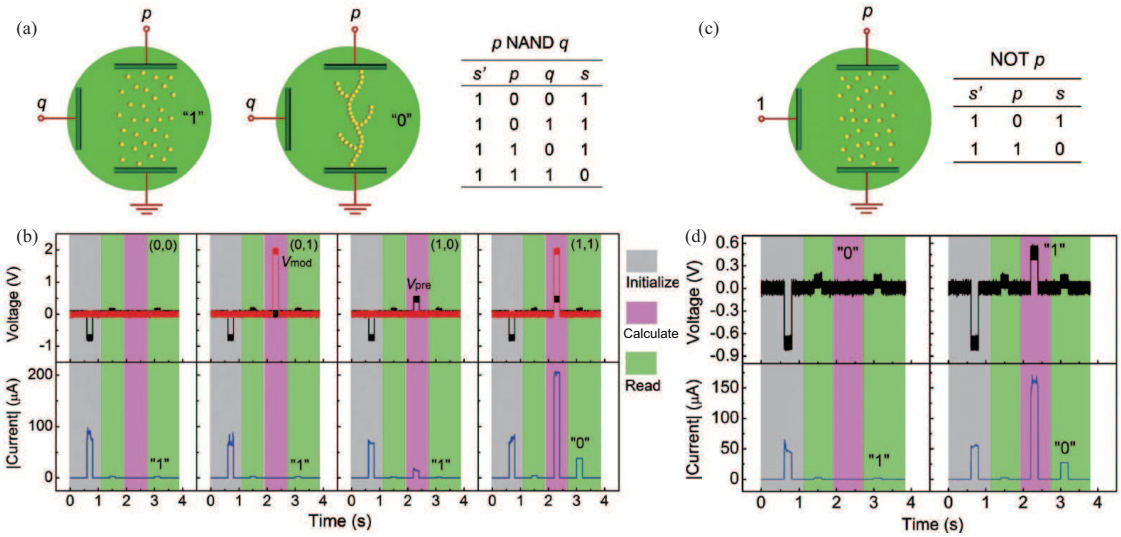


Figure 40 (Color online) Based on the vertical heterosynaptic devices. (a) and (c) Operation and truth table of NAND logic and NOT logic. (b) and (d) The implementation of nonvolatile NAND logic and NOT logic function. Ref. [62] @Copyright 2017 John Wiley & Sons, Inc.

efficiently.

4.3 Digital logic in memory

Logic in memory as known as nonvolatile logic can effectively implement future non-von Neumann architecture computing. And traditional CMOS logic operations are often based on a complementary transistor. The input and output of the logic circuit are both high and low voltages which are often accompanied by volatility, and the logic state cannot be maintained after power off. Logic based on non-volatile memristors can solve this problem, as its logic output is the resistance of the device. And the implementation of logic operations is realized by the transition between the HRS and LRS of the memristor. The memristor in the array or chip can perform distributed parallel computing, which will greatly improve the efficiency of computing [231, 232]. Such a mode of in-situ integration of logic and memory eliminates the need for traditional computer architecture to transfer calculation results to memory or external storage, which can effectively reduce the load of frequent data transmission and the power consumption of information processing as well as improve the speed and efficiency of information processing.

The above-mentioned vertical heterosynaptic devices manufactured by Yang et al. [62] can implement more efficient non-volatile Boolean logic. Figures 40(a) and (c) demonstrate the operation scheme and truth table of NAND logic and NOT logic. Besides, the experimental implementations are displayed in Figures 40(b) and (d), respectively. It is worth noting that the realization of NAND in a traditional 2-terminal memristor needs at least 3 logic cycles [233], while the AND logic and NOT logic in this work were implemented in less than 2 logic cycles. This can be explained by the participation of the modulatory terminal.

Furthermore, Xu et al. [234] realized all 16 Boolean logic in up to 3 logic cycles. The devices were fabricated by inserting a SiO_2 layer (3 nm) between two Ta_2O_5 layers (10 nm) to reduce the threshold voltages (Figure 41(a)). The operation speed of the unipolar devices is displayed in Figure 41(b). Figure 41(c) shows the definition of input and output variables. It should be pointed out that the logic functions were implemented in this work without containing selectors or diodes. Figure 41(d) demonstrates the implementation of the XOR and NAND functions under the pulse measurements. Lastly, the 1-bit binary full adder was experimentally demonstrated with 5 unipolar devices within 8 logic cycles as shown in Figure 41(e). Subsequently, within 6 computing steps, Yuan et al. [235] achieved a 1-bit full adder with only 3 devices and a 2-bit multiplier with 5 devices based on bipolar memristor.

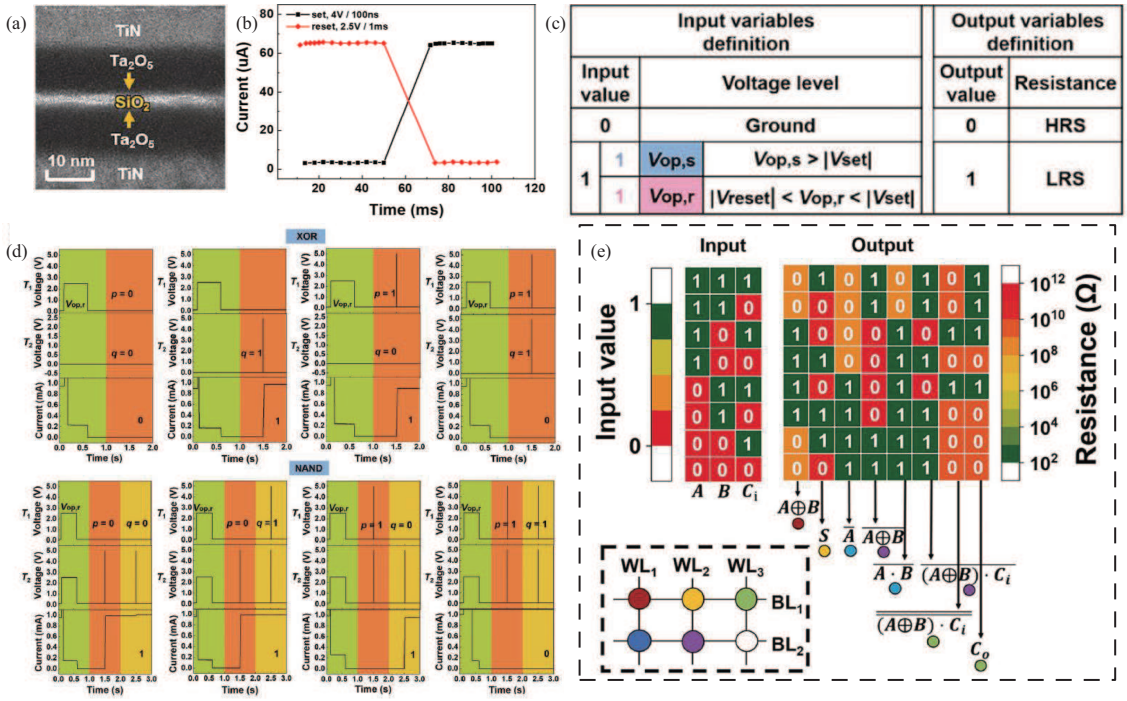


Figure 41 (Color online) (a) TEM image of the $TiN/Ta_2O_5/SiO_2/Ta_2O_5/TiN$ device, scale bar: 10 nm; (b) RS under a single pulse; (c) input and output variable definitions; (d) the implementation of the XOR and NAND functions under pulse measurements; (e) the realization of a 1-bit binary full adder. Ref. [234] @Copyright 2019 John Wiley & Sons, Inc.

4.4 Hardware security

With the unprecedented development of interconnected networks, a large amount of information transmitted on the networks is facing major security challenges [236]. Traditional encryption methods usually depend on the keys stored in NVRS memory, which are susceptible to physical and side-channel attacks to be decrypted [237, 238]. Therefore, the physically unclonable function (PUF) has attracted great attention recently due to its unpredictability caused by the changes of the silicon process and the inherent physical variety [239–241]. Memristors stand out among various emerging devices due to their potential stochastic operation [242–244].

Encryption and decryption of information are a way to ensure information security. Novel hardware encryption was demonstrated by Xu et al. [234] based on the unipolar memristive array as schematically shown in Figure 42(a). The high efficiency of the unipolar devices on the realization of the XNOR function makes it possible to construct novel encryption hardware, in which the encryption keys are generated according to the inherent randomness of HRS to ensure information security in a simple and effective method [245, 246]. Figures 42(b)–(d) present true ciphertext of "W", "L", and "R" obtained by encrypting "P", "K", and "U", respectively. The recovery of "P", "K", and "U" was successfully realized through the decryption process as shown in Figures 42(e)–(g).

Hamming code is one of the most famous block coding methods, which can perform error detection and correction on data blocks [247]. By design, this algorithm can effectively rectify any errors caused by noisy data transmission. As the encoded information has the smallest redundancy, called a codeword, with a length of n bits, and the length of the parity bit is $(n-k)$ bits, where k is the length of the message expected to be encoded [248], the Hamming code has lower space overhead and makes the message more secure.

Sun et al. [249] experimentally demonstrated hardware error corrections based on a network of unipolar memristors as shown in Figure 43(a). And Figures 43(b) and (c) displayed the output syndrome vectors z, which is (0 0 0) or (1 1 1) respectively when the corresponding input codeword is 1111111 or 1100111. The results indicate that this network has the ability to correct errors.

True random number generators (TRNGs) are critical for the applications in encryption [246, 250]. Zhang et al. [251] utilized an alternate reading scheme in the TRNG circuit to improve the unbiasedness of the random numbers. The response time of the TaO_x based memristors utilized in this work is almost

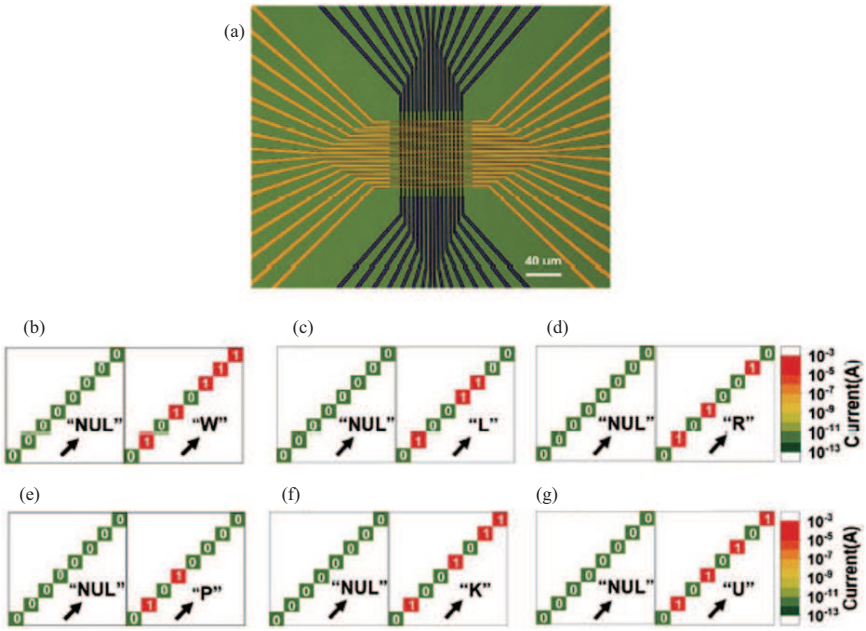


Figure 42 (Color online) (a) SEM of 16 × 16 memristive array, scale bar: 40 μm; (b)–(d) The encryption of "P", "K", and "U" produces the ciphertexts of "W", "L", and "R"; (e)–(g) The decryption of "P", "K", and "U". Ref. [234] @Copyright 2019 John Wiley & Sons, Inc.

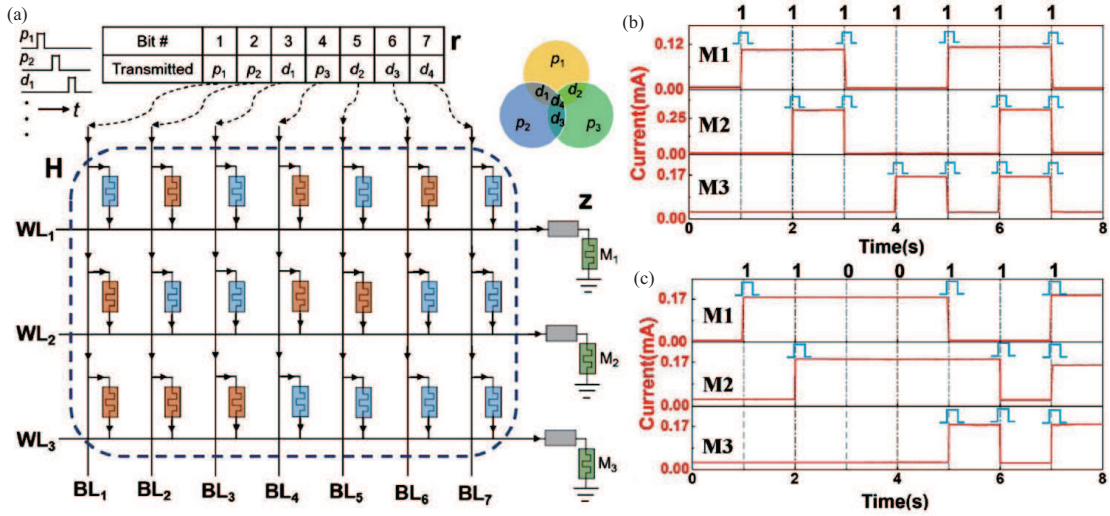


Figure 43 (Color online) (a) Schematic of the network for Hamming code error correction; the corresponding states for three memristors as the input codeword is (b) 1111111 and (c) 1100111. Ref. [249] @Copyright 2019 IEEE.

30 ns as shown in Figure 44(a). And Figure 44(b) shows a good endurance of the device, which guarantees the length of the generated bitstream. As the HRS of the device is sensitive to the location where the CF breaks, it is more stochastic and unpredictable than the LRS [252]. The circuit diagram of the TRNG is depicted in Figure 44(c). Furthermore, an alternate reading scheme was put forward, where the polarity of V_R is reversed in each cycle. Under the alternating read schemes, the endurance of V'_{out} in the circuit is more uniform at about 23000 bits (Figure 44(d)). As displayed in Figures 44(e) and (f), the bitmap of V_{out} under the proposed read scheme has a less dark area, indicating the superior unbiasedness of the alternating read scheme.

Subsequently, a novel physically transient TRNG based on MgO switching device was realized by Dang et al. [253]. With the connection of two VTS memristors, the circuit of RNG was built as shown in Figure 45(a). It should be noted that due to the volatility of the devices, there is no need to perform a reset process in each cycle, which greatly simplifies the operation of the circuit. Figure 45(b) shows the generation of a random bitstream in 19 bits in detail. TRNGs have a wide range of applications.

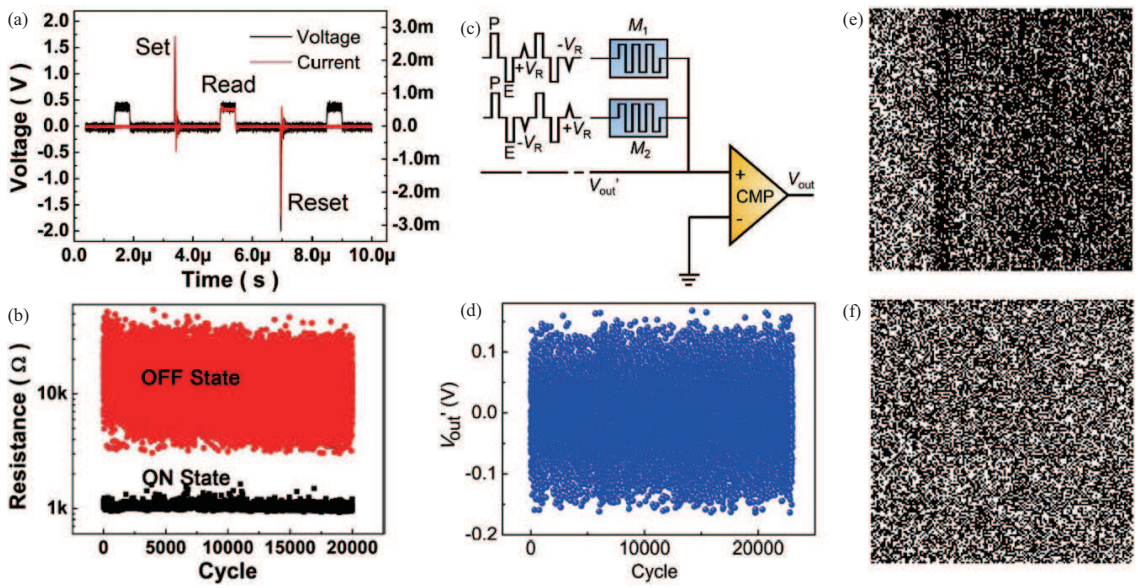


Figure 44 (Color online) (a) Pulse respond behavior; (b) endurance property; (c) circuit diagram of the TRNG; (d) the endurance of $V'_{\rm out}$; (e) and (f) the bitmap of $V_{\rm out}$ with 146 × 146 random bits in conventional read scheme and proposed read scheme, respectively. Ref. [251] @Copyright 2017 Elsevier.

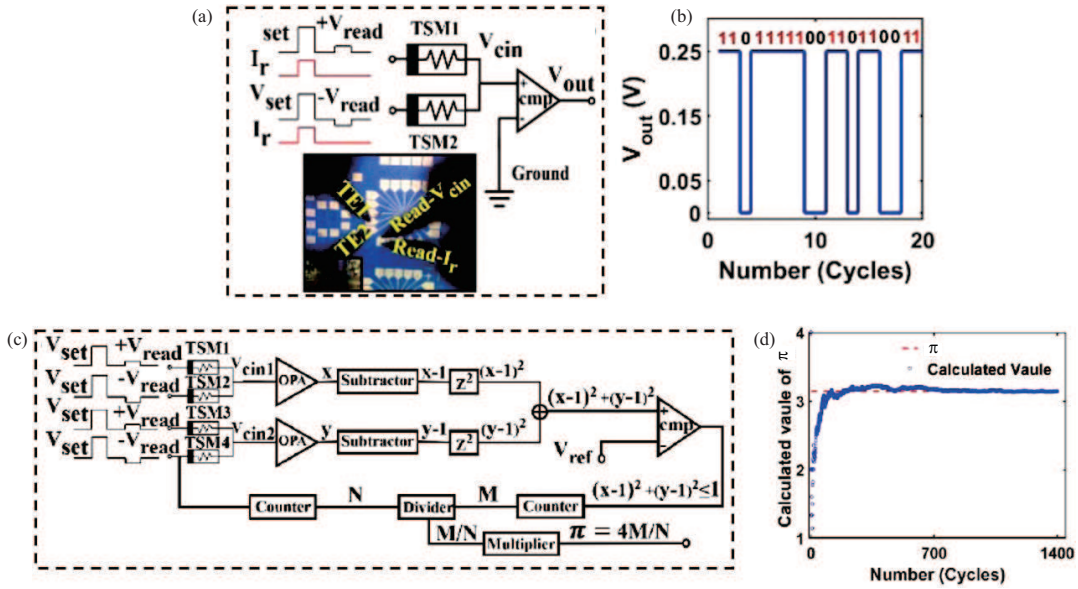


Figure 45 (Color online) (a) Circuit diagram of the RNG; (b) the generation of random bitstream in 19 bits; (c) schematic of the computing module for calculation of the π value; (d) the result of calculated π values. Ref. [253] @Copyright 2019 IEEE.

And the Monte Carlo calculation module based on two TRNGs shown in Figure 45(c) has the ability to calculate π value (Figure 45(d)) [252, 254].

5 Remaining challenges and outlook

Here we have given an overview of recent developments in in-memory computing based on emerging non-volatile memories. Although there have been encouraging progress in in-memory computing architecture, there are still many remaining challenges. The first challenge lies in the scale of on-chip memory compared with that of the neural network models. Nowadays the size of the neural network model doubles every 3.4 months so that the capacity of on-chip memory must be further increased in order to efficiently support the advanced neural network applications. The second challenge is that all the existing devices

for in-memory computing architecture have their shortcomings, and a perfect device is not present yet. Therefore, further device engineering is still needed to better support the application. The third challenge is that besides device optimizations, extensive efforts need to be devoted to optimizing the interface circuits between the digital controlling circuits and the analog MAC arrays, otherwise the advantage of in-memory computing may be compromised. A possible future direction in further enhancing the efficiency of the system might be to utilize memristors not only for weight elements or synapses but also for output neurons, therefore exploiting memristors to their full potential.

It is worth noting that the requirements on device performance for inference and training are very different, and the optimization of the device should be implemented with reference to the particular application. In order to do inference, the most important requirements are high weight precision to guarantee network performance, long retention in the storage of weight parameters, low read energy, and low device variations. In stark contrast, training requires high linearity and high symmetry when modulating the state of the weight element and high endurance, because the devices need to be updated much more frequently than inference, and the devices need to be written fast with low programming energy. Certainly, both inference and training have common requirements in fast read and high device yield, but overall it is obvious that training is more demanding than inference.

Currently, a perfect device that can meet all the above requirements is not yet present. Embedded flash has low write speed, high write energy, and low endurance, which means it is not suitable for training. RRAM, PCM, and MRAM are attractive high-performance device technologies for in-memory computing, and they all fall into the general category of memristors. RRAM suffers from variation issues. PCM also suffers from variation issues, and it has low linearity and symmetry. MRAM has a small on/off ratio, despite the best endurance. Recently, FeFET has also been studied for in-memory computing and has shown ultralow power, but its endurance is still a problem. In order to support big AI models, the on-chip memory capacity needs to be as big as possible. While flash almost stops scaling, all the other device technologies have reached 20 or 22 nm nodes and have the potential in scaling to more advanced nodes. Therefore, we believe with continued device optimization and integration, in-memory computing based on emerging nonvolatile memories could have a great prospect in enabling edge platforms and mobile devices with high intelligence and energy efficiency.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2017YFA0207600), National Natural Science Foundation of China (Grant Nos. 61925401, 92064004, 61927901), the Project supported by PKU-Baidu Fund (Grant Nos. 2019BD002, 2020BD010), and the 111 Project (Grant No. B18001). Yuchao YANG acknowledges the support from the Fok Ying-Tong Education Foundation, Beijing Academy of Artificial Intelligence (BAAI), and the Tencent Foundation through the XPLORER PRIZE.

References

1 Wulf W A, McKee S A. Hitting the memory wall: implications of the obvious. ACM SIGARCH Comput Archit News, 1995, 23: 20–24
2 Horowitz M. Computing's energy problem (and what we can do about it). In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2014. 10–14
3 Backus J. Can programming be liberated from the von Neumann style? Commun ACM, 1978, 21: 613–641
4 Merolla P A, Arthur J V, Alvarez-Icaza R, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science, 2014, 345: 668–673
5 Waldrop M M. The chips are down for Moore's law. Nature, 2016, 530: 144–147
6 Mutlu O, Ghose S, Gómez-Luna J, et al. Processing data where it makes sense: enabling in-memory computation. Microprocessors MicroSyst, 2019, 67: 28–41
7 Alpern B, Carter L, Feig E, et al. The uniform memory hierarchy model of computation. Algorithmica, 1994, 12: 72–109
8 Balasubramonian R, Albonesi D, Buyuktosunoglu A, et al. Memory hierarchy reconfiguration for energy and performance in general-purpose processor architectures. In: Proceedings of IEEE/ACM International Symposium on Microarchitecture, 2000. 245–257
9 Keckler S W, Dally W J, Khailany B, et al. GPUs and the future of parallel computing. IEEE Micro, 2011, 31: 7–17
10 Jouppi N P, Young C, Patil N, et al. In-datacenter performance analysis of a tensor processing unit. In: Proceedings of ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA), 2017. 1–12
11 Sze V, Chen Y H, Yang T J, et al. Efficient processing of deep neural networks: a tutorial and survey. Proc IEEE, 2017, 105: 2295–2329
12 Patterson D, Anderson T, Cardwell N, et al. A case for intelligent RAM. IEEE Micro, 1997, 17: 34–44
13 Ahn J, Yoo S, Mutlu O, et al. PIM-enabled instructions: a low-overhead, locality-aware processing-in-memory architecture. In: Proceedings of ACM/IEEE 42nd Annual International Symposium on Computer Architecture (ISCA), 2015. 336–348
14 Sebastian A, Le Gallo M, Khaddam-Aljameh R, et al. Memory devices and applications for in-memory computing. Nat Nanotech, 2020, 15: 529–544
15 Ielmini D, Wong H S P. In-memory computing with resistive switching devices. Nat Electron, 2018, 1: 333–343
16 Wong H S P, Salahuddin S. Memory leads the way to better computing. Nat Nanotech, 2015, 10: 191–194
17 Zhu J, Zhang T, Yang Y, et al. A comprehensive review on emerging artificial neuromorphic devices. Appl Phys Rev, 2020, 7: 011312
18 Ma W, Zidan M A, Lu W D. Neuromorphic computing with memristive devices. Sci China Inf Sci, 2018, 61: 060422

19 Li Y, Zhou Y X, Wang Z R, et al. Memcomputing: fusion of memory and computing. Sci China Inf Sci, 2018, 61: 060424
20 Haario H, Laine M, Mira A, et al. DRAM: efficient adaptive MCMC. Stat Comput, 2006, 16: 339–354
21 Jacob B, Ng S, Wang D. Memory Systems: Cache, DRAM, Disk. San Francisco: Morgan Kaufmann, 2010
22 Chung Y, Song S H. Implementation of low-voltage static RAM with enhanced data stability and circuit speed. MicroElectron J, 2009, 40: 944–951
23 Lanza M, Wong H S P, Pop E, et al. Recommended methods to study resistive switching devices. Adv Electron Mater, 2019, 5: 1800143
24 Raoux S, Burr G W, Breitwisch M J, et al. Phase-change random access memory: a scalable technology. IBM J Res Dev, 2008, 52: 465–479
25 Scott J F, de Araujo C A P. Ferroelectric memories. Science, 1989, 246: 1400–1405
26 Bez R, Camerlenghi E, Modelli A, et al. Introduction to flash memory. Proc IEEE, 2003, 91: 489–502
27 Goldhaber-Gordon D, Montemerlo M S, Love J C, et al. Overview of nanoelectronic devices. Proc IEEE, 1997, 85: 521–540
28 Chen A, Hutchby J, Zhirnov V, et al. Emerging Nanoelectronic Devices. Hoboken: John Wiley & Sons, 2014
29 Chua L. Memristor-the missing circuit element. IEEE Trans Circuit Theor, 1971, 18: 507–519
30 Chua L. Resistance switching memories are memristors. In: Handbook of Memristor Networks. Cham: Springer, 2019. 197–230
31 Wang Z, Joshi S, Savel'ev S E, et al. Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. Nat Mater, 2017, 16: 101–108
32 Jo S H, Chang T, Ebong I, et al. Nanoscale memristor device as synapse in neuromorphic systems. Nano Lett, 2010, 10: 1297–1301
33 Ravichandran V, Li C, Banagozar A, et al. Artificial neural networks based on memristive devices. Sci China Inf Sci, 2018, 61: 060423
34 Yang C J, Adhikari S P, Kim H. Excitatory and inhibitory actions of a memristor bridge synapse. Sci China Inf Sci, 2018, 61: 060427
35 Tang J, Yuan F, Shen X, et al. Bridging biological and artificial neural networks with emerging neuromorphic devices: fundamentals, progress, and challenges. Adv Mater, 2019, 31: 1902761
36 Kumar S, Graves C E, Strachan J P, et al. Direct observation of localized radial oxygen migration in functioning tantalum oxide memristors. Adv Mater, 2016, 28: 2772–2776
37 Kumar S, Wang Z, Huang X, et al. Conduction channel formation and dissolution due to oxygen thermophoresis/diffusion in hafnium oxide memristors. ACS Nano, 2016, 10: 11205–11210
38 Wuttig M, Yamada N. Phase-change materials for rewriteable data storage. Nat Mater, 2007, 6: 824–832
39 Lencer D, Salinga M, Grabowski B, et al. A map for phase-change materials. Nat Mater, 2008, 7: 972–977
40 Martin L W, Rappe A M. Thin-film ferroelectric materials and their applications. Nat Rev Mater, 2017, 2: 16087
41 Kim S J, Mohan J, Summerfelt S R, et al. Ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ thin films: a review of recent advances. JOM, 2019, 71: 246–255
42 Novoselov K S. Electric field effect in atomically thin carbon films. Science, 2004, 306: 666–669
43 Dean C R, Young A F, Meric I, et al. Boron nitride substrates for high-quality graphene electronics. Nat Nanotech, 2010, 5: 722–726
44 Stuart M A C, Huck W T S, Genzer J, et al. Emerging applications of stimuli-responsive polymer materials. Nat Mater, 2010, 9: 101–113
45 Laoutid F, Bonnaud L, Alexandre M, et al. New prospects in flame retardant polymer materials: from fundamentals to nanocomposites. Mater Sci Eng-R-Rep, 2009, 63: 100–125
46 Pan F, Gao S, Chen C, et al. Recent progress in resistive random access memories: materials, switching mechanisms, and performance. Mater Sci Eng-R-Rep, 2014, 83: 1–59
47 Lee S R, Kim Y B, Chang M, et al. Multi-level switching of triple-layered TaO_x RRAM with excellent reliability for storage class memory. In: Proceedings of Symposium on VLSI Technology (VLSIT), 2012. 71–72
48 Li J, Yang Y, Yin M, et al. Electrochemical and thermodynamic processes of metal nanoclusters enabled biorealistic synapses and leaky-integrate-and-fire neurons. Mater Horiz, 2020, 7: 71–81
49 Dan Y, Poo M. Spike timing-dependent plasticity of neural circuits. Neuron, 2004, 44: 23–30
50 van Rossum M C W, Bi G Q, Turrigiano G G. Stable Hebbian learning from spike timing-dependent plasticity. J Neurosci, 2000, 20: 8812–8821
51 Caporale N, Dan Y. Spike timing-dependent plasticity: a Hebbian learning rule. Annu Rev Neurosci, 2008, 31: 25–46
52 Lu Y, Liu K, Yang J, et al. Highly uniform two-terminal artificial synapses based on polycrystalline $Hf_{0.5}Zr_{0.5}O2$ for sparsified back propagation networks. Adv Electron Mater, 2020, 6: 2000204
53 Cheng C, Li Y, Zhang T, et al. Bipolar to unipolar mode transition and imitation of metaplasticity in oxide based memristors with enhanced ionic conductivity. J Appl Phys, 2018, 124: 152103
54 Abraham W C. Metaplasticity: tuning synapses and networks for plasticity. Nat Rev Neurosci, 2008, 9: 387
55 Tan Z H, Yang R, Terabe K, et al. Synaptic metaplasticity realized in oxide memristive devices. Adv Mater, 2016, 28: 377–384
56 Hao Y, Xiang S Y, Han G, et al. Recent progress of integrated circuits and optoelectronic chips. Sci China Inf Sci, 2021, 64: 201401
57 Tan H, Liu G, Zhu X, et al. An optoelectronic resistive switching memory with integrated demodulating and arithmetic functions. Adv Mater, 2015, 27: 2797–2803
58 Ye C, Peng Q, Li M, et al. Multilevel conductance switching of memory device through photoelectric effect. J Am Chem Soc, 2012, 134: 20053–20059
59 Dang B, Ma L, Yan L, et al. Physically transient optic-neural synapse for secure in-sensor computing. IEEE Electron Device Lett, 2020, 41: 1641–1644
60 Srikant V, Clarke D R. On the optical band gap of zinc oxide. J Appl Phys, 1998, 83: 5447–5451
61 Seo S, Jo S H, Kim S, et al. Artificial optic-neural synapse for colored and color-mixed pattern recognition. Nat Commun, 2018, 9: 5106
62 Yang Y, Yin M, Yu Z, et al. Multifunctional nanoionic devices enabling simultaneous heterosynaptic plasticity and efficient in-memory boolean logic. Adv Electron Mater, 2017, 3: 1700032
63 Lai S, Lowrey T. OUM-A 180 nm nonvolatile memory cell element technology for stand alone and embedded applications. In: Proceedings of International Electron Devices Meeting, 2001. 1–4

64 Rao F, Ding K, Zhou Y, et al. Reducing the stochasticity of crystal nucleation to enable subnanosecond memory writing. Science, 2017, 358: 1423–1427
65 Im D H, Lee J I, Cho S L, et al. A unified 7.5 nm dash-type confined cell for high performance PRAM device. In: Proceedings of IEEE International Electron Devices Meeting, 2008. 1–4
66 Kolobov A V, Fons P, Frenkel A I, et al. Understanding the phase-change mechanism of rewritable optical media. Nat Mater, 2004, 3: 703–708
67 Khwa W S, Wu J Y, Su T H, et al. A novel inspection and annealing procedure to rejuvenate phase change memory from cycling-induced degradations for storage class memory applications. In: Proceedings of IEEE International Electron Devices Meeting, 2014. 1–4
68 Song Z T, Cai D L, Li X, et al. High endurance phase change memory chip implemented based on carbon-doped $Ge_2Sb_2Te_5$ in 40 nm node for embedded application. In: Proceedings of IEEE International Electron Devices Meeting, 2018. 1–4
69 Lu Y M, Li X, Yan L H, et al. Accelerated local training of CNNs by optimized direct feedback alignment based on stochasticity of 4 Mb C-doped $Ge_2Sb_2Te_5$ PCM chip in 40 nm node. In: Proceedings of IEEE International Electron Devices Meeting, 2020. 1–4
70 Garcia V, Fusil S, Bouzehouane K, et al. Giant tunnel electroresistance for non-destructive readout of ferroelectric states. Nature, 2009, 460: 81–84
71 Pantel D, Goetze S, Hesse D, et al. Room-temperature ferroelectric resistive switching in ultrathin $Pb(Zr_{0.2}Ti_{0.8})O_3$ films. ACS Nano, 2011, 5: 6032–6038
72 Li Z, Guo X, Lu H B, et al. An epitaxial ferroelectric tunnel junction on silicon. Adv Mater, 2014, 26: 7185–7189
73 Chang P Y, Du G, Liu X Y. Design space for stabilized negative capacitance in HfO2 ferroelectric-dielectric stacks based on phase field simulation. Sci China Inf Sci, 2021, 64: 122402
74 Park M H, Lee Y H, Kim H J, et al. Ferroelectricity and antiferroelectricity of doped thin HfO_2-based films. Adv Mater, 2015, 27: 1811–1831
75 Yoong H Y, Wu H, Zhao J, et al. Epitaxial ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ thin films and their implementations in memristors for brain-inspired computing. Adv Funct Mater, 2018, 28: 1806037
76 Mikheev V, Chouprik A, Lebedinskii Y, et al. Memristor with a ferroelectric HfO_2 layer: in which case it is a ferroelectric tunnel junction. Nanotechnology, 2020, 31: 215205
77 Chen C, Yang M, Liu S, et al. Bio-inspired neurons based on novel leaky-FeFET with ultra-low hardware cost and advanced functionality for all-ferroelectric neural network. In: Proceedings of Symposium on VLSI Technology, 2019. 136–137
78 Pirrotta O, Larcher L, Lanza M, et al. Leakage current through the poly-crystalline HfO_2: trap densities at grains and grain boundaries. J Appl Phys, 2013, 114: 134503
79 Luo Q, Cheng Y, Yang J, et al. A highly CMOS compatible hafnia-based ferroelectric diode. Nat Commun, 2020, 11: 1391
80 Seo J, Brezzo B, Liu Y, et al. A 45 nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons. In: Proceedings of IEEE Custom Integrated Circuits Conference (CICC), 2011. 1–4
81 Indiveri G, Linares-Barranco B, Hamilton T J, et al. Neuromorphic silicon neuron circuits. Front Neurosci, 2011, 5: 73
82 Liu Y H, Wang X J. Spike-frequency adaptation of a generalized leaky integrate-and-fire model neuron. J Comput Neuroscience, 2001, 10: 25–45
83 Andrade R, Foehring R C, Tzingounis A V. The calcium-activated slow AHP: cutting through the Gordian knot. Front Cell Neurosci, 2012, 6: 47
84 Sharpee T O, Sugihara H, Kurgansky A V, et al. Adaptive filtering enhances information transmission in visual cortex. Nature, 2006, 439: 936–942
85 Luo J, Yu L, Liu T, et al. Capacitor-less stochastic leaky-FeFET neuron of both excitatory and inhibitory connections for SNN with reduced hardware cost. In: Proceedings of IEEE International Electron Devices Meeting, 2019. 1–4
86 Ali T, Polakowski P, Riedel S, et al. High endurance ferroelectric hafnium oxide-based FeFET memory without retention penalty. IEEE Trans Electron Devices, 2018, 65: 3769–3774
87 García H, Dueñas S, Castán H, et al. Influence of interlayer trapping and detrapping mechanisms on the electrical characterization of hafnium oxide/silicon nitride stacks on silicon. J Appl Phys, 2008, 104: 094107
88 Ali T, Polakowski P, Kähnel K, et al. A multilevel FeFET memory device based on laminated HSO and HZO ferroelectric layers for high-density storage. In: Proceedings of IEEE International Electron Devices Meeting, 2019. 1–4
89 Muller J, Polakowski P, Paul J, et al. Integration challenges of ferroelectric hafnium oxide based embedded memory. ECS Trans, 2015, 69: 85–95
90 Sarkar D, Xie X, Liu W, et al. A subthermionic tunnel field-effect transistor with an atomically thin channel. Nature, 2015, 526: 91–95
91 Wang Q H, Kalantar-Zadeh K, Kis A, et al. Electronics and optoelectronics of two-dimensional transition metal dichalcogenides. Nat Nanotech, 2012, 7: 699–712
92 Das S, Robinson J A, Dubey M, et al. Beyond graphene: progress in novel two-dimensional materials and van der Waals solids. Annu Rev Mater Res, 2015, 45: 1–27
93 Yang H, Xiao M, Cui Y, et al. Nonvolatile memristor based on heterostructure of 2D room-temperature ferroelectric α-In_2Se_3 and WSe_2. Sci China Inf Sci, 2019, 62: 220404
94 Tian H, Guo Q, Xie Y, et al. Anisotropic black phosphorus synaptic device for neuromorphic applications. Adv Mater, 2016, 28: 4991–4997
95 Xia F, Wang H, Jia Y. Rediscovering black phosphorus as an anisotropic layered material for optoelectronics and electronics. Nat Commun, 2014, 5: 4458
96 Schulman D S, Arnold A J, Das S. Contact engineering for 2D materials and devices. Chem Soc Rev, 2018, 47: 3037–3058
97 Zhu J, Yang Y, Jia R, et al. Ion gated synaptic transistors based on 2D van der Waals crystals with tunable diffusive dynamics. Adv Mater, 2018, 30: 1800195
98 Upadhyayula L C, Loferski J J, Wold A, et al. Semiconducting properties of single crystals of n- and p-type tungsten diselenide (WSe_2). J Appl Phys, 1968, 39: 4736–4740
99 Kuzminskii Y V, Voronin B M, Redin N N. Iron and nickel phosphorus trisulfides as electroactive materials for primary lithium batteries. J Power Sources, 1995, 55: 133–141
100 Barj M, Sourisseau C, Ouvrard G, et al. Infrared studies of lithium intercalation in the $FePS_3$ and $NiPS_3$ layer-type compounds. Solid State Ion, 1983, 11: 179–183
101 Bao L, Zhu J, Yu Z, et al. Dual-gated MoS_2 neuristor for neuromorphic computing. ACS Appl Mater Interfaces, 2019, 11: 41482–41489

102 Pan L, Ji Z, Yi X, et al. Metal-organic framework nanofilm for mechanically flexible information storage applications. Adv Funct Mater, 2015, 25: 2677–2685
103 Fang Y K, Liu C L, Li C, et al. Synthesis, morphology, and properties of poly(3-hexylthiophene)-block-poly(vinylphenyl oxadiazole) donor-acceptor rod-coil block copolymers and their memory device applications. Adv Funct Mater, 2010, 20: 3012–3024
104 Ji Y, Cho B, Song S, et al. Stable switching characteristics of organic nonvolatile memory on a bent flexible substrate. Adv Mater, 2010, 22: 3071–3075
105 Hsu J M, Rieth L, Normann R A, et al. Encapsulation of an integrated neural interface device with Parylene C. IEEE Trans Biome Eng, 2008, 56: 23–29
106 Kahouli A, Sylvestre A, Ortega L, et al. Structural and dielectric study of parylene C thin films. Appl Phys Lett, 2009, 94: 152901
107 Cai Y, Tan J, YeFan L, et al. A flexible organic resistance memory device for wearable biomedical applications. Nanotechnology, 2016, 27: 275206
108 Lin M, Chen Q, Wang Z, et al. Flexible polymer device based on parylene-C with memory and temperature sensing functionalities. Polymers, 2017, 9: 310
109 Zhang Z, Wang Z, Shi T, et al. Memory materials and devices: from concept to application. InfoMat, 2020, 2: 261–290
110 Schönhals A, Rosário C M M, Hoffmann-Eifert S, et al. Role of the electrode material on the RESET limitation in oxide ReRAM devices. Adv Electron Mater, 2018, 4: 1700243
111 Valov I, Waser R, Jameson J R, et al. Electrochemical metallization memories–fundamentals, applications, prospects. Nanotechnology, 2011, 22: 254003
112 Lim E W, Ismail R. Conduction mechanism of valence change resistive switching memory: a survey. Electronics, 2015, 4: 586–613
113 Moors M, Adepalli K K, Lu Q, et al. Resistive switching mechanisms on TaO_x and $SrRuO_3$ thin-film surfaces probed by scanning tunneling microscopy. ACS Nano, 2016, 10: 1481–1492
114 Yang J J, Inoue I H, Mikolajick T, et al. Metal oxide memories based on thermochemical and valence change mechanisms. MRS Bull, 2012, 37: 131–137
115 Yang J J, Pickett M D, Li X, et al. Memristive switching mechanism for metal/oxide/metal nanodevices. Nat Nanotech, 2008, 3: 429–433
116 Ielmini D. Resistive switching memories based on metal oxides: mechanisms, reliability and scaling. Semicond Sci Technol, 2016, 31: 063002
117 Akinaga H, Shima H. Resistive random access memory (ReRAM) based on metal oxides. Proc IEEE, 2010, 98: 2237–2251
118 Lee J, Lu W D. On-demand reconfiguration of nanomaterials: when electronics meets ionics. Adv Mater, 2018, 30: 1702770
119 Grundmeier G, Schmidt W, Stratmann M. Corrosion protection by organic coatings: electrochemical mechanism and novel methods of investigation. Electrochim Acta, 2000, 45: 2515–2533
120 Liu Q, Long S, Lv H, et al. Controllable growth of nanoscale conductive filaments in solid-electrolyte-based ReRAM by using a metal nanocrystal covered bottom electrode. ACS Nano, 2010, 4: 6162–6168
121 Yang Y, Gao P, Li L, et al. Electrochemical dynamics of nanoscale metallic inclusions in dielectrics. Nat Commun, 2014, 5: 4232
122 Yang Y, Zhang X, Qin L, et al. Probing nanoscale oxygen ion motion in memristive systems. Nat Commun, 2017, 8: 1–10
123 Liu K, Qin L, Zhang X, et al. Interfacial redox processes in memristive devices based on valence change and electrochemical metallization. Faraday Discuss, 2019, 213: 41–52
124 Kwon D H, Kim K M, Jang J H, et al. Atomic structure of conducting nanofilaments in TiO_2 resistive switching memory. Nat Nanotech, 2010, 5: 148–153
125 Valov I, Linn E, Tappertzhofen S, et al. Nanobatteries in redox-based resistive switches require extension of memristor theory. Nat Commun, 2013, 4: 1771
126 Tappertzhofen S, Valov I, Tsuruoka T, et al. Generic relevance of counter charges for cation-based nanoscale resistive switching memories. ACS Nano, 2013, 7: 6396–6402
127 Guan W, Long S, Liu Q, et al. Nonpolar nonvolatile resistive switching in Cu Doped ZrO_2. IEEE Electron Device Lett, 2008, 29: 434–437
128 Chae S C, Lee J S, Kim S, et al. Random circuit breaker network model for unipolar resistance switching. Adv Mater, 2008, 20: 1154–1159
129 Strukov D B, Alibart F, Williams R S. Thermophoresis/diffusion as a plausible mechanism for unipolar resistive switching in metal-oxide-metal memristors. Appl Phys A, 2012, 107: 509–518
130 Murgatroyd P N. Theory of space-charge-limited current enhanced by Frenkel effect. J Phys D-Appl Phys, 1970, 3: 151–156
131 Liu Q, Liu Z, Zhang X, et al. Organic photovoltaic cells based on an acceptor of soluble graphene. Appl Phys Lett, 2008, 92: 223303
132 Carbone A, Kotowska B K, Kotowski D. Space-charge-limited current fluctuations in organic semiconductors. Phys Rev Lett, 2005, 95: 236601
133 Hill R M. Poole-Frenkel conduction in amorphous solids. Philos Mag, 1971, 23: 59–86
134 Kim W, Park S I, Zhang Z, et al. Current conduction mechanism of nitrogen-doped AlO_x RRAM. IEEE Trans Electron Dev, 2014, 61: 2158–2163
135 Jeong D S, Hwang C S. Tunneling-assisted Poole-Frenkel conduction mechanism in HfO_2 thin films. J Appl Phys, 2005, 98: 113701
136 Chen Y C, Chen C F, Chen C T, et al. An access-transistor-free (0T/1R) non-volatile resistance random access memory (RRAM) using a novel threshold switching, self-rectifying chalcogenide device. In: Proceedings of IEEE International Electron Devices Meeting, 2003. 1–4
137 Mazumder P, Kang S M, Waser R. Memristors: devices, models, and applications. Proc IEEE, 2012, 100: 1911–1919
138 Jiang W, Xie B, Liu C C, et al. Integrating memristors and CMOS for better AI. Nat Electron, 2019, 2: 376–377
139 Cai F, Correll J M, Lee S H, et al. A fully integrated reprogrammable memristor-CMOS system for efficient multiply-accumulate operations. Nat Electron, 2019, 2: 290–299
140 Yang X, Fang Y, Yu Z, et al. Nonassociative learning implementation by a single memristor-based multi-terminal synaptic device. Nanoscale, 2016, 8: 18897–18904
141 Scott J C. Is there an immortal memory? Science, 2004, 304: 62–63
142 Baek I G, Kim D C, Lee M J, et al. Multi-layer cross-point binary oxide resistive memory (OxRRAM) for post-NAND

storage application. In: Proceedings of IEEE International Electron Devices Meeting, 2005. 750–753
143 Seok J Y, Song S J, Yoon J H, et al. A review of three-dimensional resistive switching cross-bar array memories from the integration and materials property points of view. Adv Funct Mater, 2014, 24: 5316–5339
144 Hsu C W, Wang I T, Lo C L, et al. Self-rectifying bipolar TaO_x/TiO_2 RRAM with superior endurance over 1012 cycles for 3D high-density storage-class memory. In: Proceedings of Symposium on VLSI Technology, 2013. 166–167
145 Bai Y, Wu H, Wang K, et al. Stacked 3D RRAM array with graphene/CNT as edge electrodes. Sci Rep, 2015, 5: 13785
146 Yoon H S, Baek I G, Zhao J, et al. Vertical cross-point resistance change memory for ultra-high density non-volatile memory applications. In: Proceedings of Symposium on VLSI Technology, 2009. 26–27
147 Deng Y, Chen H Y, Gao B, et al. Design and optimization methodology for 3D RRAM arrays. In: Proceedings of IEEE International Electron Devices Meeting, 2013. 1–4
148 Yu M, Fang Y, Wang Z, et al. Encapsulation layer design and scalability in encapsulated vertical 3D RRAM. Nanotechnology, 2016, 27: 205202
149 Lee S M, Cahill D G. Heat transport in thin dielectric films. J Appl Phys, 1997, 81: 2590–2595
150 Chen Y S, Lee H Y, Chen P S, et al. Good endurance and memory window for Ti/HfO_x pillar RRAM at 50-nm scale by optimal encapsulation layer. IEEE Electron Device Lett, 2011, 32: 390–392
151 Chen Q, Wang Z, Yu M, et al. Thermal effect in ultra-high density 3D vertical and horizontal RRAM array. Phys Scr, 2019, 94: 045001
152 Li S, Niu D, Malladi K T, et al. Drisa: a DRAM-based reconfigurable in-situ accelerator. In: Proceedings of the 50th Annual IEEE/ACM International Symposium Microarchit (MICRO), 2017. 288–301
153 Seshadri V, Lee D, Mullins T, et al. Ambit: in-memory accelerator for bulk bitwise operations using commodity DRAM technology. In: Proceedings of the 50th Annual IEEE/ACM International Symposium Microarchit (MICRO), 2017. 273–287
154 Yin S, Jiang Z, Seo J S, et al. XNOR-SRAM: in-memory computing SRAM Macro for binary/ternary deep neural networks. IEEE J Solid-State Circ, 2020, 55: 1733–1743
155 Irom F, Nguyen D N. Single event effect characterization of high density commercial NAND and NOR nonvolatile flash memories. IEEE Trans Nucl Sci, 2007, 54: 2547–2553
156 Xiang Y, Huang P, Han R, et al. Hardware implementation of energy efficient deep learning neural network based on nanoscale flash computing array. Adv Mater Technol, 2019, 4: 1800720
157 Chiu F C. A review on conduction mechanisms in dielectric films. Adv Mater Sci Eng, 2014, 2014: 1–18
158 Cheong W, Yoon C, Woo S, et al. A flash memory controller for 15 μs ultra-low-latency SSD using high-speed 3D NAND flash with 3 μs read time. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2018. 338–340
159 Miura N, Take Y, Saito M, et al. A 2.7 $Gb/s/mm_2$ 0.9 pJ/b/chip 1coil/channel ThruChip interface with coupled-resonator-based CDR for NAND Flash memory stacking. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2011. 490–492
160 Guo X, Bayat F M, Bavandpour M, et al. Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology. In: Proceedings of IEEE International Electron Devices Meeting, 2017. 1–4
161 Bayat F M, Guo X, Klachko M, et al. Model-based high-precision tuning of NOR flash memory cells for analog computing applications. In: Proceedings of the 74th Annual Device Research Conference (DRC), 2016. 1–2
162 Mahmoodi M R, Strukov D. An ultra-low energy internally analog, externally digital vector-matrix multiplier based on NOR flash memory technology. In: Proceedings of the 55th ACM/ESDA/IEEE Design Automation Conference (DAC), 2018. 1–6
163 Han R, Huang P, Xiang Y, et al. A novel convolution computing paradigm based on NOR flash array with high computing speed and energy efficiency. IEEE Trans Circ Syst I, 2019, 66: 1692–1703
164 Lee S T, Lee J H. Neuromorphic computing using NAND flash memory architecture with pulse width modulation scheme. Front Neurosci, 2020, 14: 571292
165 Govoreanu B, Kar G S, Chen Y Y, et al. 10×10 nm^2 Hf/HfO_x crossbar resistive RAM with excellent performance, reliability and low-energy operation. In: Proceedings of IEEE International Electron Devices Meeting, 2011. 1–4
166 Torrezan A C, Strachan J P, Medeiros-Ribeiro G, et al. Sub-nanosecond switching of a tantalum oxide memristor. Nanotechnology, 2011, 22: 485203
167 Xiong F, Liao A D, Estrada D, et al. Low-power switching of phase-change materials with carbon nanotube electrodes. Science, 2011, 332: 568–570
168 Florent K, Pesic M, Subirats A, et al. Vertical ferroelectric HfO_2 FET based on 3-D NAND architecture: towards dense low-power memory. In: Proceedings of IEEE International Electron Devices Meeting, 2018. 1–4
169 Dünkel S, Trentzsch M, Richter R, et al. A FeFET based super-low-power ultra-fast embedded NVM technology for 22 nm FDSOI and beyond. In: Proceedings of IEEE International Electron Devices Meeting, 2017. 1–4
170 Zhang F, Zhang H, Shrestha P R, et al. An ultra-fast multi-level $MoTe_2$-based RRAM. In: Proceedings of IEEE International Electron Devices Meeting, 2018. 1–4
171 Xu W, Min S Y, Hwang H, et al. Organic core-sheath nanowire artificial synapses with femtojoule energy consumption. Sci Adv, 2016, 2: e1501326
172 Huang R, Cai Y, Liu Y, et al. Resistive switching in organic memory devices for flexible applications. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2014. 838–841
173 Yan B N, Chen Y R, Li H. Challenges of memristor based neuromorphic computing system. Sci China Inf Sci, 2018, 61: 060425
174 Zhou Y, Wu H, Gao B, et al. Associative memory for image recovery with a high-performance memristor array. Adv Funct Mater, 2019, 29: 1900155
175 Gao B, Chen B, Zhang F, et al. A novel defect-engineering-based implementation for high-performance multilevel data storage in resistive switching memory. IEEE Trans Electron Dev, 2013, 60: 1379–1383
176 Kim W, Menzel S, Wouters D J, et al. 3-bit multilevel switching by deep reset phenomenon in Pt/W/TaOX/Pt-ReRAM devices. IEEE Electron Device Lett, 2016, 37: 564–567
177 Li J, Duan Q, Zhang T, et al. Tuning analog resistive switching and plasticity in bilayer transition metal oxide based memristive synapses. RSC Adv, 2017, 7: 43132–43140
178 Ren P, Wang R, Ji Z, et al. New insights into the design for end-of-life variability of NBTI in scaled high-κ/metal-gate technology for the nano-reliability era. In: Proceedings of IEEE International Electron Devices Meeting, 2014. 1–4
179 Fang Y, Yu Z, Wang Z, et al. Improvement of HfO_x-based RRAM device variation by inserting ALD TiN buffer layer. IEEE Electron Device Lett, 2018, 39: 819–822

180 Kim S, Choi S H, Lu W. Comprehensive physical model of dynamic resistive switching in an oxide memristor. ACS Nano, 2014, 8: 2369–2376
181 Bi G, Poo M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J Neurosci, 1998, 18: 10464–10472
182 Wang Z, Yin M, Zhang T, et al. Engineering incremental resistive switching in TaO_x based memristors for brain-inspired computing. Nanoscale, 2016, 8: 14015–14022
183 Kannan S, Rajendran J, Karri R, et al. Sneak-path testing of crossbar-based nonvolatile random access memories. IEEE Trans Nanotechnol, 2013, 12: 413–426
184 Zidan M A, Fahmy H A H, Hussain M M, et al. Memristor-based memory: the sneak paths problem and solutions. Microelectron J, 2013, 44: 176–183
185 Fang Y C, Wang Z W, Cheng C D, et al. Investigation of NbO_x-based volatile switching device with self-rectifying characteristics. Sci China Inf Sci, 2019, 62: 229401
186 Sun P, Lu N, Li L, et al. Thermal crosstalk in 3-dimensional RRAM crossbar array. Sci Rep, 2015, 5: 13504
187 Kim W, Rösgen B, Breuer T, et al. Nonlinearity analysis of TaO_x redox-based RRAM. Microelectron Eng, 2016, 154: 38–41
188 Wang Z, Kang J, Yu Z, et al. Modulation of nonlinear resistive switching behavior of a TaO_x-based resistive device through interface engineering. Nanotechnology, 2017, 28: 055204
189 Linn E, Rosezin R, Kügeler C, et al. Complementary resistive switches for passive nanocrossbar memories. Nat Mater, 2010, 9: 403–406
190 Huang Y, Huang R, Pan Y, et al. A new dynamic selector based on the bipolar RRAM for the crossbar array application. IEEE Trans Electron Devices, 2012, 59: 2277–2280
191 Yu M, Fang Y, Wang Z, et al. Self-selecodulation of TaO_x resistive switching random access memory with bottom electrode of highly doped Si. J Appl Phys, 2016, 119: 195302
192 Wang Z, Kang J, Bai G, et al. Self-selective resistive device with hybrid switching mode for passive crossbar memory application. IEEE Electron Device Lett, 2020, 41: 1009–1012
193 Dönges S A, Khatib O, O'Callahan B T, et al. Ultrafast nanoimaging of the photoinduced phase transition dynamics in VO_2. Nano Lett, 2016, 16: 3029–3035
194 Wang Z, Rao M, Midya R, et al. Threshold switching of Ag or Cu in dielectrics: materials, mechanism, and applications. Adv Funct Mater, 2018, 28: 1704862
195 Chen Q, Lin M, Wang Z, et al. Low power parylene-based memristors with a graphene barrier layer for flexible electronics applications. Adv Electron Mater, 2019, 5: 1800852
196 Chi P, Li S, Xu C, et al. PRIME: a novel processing-in-memory architecture for neural network computation in ReRAM-based main memory. ACM SIGARCH Comput Archit News, 2016, 44: 27–39
197 Chen Y H, Krishna T, Emer J S, et al. Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE J Solid-State Circ, 2017, 52: 127–138
198 Manukian H, Traversa F L, Di Ventra M. Accelerating deep learning with memcomputing. Neural Networks, 2019, 110: 1–7
199 Krogh A. What are artificial neural networks? Nat Biotechnol, 2008, 26: 195–197
200 Krenker A, Bešter J, Kos A. Introduction to the artificial neural networks. In: Artificial Neural Networks: Methodological Advances and Biomedical Applications. Rijeka: InTech, 2011. 1–18
201 Minsky M, Papert S A. Perceptrons: An Introduction to Computational Geometry. Cambridge: MIT Press, 2017
202 Sutskever I, Martens J, Hinton G E. Generating text with recurrent neural networks. In: Proceedings of International Conference on Machine Learning, 2011
203 Schuster M, Paliwal K K. Bidirectional recurrent neural networks. IEEE Trans Signal Process, 1997, 45: 2673–2681
204 Abdel-Hamid O, Deng L, Yu D. Exploring convolutional neural network structures and optimization techniques for speech recognition. Interspeech. 2013, 11: 73–75
205 Dumoulin V, Visin F. A guide to convolution arithmetic for deep learning. 2016. ArXiv:1603.07285
206 Yao P, Wu H, Gao B, et al. Face classification using electronic synapses. Nat Commun, 2017, 8: 15199
207 Li C, Belkin D, Li Y, et al. Efficient and self-adaptive in-situ learning in multilayer memristor neural networks. Nat Commun, 2018, 9: 2385
208 Ambrogio S, Narayanan P, Tsai H, et al. Equivalent-accuracy accelerated neural-network training using analogue memory. Nature, 2018, 558: 60–67
209 Wang Z, Zheng Q, Kang J, et al. Self-activation neural network based on self-selective memory device with rectified multilevel states. IEEE Trans Electron Dev, 2020, 67: 4166–4171
210 Zheng Q, Wang Z, Gong N, et al. Artificial neural network based on doped HfO_2 ferroelectric capacitors with multilevel characteristics. IEEE Electron Dev Lett, 2019, 40: 1309–1312
211 Kang J, Yu Z, Wu L, et al. Time-dependent variability in RRAM-based analog neuromorphic system for pattern recognition. In: Proceedings of IEEE International Electron Devices Meeting, 2017. 1–4
212 Yu Z, Wang Z, Kang J, et al. Early-stage fluctuation in low-power analog resistive memory: impacts on neural network and mitigation approach. IEEE Electron Dev Lett, 2020, 41: 940–943
213 Boureau Y L, Ponce J, LeCun Y. A theoretical analysis of feature pooling in visual recognition. In: Proceedings of International Conference on Machine Learning, 2010. 111–118
214 Wang Z, Li C, Lin P, et al. In situ training of feed-forward and recurrent convolutional memristor networks. Nat Mach Intell, 2019, 1: 434–442
215 Yao P, Wu H, Gao B, et al. Fully hardware-implemented memristor convolutional neural network. Nature, 2020, 577: 641–646
216 Li C, Wang Z, Rao M, et al. Long short-term memory networks in memristor crossbar arrays. Nat Mach Intell, 2019, 1: 49–57
217 Wang Y, Yu L, Wu S, et al. Memristor-based biologically plausible memory based on discrete and continuous attractor networks for neuromorphic systems. Adv Intell Syst, 2020, 2: 2000001
218 Yang K, Duan Q, Wang Y, et al. Transiently chaotic simulated annealing based on intrinsic nonlinearity of memristors for efficient solution of optimization problems. Sci Adv, 2020, 6: eaba9901
219 Cai F, Kumar S, van Vaerenbergh T, et al. Power-efficient combinatorial optimization using intrinsic noise in memristor Hopfield neural networks. Nat Electron, 2020, 3: 409–418
220 Lu J, Wu Z, Zhang X, et al. Quantitatively evaluating the effect of read noise in memristive Hopfield network on solving traveling salesman problem. IEEE Electron Dev Lett, 2020, 41: 1688–1691

221 Ghosh-dastidar S, Adeli H. Spiking neural networks. Int J Neur Syst, 2009, 19: 295–308
222 Wade J J, McDaid L J, Santos J A, et al. SWAT: a spiking neural network training algorithm for classification problems. IEEE Trans Neural Netw, 2010, 21: 1817–1830
223 Xiang Y, Huang P, Han R, et al. Efficient and robust spike-driven deep convolutional neural networks based on NOR flash computing array. IEEE Trans Electron Dev, 2020, 67: 2329–2335
224 Bao L, Wang Z, Yu Z, et al. Rotational pattern recognition by spiking correlated neural network based on dualgated MoS_2 neuristor. Adv Intell Syst, 2020, 2: 2000102
225 Wang Z, Joshi S, Savel'ev S, et al. Fully memristive neural networks for pattern classification with unsupervised learning. Nat Electron, 2018, 1: 137–145
226 Zhang X M, Lu J, Wang Z R, et al. Hybrid memristor-CMOS neurons for in-situ learning in fully hardware memristive spiking neural networks. Sci Bull, 2021, 66: 1624–1633
227 Duan Q, Jing Z, Zou X, et al. Spiking neurons with spatiotemporal dynamics and gain modulation for monolithically integrated memristive neural networks. Nat Commun, 2020, 11: 1–13
228 Zhang X, Wang Z, Song W, et al. Experimental demonstration of conversion-based SNNs with 1T1R Mott neurons for neuromorphic inference. In: Proceedings of IEEE International Electron Devices Meeting, 2019. 1–4
229 Zhang X, Wu Z, Lu J, et al. Fully memristive SNNs with temporal coding for fast and low-power edge computing. In: Proceedings of IEEE International Electron Devices Meeting, 2020. 1–4
230 Dang B, Liu K, Zhu J, et al. Stochastic neuron based on IGZO Schottky diodes for neuromorphic computing. APL Mater, 2019, 7: 071114
231 Borghetti J, Snider G S, Kuekes P J, et al. 'Memristive' switches enable 'stateful' logic operations via material implication. Nature, 2010, 464: 873–876
232 Siemon A, Breuer T, Aslam N, et al. Realization of boolean logic functionality using redox-based memristive devices. Adv Funct Mater, 2015, 25: 6414–6423
233 Linn E, Rosezin R, Tappertzhofen S, et al. Beyond von Neumann-logic operations in passive crossbar arrays alongside memory operations. Nanotechnology, 2012, 23: 305205
234 Xu L, Yuan R, Zhu Z, et al. Memristor-based efficient in-memory logic for cryptologic and arithmetic applications. Adv Mater Technol, 2019, 4: 1900212
235 Yuan R, Ma M, Xu L, et al. Efficient 16 Boolean logic and arithmetic based on bipolar oxide memristors. Sci China Inf Sci, 2020, 63: 202401
236 Damiani E, Di Vimercati S D C, Samarati P. New paradigms for access control in open environments. In: Proceedings of IEEE International Symposium on Signal Processing & Information Technology, 2005. 540–545
237 Sadeghi A R, Naccache D. Towards Hardware-Intrinsic Security. Berlin: Springer, 2010
238 Konstantinou C, Maniatakos M, Saqib F, et al. Cyber-physical systems: a security perspective. In: Proceedings of the 20th IEEE European Test Symposium (ETS), 2015. 1–8
239 Yu M D, Sowell R, Singh A, et al. Performance metrics and empirical results of a PUF cryptographic key generation ASIC. In: Proceedings of IEEE International Symposium on Hardware-oriented Security & Trust, 2012. 108–115
240 Suh G E, Devadas S. Physical unclonable functions for device authentication and secret key generation. In: Proceedings of the 44th ACM/IEEE Design Automation Conference, 2007. 9–14
241 Holcomb D E, Burleson W P, Fu K. Power-up SRAM State as an identifying fingerprint and source of true random numbers. IEEE Trans Comput, 2009, 58: 1198–1210
242 Gao L, Chen P Y, Liu R, et al. Physical unclonable function exploiting sneak paths in resistive cross-point array. IEEE Trans Electron Dev, 2016, 63: 3109–3115
243 Nili H, Adam G C, Hoskins B, et al. Hardware-intrinsic security primitives enabled by analogue state and nonlinear conductance variations in integrated memristors. Nat Electron, 2018, 1: 197–202
244 Jiang H, Li C, Zhang R, et al. A provable key destruction scheme based on memristive crossbar arrays. Nat Electron, 2018, 1: 548–554
245 Gaba S, Sheridan P, Zhou J, et al. Stochastic memristive devices for computing and neuromorphic applications. Nanoscale, 2013, 5: 5872–5878
246 Jiang H, Belkin D, Savel'ev S E, et al. A novel true random number generator based on a stochastic diffusive memristor. Nat Commun, 2017, 8: 882
247 Hamming R W. Error detecting and error correcting codes. Bell Syst Tech J, 1950, 29: 147–160
248 Mstafa R J, Elleithy K M. A highly secure video steganography using Hamming code (7, 4). In: Proceedings of IEEE Long Island Systems, Applications and Technology (LISAT) Conference, 2014. 1–6
249 Sun X, Zhang T, Cheng C, et al. A memristor-based in-memory computing network for hamming code error correction. IEEE Electron Device Lett, 2019, 40: 1080–1083
250 Ben-Romdhane M, Graba T, Danger J L, et al. Design methodology of an ASIC TRNG based on an open-loop delay chain. In: Proceedings of IEEE 11th International New Circuits and Systems Conference (NEWCAS), 2013. 1–4
251 Zhang T, Yin M, Xu C, et al. High-speed true random number generation based on paired memristors for security electronics. Nanotechnology, 2017, 28: 455202
252 Yu S, Guan X, Wong H S P. On the stochastic nature of resistive switching in metal oxide RRAM: physical modeling, Monte Carlo simulation, and experimental characterization. In: Proceedings of International Electron Devices Meeting, 2011. 1–4
253 Dang B, Sun J, Zhang T, et al. Physically transient true random number generators based on paired threshold switches enabling Monte Carlo method applications. IEEE Electron Device Lett, 2019, 40: 1096–1099
254 Xue Y Y, Wang Z J, Chen W, et al. Modeling dark signal of CMOS image sensors irradiated by reactor neutron using Monte Carlo method. Sci China Inf Sci, 2018, 61: 062405