Review Article

Nano Research
2020, 13(5): 1228–1243

ISSN 1998-0124 CN 11-5974/O4
https://doi.org/10.1007/s12274-020-2616-0

Nanoscale resistive switching devices for memory and computing applications

Seung Hwan Lee, Xiaojian Zhu, and Wei D. Lu (✉)

Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan 48109, USA

Received: 25 October 2019 / **Revised:** 3 December 2019 / **Accepted:** 18 December 2019

ABSTRACT

With the slowing down of the Moore's law and fundamental limitations due to the von-Neumann bottleneck, continued improvements in computing hardware performance become increasingly more challenging. Resistive switching (RS) devices are being extensively studied as promising candidates for next generation memory and computing applications due to their fast switching speed, excellent endurance and retention, and scaling and three-dimensional (3D) stacking capability. In particular, RS devices offer the potential to natively emulate the functions and structures of synapses and neurons, allowing them to efficiently implement neural networks (NNs) and other in-memory computing systems for data intensive applications such as machine learning tasks. In this review, we will examine the mechanisms of RS effects and discuss recent progresses in the application of RS devices for memory, deep learning accelerator, and more faithful brain-inspired computing tasks. Challenges and possible solutions at the device, algorithm, and system levels will also be discussed.

KEYWORDS

resistive switching, oxygen vacancy, metal cation, memory application, in-memory computing, bio-inspired application

1 Introduction

Over the last several decades, successive downscaling of the complementary metal-oxide-semiconductor (CMOS) transistors according to Moore's law has contributed to the exponential progress in computing and information technology [1]. The number of transistors per microprocessor chip has doubled about every two years with increased clock speed or rate of executing instructions [2]. A rule of thumb has even dominated the semiconductor industry to deliberately stay on track of the Moore's law. However, Moore's law is finally slowing down. For one, heat generated in the highly scaled integrated chip with extremely fast clock rates has already limited performance improvement and stopped the increase of clock speed a decade ago [3]. Furthermore, transistor size is already approaching the fundamental physical limit of around 2–3 nm beyond which quantum tunneling and quantum uncertainties will make transistors unreliable to operate in conventional circuits [4–6]. Even worse, the performance gap between processing and memory units has increased dramatically, and the data movement between these units in the conventional von-Neumann architecture starts to be the most dominate factor for energy consumption and system throughput [7, 8]. This problem, widely known as the von-Neumann bottleneck, will be further exacerbated in the data-intensive applications such as machine learning tasks. To address these issues, new materials, devices and new computing architectures are now extensively investigated to complement and possibly replace conventional CMOS devices and circuits.

One such device is non-volatile resistive switching (RS) device [9–13]. Instead of using electrons and holes to store data in conventional memory units, RS devices store data by reconfiguring the internal ion (oxygen ion or metal cation) distribution in nanoscale solid state films [14, 15]. A typical RS device stores data in the form of different values of resistance, and has a simple two-terminal resistor-like structure with a functional thin film sandwiched by two electrodes, as schematically shown in Fig. 1(a). Programming the device involves changing the oxygen vacancy (V_O) or metal cation distribution in the switching layer, through field-driven ionic drift and diffusion and electrochemical processes. This type of devices is often called redox memory devices, where the ion migration and associated redox processes can lead to conducting filament (CF) formation or rupture in the switching layer, thus changing the overall device resistance.

RS devices first received broad interest since they can effectively reduce the performance gap between existing memory (i.e. static random-access memory (SRAM) and dynamic random-access memory (DRAM)) and non-volatile data storage (i.e. solid state drive (SSD) or hard disk drive (HDD)) solutions, and can potentially replace them [16–18] due to the fast switching speed [19], low power operation [20], scalability [21, 22] and good reliability [23, 24] offered by RS devices. For instance, the Optane memory announced by Intel and Micron in 2015 is a non-volatile memory based on the broadly defined RS effects. It can be up to 1,000 times faster than today's NAND flash and can thus be used as a storage class memory (SCM) that sits between memory and storage to complement the gaps between these systems in terms of performance and density [25]. Due to the simple two-terminal structure and the ability of three-dimensional (3D) stacking, RS memories can offer very high memory density without relying on extreme scaling.

Address correspondence to wluee@umich.edu

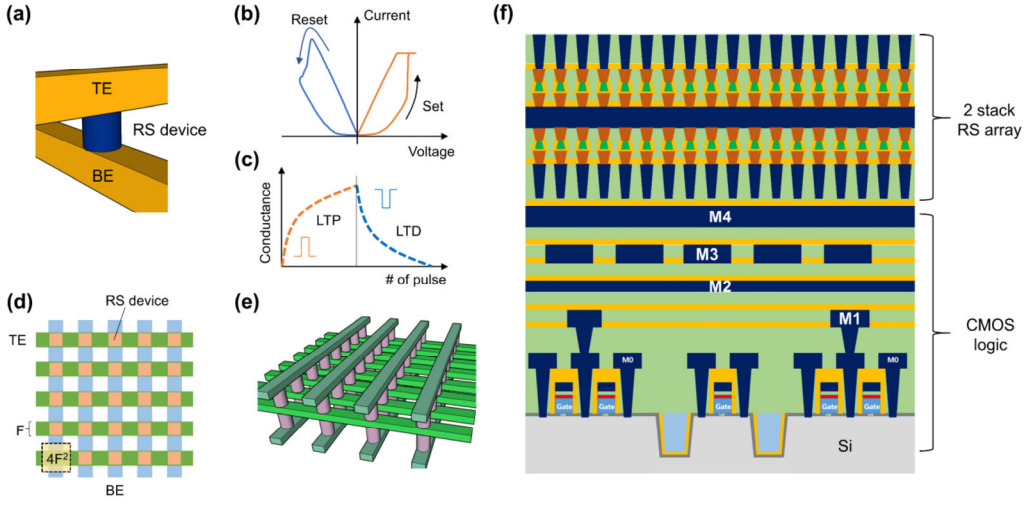


Figure 1 RS devices and integrated systems. (a) Schematic of two-terminal RS device structure, consisting of a switching layer sandwiched by top and bottom electrodes. (b) Typical bipolar resistive switching behavior. After the Set operation, the device is changed to the LRS (ON), while Reset operation leads to the transition to HRS (OFF). (c) LTP and LTD behaviors in a typical RS device. These effects are obtained with successive positive or negative pulses, respectively. (d) A crossbar array structure, in which a device is formed at each crosspoint with cell size of $4F^2$. (e) Schematic of a 2 stack crossbar architecture. The 2D crossbar arrays are stacked on top of each other. (f) Schematic of 3D integration of a 2 stack crossbar array with CMOS circuitry underneath.

Beyond memory applications, recent efforts have also focused on using RS devices and arrays to perform computing directly at the site where data is stored, known as in-memory computing [9, 26–28]. This approach fundamentally addresses the energy consumption and speed issues due to the von-Neumann bottleneck, by eliminating constant data movements between process and memory units. For example, an array of RS devices formed in the crossbar form allows the direct mapping of a neural network structure to the hardware, allowing compute-intensive operations such as vector matrix multiplication (VMM) to be performed natively through Ohm's law and Kirchhoff 's law, in place and in parallel [29–31]. Since VMM is the most used operations in typical machine-learning algorithms, hardware built with RS crossbars can offer orders of magnitude improvements in performance compared with state-of-the-art. Beyond VMM, recent studies have also shown that the internal dynamic ionic processes during RS can be employed to faithfully emulate many biological processes and functions that are critical for learning and memory, allowing even more efficient neuromorphic systems to be implemented using solid-state devices and networks [9, 32, 33].

In this Review Article, we will first examine the switching mechanisms of redox-based RS effects. Afterwards, applications of RS devices for memory and computing applications, including conventional deep neural network (DNN) implementations and more bio-realistic neuromorphic computing systems, will be discussed. Finally, challenges and new concepts to solve device non-ideality and system scaling will be analyzed, with an emphasis on a tiled architecture that can be implemented with practical RS devices and arrays.

2 RS device operation and structure: From single device to 3D integration

RS devices exhibit reversible and nonvolatile resistance changes, upon the applications of properly designed external stimuli [15, 34, 35]. For instance, its resistance can be modulated by voltage and current pulses, switching between two or more levels [36, 37]. In binary memory or logic applications [10, 38], the RS device can be either in the high resistance state (HRS) OFF state, or the low resistance state (LRS) ON state, depending on the applied programming voltage (Fig. 1(b)). The programming operation that changes the RS device state from HRS to LRS is called Set, while the opposite programming operation is called Reset. In addition, an initialization process known as the forming process may be required to condition the device before reliable RS behaviors are obtained [36]. With optimized materials and switching pulse configurations, an RS device can also show multiple resistance states such as 2^n levels, which can further increase the memory density by a factor of n [39–41]. Incremental conductance changes, often characterized as long-term potentiation (LTP) and long-term depression (LTD) effects, can also be achieved in some RS devices, which are often called memristors, by repeatedly applying moderate positive and negative pulse trains to the device (Fig. 1(c)) [32]. These properties are particularly attractive for neuromorphic applications, since having a large number of states allows weight storage in high precision, while the incremental conductance changes enable online learning, respectively.

RS device has a simple form of a two-terminal structure which allows device integration in a crossbar form with high density and connectivity. For instance, an RS device can be formed at each crosspoint between the top (rows) and bottom (columns) electrodes, with an effective cell area of $4F^2$ (F: the smallest feature size), which guarantees the layout with maximum cell area efficiency and memory density in any planar design, as schematically shown in Fig. 1(d). It should be noted that to operate a crossbar array properly, selector devices or RS devices with intrinsic highly-nonlinear I–V behavior are required, in order to control total power dissipation through unselected RS devices, and to minimize parasitic effects caused by line resistance [42]. Recently, several types of two-terminal selector devices with high on/off ratios have been explored for large-scale crossbar array implementations [43–46]. However, adding the selector device to RS device should be carefully analyzed, considering the voltage divider effect between the two devices in the presence of device variability. Otherwise, the read voltage window margin can be significantly affected, and the allowed crossbar array size will be significantly reduced. Integration of RS devices with MOS transistors can be an alternative to achieve a large crossbar array. Connection with a transistor to the RS device in series (1T1R) effectively suppresses the sneak path

current, leading to accurate reading and programming operations by controlling the third terminal, the gate voltage, even in a very large crossbar array [47]. However, the use of transistors can increase the effective cell size, and is also not compatible with 3D stacking which will limit the memory density the system can ultimately achieve.

To further increase the memory density, 3D stackable structures are essential. The 2D crossbar array can be stacked on top of each other in which the storage density increases as the number of stacked layers increases. For example, Fig. 1(e) shows 2 stacks of crossbar arrays, increasing the memory density by a factor of 2. Vertical side-wall structures can also be fabricated, achieving similar increase in memory density, depending on how many cells can be fabricated along a vertical electrode [48–50]. The 3D stackable crossbar arrays or vertical arrays should be fabricated on the CMOS circuitry which can include decoders and sense amplifiers to form a complete memory system [51], or mixed-signal interface, SRAM, and logic circuitry for computing applications (Fig. 1(f)) [52]. In general, RS device fabrication can be compatible with the low temperature back-end of line (BEOL) process. As a result, integration of RS arrays can be achieved in the same fab after the front-end CMOS circuitry fabrication has completed, leading to high wafer yield and low cost.

3 Mechanisms of redox-based RS devices

3.1 Oxygen-ion based RS devices

In an oxide-based resistive switching device, often called valence change memory (VCM) or oxide-based resistive random access memory (oxide-RRAM or ReRAM), RS is caused by the redistribution of oxygen vacancies (V_Os) in the switching layer acting as a solid electrolyte [10, 53]. In general, these devices may consist of two oxide layers, a V_O-rich layer and a V_O-poor layer, sandwiched by a pair of inert electrodes. The V_O-rich layer can be a deposited non-stoichiometric suboxide or formed at the interface between a reactive metal layer and the switching layer (the V_O-poor layer). The V_O-rich layer has a high conductivity and acts as a V_O source during RS. The V_O-poor layer is usually stoichiometric and insulating when deposited. Under high electric field and followed by increased local temperature due to Joule heating, V_Os can migrate from the V_O-rich layer to the V_O-poor layer, leading to the formation of local V_O-rich conduction channels (CFs) and switching the device to the LRS. The reverse process breaks the CFs and switches the device back to the HRS.

Figures 2(a) and 2(b) show results from high-angle annular dark-field (HAADF) scanning transmission electron microscopy (STEM) studies on a Pt/SiO_2/Ta_2O_{5-x} RS device at HRS and LRS, respectively, using an *in-situ* experimental setup [54]. The conduction channel appears as a brighter region in the image, suggesting it contains more atoms with large atomic numbers (Ta in this case) since the intensity in the STEM mode strongly depends on Z (Z: atomic number or proton number) and the atomic density. Horizontal electron energy loss spectroscopy (EELS) line scan analysis also shows that the local oxygen concentration after the Set operation (LRS) is significantly reduced compared to the Reset operation (HRS), verifying the

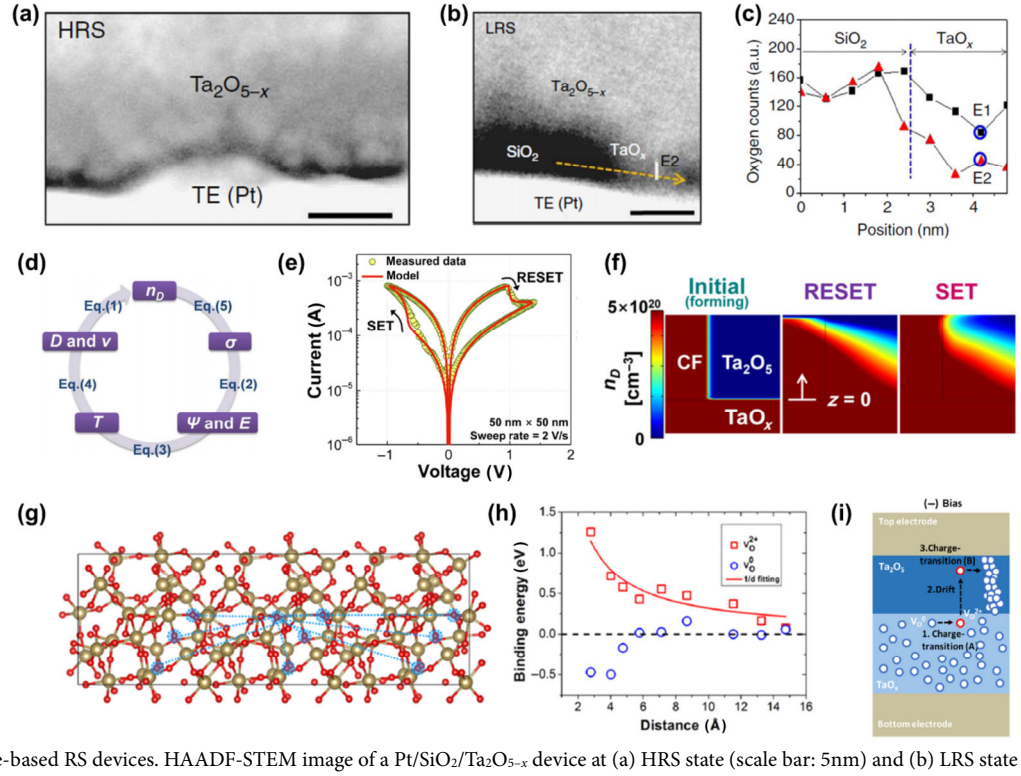


Figure 2 Oxide-based RS devices. HAADF-STEM image of a Pt/SiO_2/Ta_2O_{5-x} device at (a) HRS state (scale bar: 5nm) and (b) LRS state (scale bar: 2 nm). (c) Horizontal EELS line scans with the corresponding oxygen profiles taken at LRS (red symbol) and HRS (black symbol) moving along the arrow shown in (b). Reproduced with permission from Ref. [54], © Springer Nature 2013. (d) Coupled partial differential equations (PDEs) used in a device model that can explain the RS effects. The PDEs are solved self-consistently by a numerical solver (COMSOL). (e) Simulated DC I–V characteristic (solid lines) during RS, showing good agreement with experimental measurements (circles). (f) Simulated 2D maps of V_O concentration (n_D) in the initial state, HRS and LRS. Reproduced with permission from Ref. [55], © American Chemical Society 2014. (g) Atomistic configuration of an amorphous-Ta_2O_5 supercell used to study the interactions of V_O pairs in DFT calculations. Tantalum (Ta) and oxygen (O) atoms are represented by gold and red spheres, respectively. (h) Calculated binding energy of V_O pairs as a function of the distance between the V_Os. Charged V_Os (V_O^{2+}s) show a strong Coulomb repulsion, whereas neutral V_Os (V_O^0S) exhibit a short-range attractive interaction. (i) Illustration of the CF formation processes, including the charge transition and ionic migration steps in a bilayer Ta_2O_5-based RS device. Reproduced with permission from Ref. [60], © American Chemical Society 2019.

role of V_O migration during the operation of oxide-based RS devices (Fig. 2(c)).

Understanding internal dynamics of oxide-based RS devices such as how the filament evolves as V_Os drift and diffuse is essential for continued device optimization and design for both memory and neuromorphic applications [55–58]. Recently, accurate dynamic models have been developed to predict V_O migration with a set of partial differential equations (PDEs) in a self-consistent manner (Fig. 2(d)) [55]. With properly chosen parameters and dynamic equations, the simulated *I*–*V* behavior can match very well with the experimentally measured data, as shown in Fig. 2(e). Here, the filaments correspond to regions with high n_D ($> 5 \times 10^{20}$ cm^{-3}) where the local electrical conduction becomes metallic. In addition, the evolution of V_O concentration (n_D) during RS was successfully captured by the model. For example, the model revealed that the Reset transition creates a small gap (~ 1 nm) with depleted V_O concentration near the TE, thus increasing device resistance. On the other hand, refilling the gap during Set with a negative voltage recovers the CF and results in switching to the LRS (Fig. 2(f)).

The interactions between V_Os and the microscopic processes that are required for stable RS have also been studied. Using first-principles calculations based on the density functional theory (DFT) implemented in the Vienna *ab initio* simulation package (VASP) (Fig. 2(g)) [59], it was found that strong repulsive interaction exists between charged vacancies (V_O^{2+}), while neutral vacancies (V_O^0) experience short-range attraction that facilitates aggregation of the V_Os to form a stable conductive filament (Fig. 2(h)) [60]. This effect explains the long retention time in oxide-RRAM. However, neutral vacancies will not be driven by electric field and will not result in successful Set. To explain this discrepancy, a series of charge transition processes were revealed by the first-principles analysis and experiments to consistently explain both the aggregation effect that leads to stable CFs and the field driven drift of charged V_Os during RS. First, under sufficient bias voltage, isolated V_O^0s near the anode can be ionized to V_O^{2+}s that lead to a lower Fermi level. The charged V_O^{2+}s then migrate toward the cathode through the drift process under the applied high electric field. As the V_O^{2+} concentration is increased near the cathode, a second charge transition process from V_O^{2+} to neural V_O^0 can occur to reduce the system energy. The attractive forces among the neural V_O^0 in turn form a stable CF, completing the Set process (Fig. 2(i)).

3.2 Metal-ion based RS devices

RS effects have also been widely observed in metal-ion based devices, also known as electrochemical metallization (ECM) or conductive bridge random-access memory (CBRAM) devices [14, 36, 61]. The RS effect originates from the electrochemical growth/dissolution of metal (e.g. Ag and Cu) filaments within the insulating layer that acts as the switching layer and plays the role of a solid electrolyte for Ag or Cu cation migration. Typical switching layers for ECM devices include SiO_2 [62, 63], Al_2O_3 [64, 65] or a-Si [66].

During RS, nanoscale metal filaments are formed in the switching layer, through electrochemical processes and ion migration processes. For instance, Ag^+ ions will be first generated from Ag atoms at the anode side through an electrochemical oxidation process under high electric field. Then, the Ag^+ ion will drift along with the electric field and become reduced (from Ag^+ ion to Ag atom) when it reaches a cathode and captures an electron. The reduced Ag atoms form Ag nanoclusters after a nucleation process, eventually leading to a continuous Ag filament and increase of the device conductance.

A key difference between ECM devices and VCM (oxide-RRAM) devices is that in VCM, RS leads to changes in the stoichiometry of the switching material itself, i.e. locally converting a stoichiometric oxide into a sub-oxide as the conducting channel. This mechanism on one hand suggests the devices will have long write/erase endurance since the electrochemical processes are reversible and the species involved in the processes are native to the switching material. On the other hand, the native defects created (V_Os) become very difficult to completely remove during the Reset process, leading to a leaky HRS state after Reset. To the contrary, in ECM, the electrochemical processes involve only the foreign species (e.g. Ag atoms and Ag^+ ions), where the switching film (e.g. SiO_2) does not directly play an active role and chemical reactions may not occur between the filament and the switching layer. As a result, the Reset process of ECM devices can be very clean and allows the device to recover the very insulating HRS state. This leads to very high on/off ratio and low switching energy. On the other hand, since extrinsic species are constantly moving in and out of the switching layer, this process may lead to physical damage to the switching film over time, e.g. in the form of plastic deformation, limiting the device's write/erase endurance when compared with VCM devices.

Ex-situ transmission electron microscopy (TEM) studies were first conducted to observe conducting filaments in Ag/SiO_2/Pt devices based on active Ag metal electrodes [62]. To switch the device from the initial high resistance state, a constant positive voltage was applied to the Ag electrode. The device current level was found to abruptly increase, corresponding to the formation of a conducting (Fig. 3(a)). After the forming process, well-defined conducting filament with a typical cone shape is revealed in the switching layer, highlighted by the arrow in Fig. 3(a). Once a single dominant filament is formed, further filament growth will be suppressed due to the reduced electric field. The partially formed filaments suggest that in this device the filament growth starts from the inert electrode (cathode) side. After a Reset process by applying a negative voltage to the Ag electrode, it is found that all filaments break at the interface between the filament and the inert electrode, and the dissolved parts of the filaments migrated back towards the Ag side (Fig. 3(b)).

The dynamics of the cation (metal ion) transport in the dielectric can depend strongly on the properties of the switching layer. Unlike sputtered SiO_2 which contains a high density of defects and offers fast diffusion paths for cation transport, a-Si based switching layer provides low cation mobility. As a result, filament growth in a-Si based RS devices can start from the active Ag electrode (anode) side instead. *In-situ* TEM studies were performed in a W/a-Si/Ag structure during resistive switching to directly observe the dynamics of filament growth [62]. Positive voltage is applied to the Ag electrode and evolution of current through the device is recorded over time (Fig. 3(c)). Real time observation of filament growth with *in-situ* TEM verified that filaments in this device start growing from the active Ag electrode with a wide base (more Ag clusters) and extended towards the inert electrode, where the filament is composed of discrete Ag metal particles (Figs. 3(d)–3(h)).

Depending on the switching material composition and microstructure, other growth modes have also been observed in the *in-situ* TEM studies [62]. Based on these observations, a universal model that can quantitively capture the filament growth mechanism in metal-ion based RS devices was developed. The model reveals the role different kinetic parameters such as ion mobility (μ) and redox reaction rate (Γ) play during RS [67]. For example, high μ and high Γ will lead to metallic filament

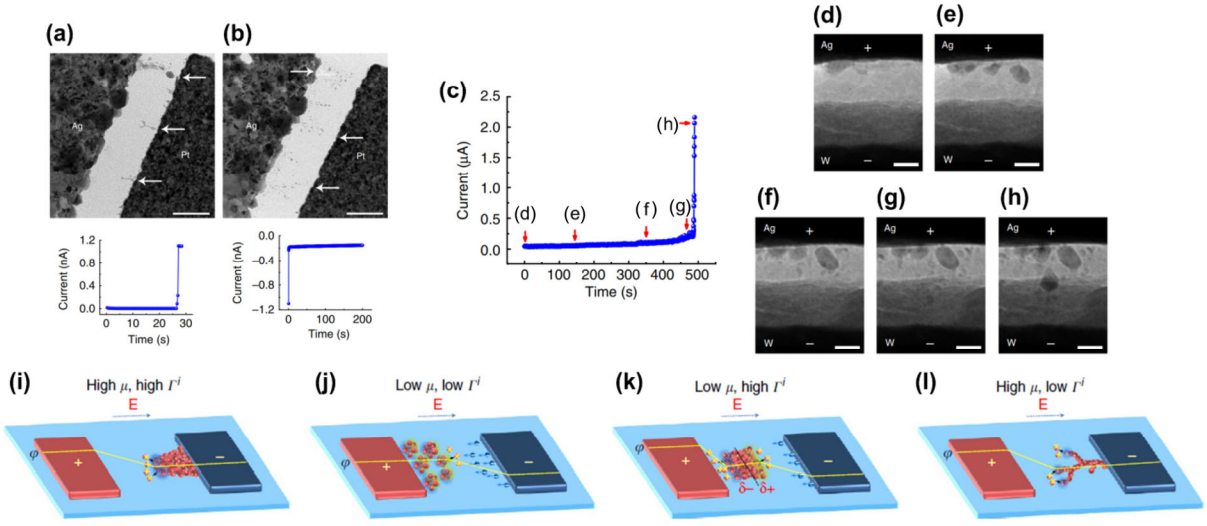


Figure 3 Metal ion-based RS switching mechanism. (a) TEM image of an Ag/SiO_2/Pt device after the forming process, showing conducting filaments. Scale bar: 200 nm. Bottom inset: corresponding *I*–*t* curve during the forming process. (b) TEM image of the same device after erasing. Scale bar: 200 nm. Bottom inset: corresponding *I*–*t* curve during erasing process. (c)–(h) *I*–*t* characteristics recorded during the forming process (c), and *in-situ* TEM images of an Ag/a-Si/W device showing conducting filament growth in the switching layer (a-Si) during the process (d)–(h). Reproduced with permission from Ref. [62], © Springer Nature 2012. (i)–(l) Schematics of filament growth dynamics with different kinetic parameters, including the redox rates (Γ) and ion mobility (μ). (i) High μ and high Γ case, filament growth starts from the inert electrode (cathode), forming a reversed cone shape. (j) Low μ and low Γ case, filament growth starts from the active electrode (anode) with discrete nanoclusters, forming a cone shape. (k) Low μ but high Γ case, filament nucleation occurs inside the dielectric and the clusters eventually re-connect with the anode. (l) High μ but low Γ case, filament growth starts from the cathode, forming a branched structure. Reproduced with permission from Ref. [67], © Springer Nature 2014.

growth from the inert electrode with a reversed cone shape (Fig. 3(i)). In contrast, with low μ and low Γ, filament growth will start from the active electrode (anode) with discrete nanoclusters, forming a typical cone shape (Fig. 3(j)). In the case of low μ but high Γ, filament nucleation occurs inside the solid electrolyte and Set is completed when the filament reconnects with the active electrode (Fig. 3(k)). Finally, with high μ and low Γ, filaments will grow from the cathode, leading to the formation of branched filaments such as the case observed in Ag/SiO_2/Pt devices (Figs. 3(a), 3(b) and 3(l)).

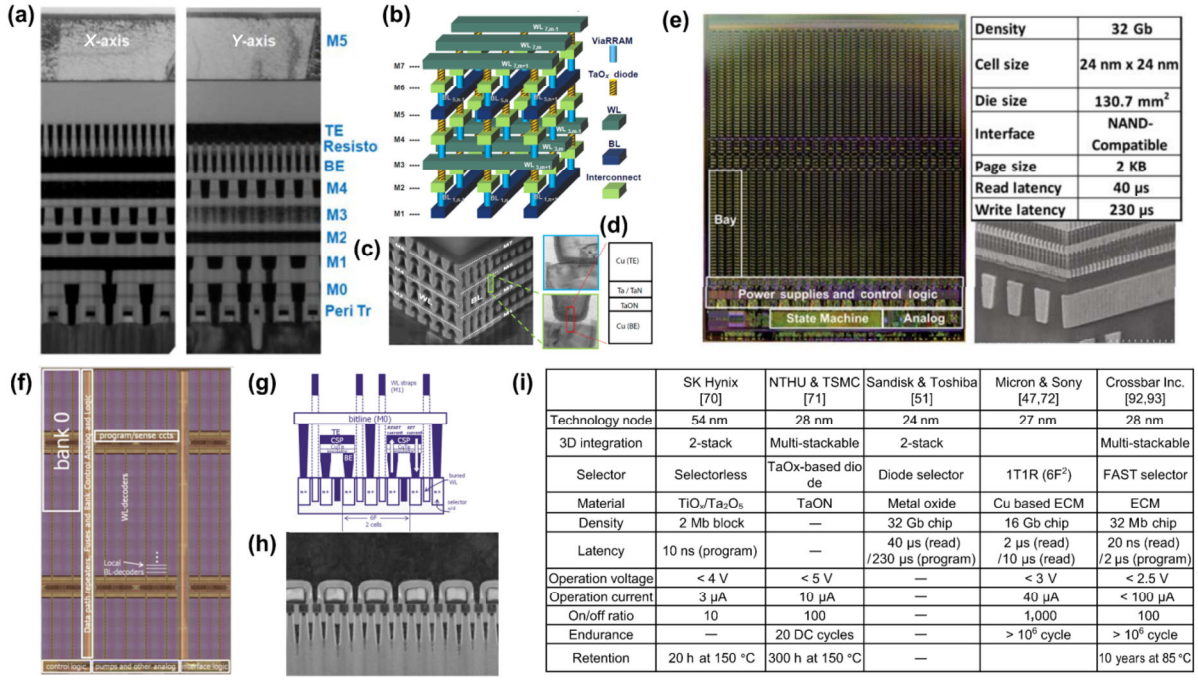
4 Resistive switching memory device applications

4.1 Memory applications

RS devices have been considered as one of the best candidates for future non-volatile memory applications [68, 69]. Unlike charge-based memories, such as dynamic random access memory (DRAM) and NAND flash memory, which suffer from performance degradation as the scaling limit is approached, non-charge based memories including RS devices can offer solutions to extend Moore's law. In particular, ionic RS devices offer outstanding performance specs including scalability, high switching speed, long retention time, high endurance, large on/off ratio, and low power operation. The ability to achieve CMOS-compatible integration with the decoder, sense amplifier, and other peripheral circuitry and form stackable crossbar array structures further makes them attractive in terms of "cost per bit", by maximizing cell area efficiency [70, 71].

In memory applications, the resistance states of RS devices represent a bit ('0' or '1') or multi bits (e.g. 2bits: '00', '01', '10', or '11'), which can be read by sensing current through the RS device during the read operation. In one example, with the 54 nm technology node, a 2 stack and $4F^2$ selector-less crossbar array prototype was implemented by SK Hynix in 2012 [70]. Shared bitline scheme was employed to further increase the cell area efficiency in the multi-layer crossbar array (Fig. 4(a)). TiO_x/Ta_2O_5 oxygen-ion based RS devices with nonlinear *I*–*V* characteristics were used to form the crossbar arrays, effectively suppressing sneak path current through the unselected cells. The CMOS-based decoders are placed under the array region to obtain highly efficient memory. In another example, 3D RS device structure based on the 28 nm CMOS logic technology node was demonstrated using the Cu single damascene backend process (Figs. 4(b) and 4(c)) [71]. Here, a TaO_x based current rectifier (selector) was used to suppress sneak current and was connected to a 30 nm × 30 nm TaON based RS device in series (Fig. 4(d)). The ultra-high density 3D crossbar can operate with very high on/off ratio (~ 10^5), good retention (no degradation after 300 h at 150 °C) and endurance.

A fully integrated test-chip for high density memory applications was also demonstrated. 2 layer 32 Gb RS memory device in 24 nm technology node was successfully integrated on CMOS circuity (Fig. 4(e)) [51]. The CMOS circuit includes the array control circuit, sense amplifier, page buffer and voltage regulator drivers, with sense amplifiers and page buffer shared among the blocks. The page size is 2 KB, and latency for read and write is 40 and 230 μs, respectively. Moreover, a 16 Gb metal cation-based test chip was also developed with performance of 200 MB/s write and 1 GB/s read by Sony/Micron (Fig. 4(f)) [47, 72]. In this work, the 1T1R structure was implemented to mitigate the sneak current effects, at the cost of increased cell size to $6F^2$ (Figs. 4(g) and 4(h)). The 16 Gb array is divided into 8 banks, each with 8 Y-strips, vertical groups of tiles with a common global bitline. A Y-strip has 16 tiles plus one redundant tile where each tile consists of 8 k+256 bitline and 2 k wordline (16 Mbit). During read and write operations in a bank, 8 tiles are activated at the same time, and all 8 banks can be enabled simultaneously, achieving read and write operation with a total of 2,048+64 bytes simultaneously in 10 μs (program) and 2 μs (sense). The details of RS device characteristics and system performances for each prototype are summarized in Fig. 4(i).



(i)

	SK Hynix [70]	NTHU & TSMC [71]	Sandisk & Toshiba [51]	Micron & Sony [47,72]	Crossbar Inc. [92,93]
Technology node	54 nm	28 nm	24 nm	27 nm	28 nm
3D integration	2-stack	Multi-stackable	2-stack		Multi-stackable
Selector	Selectorless	TaOx-based diode	Diode selector	1T1R ($6F^2$)	FAST selector
Material	TiO_x/Ta_2O_5	TaON	Metal oxide	Cu based ECM	ECM
Density	2 Mb block	—	32 Gb chip	16 Gb chip	32 Mb chip
Latency	10 ns (program)	—	40 µs (read) /230 µs (program)	2 µs (read) /10 µs (read)	20 ns (read) /2 µs (program)
Operation voltage	< 4 V	< 5 V	—	< 3 V	< 2.5 V
Operation current	3 µA	10 µA	—	40 µA	< 100 µA
On/off ratio	10	100	—	1,000	100
Endurance	—	20 DC cycles	—	> 10^6 cycle	> 10^6 cycle
Retention	20 h at 150 °C	300 h at 150 °C	—		10 years at 85 °C

Figure 4 RS devices for memory applications. (a) TEM image of a 2-stack oxide-based RS crossbar arrays integrated in the 54nm standard CMOS process. Reproduced with permission from Ref. [70], © IEEE 2012. (b) Schematic of 3D stacked cross-point RS devices between M1 and M7, using TaO_x diode as a current rectifier. (c) SEM image of 28 nm TaON based 3D cross-point RS devices. (d) TEM image of a RS cell with 30 nm × 30 nm cell size and its stack schematic. Reproduced with permission from Ref. [71], © IEEE 2013. (e) Microphotograph of a 32 Gb 2-stack RS memory in 24 nm Technology, along with its cross-sectional view and specs. Reproduced with permission from Ref. [51], © IEEE 2013. (f) Die micrograph for a 16 Gb RS memory in 27 nm Technology. Reproduced with permission from Ref. [72], © IEEE 2014. (g) Schematic of the 27 nm RS memory device cross-section. (h) Cross-sectional TEM image of the 27 nm RS device array. Reproduced with permission from Ref. [47], © IEEE 2015. (i) Summary of reported RS devices and system performance developed by different companies.

4.2 In-memory computing: Deep learning accelerator

4.2.1 *Synaptic functions*

Beyond memory applications, RS devices are considered promising candidates for bio-inspired computing and in-memory computing in general, with their abilities to store and process information at the same physical locations [9, 11, 31, 73]. When applied in computing systems, the RS devices can be used in a crossbar form to perform VMM. In this approach, the values in the matrix are stored as the analog conductance values of the RS devices in the crossbar array. The input vector is applied as voltage pulses with different pulse amplitudes or different pulse widths to the rows of the crossbar. Natively through Ohm's law and Kirchhoff's current law, the currents or charges collected at the columns of the crossbar represent the resulting VMM outputs. As a result, the compute-intensive VMM operations can be obtained in a single step, greatly improving the energy efficiency and throughput beyond the limitations of a conventional computing system.

In a conventional von-Neumann architecture, the central processing unit is physically separated from the memory, resulting in severe energy and throughput penalties due to constant data movements, especially for data-intensive applications such as machine learning (ML) tasks. For instance, a ML program, AlphaGo, developed by DeepMind defeated one of the top-ranking professional players, Sedol Lee, in the board game Go in 2016 [74]. This event is widely considered as a milestone in the progress of artificial intelligence (AI). However, the AlphaGo system used 1,202 CPUs and 176 GPUs, corresponding to power consumption of around 170 kW (1,202 × 100 W + 176 × 300 W), much larger than that of the human brain, ~ 20 W [75]. Examining the system performance showed that for ML systems such as DNNs between 80% and 90% of the execution time is spent in memory access, compared to the 10% to 20% of the execution time spent in computation [8]. On the other hand, the human brain is more than four orders of magnitude more energy efficient than all current DNN systems. Unlike DNNs, it also does not require separated neural network structures for different tasks. Part of the efficiency can be attributed to the use of analog physical basis functions in the brain rather than digital logic basis functions in the computer system, but perhaps more importantly, brains perform computing in-memory and in parallel through very low power synaptic operations. In this sense, RS devices offer a possibility to emulate the brain system, and can potentially lead to exceptional computing capability with massive parallelism at extremely low power consumption.

In the nervous system, a neuron can communicate with other neurons by passing electrical or chemical signals through synapses [76]. Each neuron can be connected with thousands of other neurons with different connection strength, i.e. synaptic weight, which determines how efficient the input spikes from one neuron (the pre-synaptic neuron) can be delivered to the receiving neuron (the post-synaptic neuron). In addition, the synaptic weight can be updated by spikes from the pre- and post-neurons, allowing the system to achieving learning and form memory. These structures and functions can be implemented with RS device networks as well, thus making it possible to achieve highly-efficient bio-inspired computing hardware. For example, with external electrical stimulations from bottom or top electrodes, the configuration of internal ions in a RS device can be modulated (Fig. 5(a)), resulting in conductance updates that can emulate synaptic weight changes. For instance, consecutive positive and negative voltage biases to an Ag:Si/Si device allow the conductance modulation in an analog fashion (Fig. 5(b)) [32]. Effects such as LTP and LTD

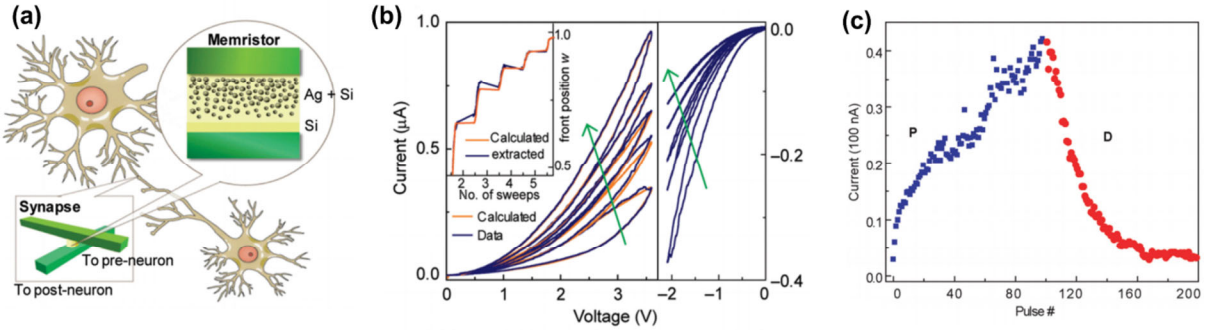


Figure 5 RS devices as synaptic elements. (a) Schematic of the concept of RS devices as synapses between neurons. (b) Experimentally measured (blue lines) and simulated (orange lines) *I*–*V* characteristics of an Ag:Si/Si device. Inset: simulated (orange lines) and extracted (blue lines) values of the normalized Ag front position (*w*) during positive DC sweeps. (c) The conductance of the Ag:Si/Si device can be incrementally increased (P) or decreased (D) by consecutive positive or negative pulses. Potentiation pulse condition: 3.2 V, 300 μs; depression pulse condition: −2.8 V, 300 μs. Reproduced with permission from Ref. [32], © American Chemical Society 2010.

can be achieved by applying 100 consecutive programming pulses, followed by 100 erase pulses, as shown in Fig. 5(c).

4.2.2 *VMM with crossbar array*

In deep learning algorithms, the VMM operation (or more basic multiply-accumulate (MAC) operation) is the core computing operation for training and inference but is very resource-expensive for conventional computing systems to implement. To accelerate VMM efficiently, the graphics processing unit has been extensively used to improve parallelism by using 1,000 s of compute cores with high-throughput connections to the memory. Algorithm studies to more efficiently map the neural networks (NNs) onto the hardware have also been conducted [77]. New hardware "accelerators", such as the tensor processing unit (TPU), were also designed to improve the efficiency of matrix operations and have enjoyed success through optimizations of the digital circuit and architecture design for these relatively narrow types of operations [78].

Unlike conventional hardware systems, RS crossbar array structures can naturally perform VMM in a single read step [27]. The NN structure can be readily mapped to the crossbar arrays, where the RS devices located at each crosspoint can store the weight matrix values as well as producing an output based on the input and the weight. The inputs and outputs of the network (or a layer in the network) are connected to rows and columns of the crossbar array, respectively. During inference, read voltage pulses are fed to the rows of the crossbar corresponding to the input signals, the VMM outputs are collected as current through the crossbar array at the columns (Fig. 6(a)) [79]. The VMM operation is completed concurrently without any data movement between processing and memory units, regardless of the matrix size, thus offering very high parallelism that leads to superior computing throughput and very high energy efficiency [9, 80]. Additionally, the conductance of RS devices can be adjusted by potentiation or depression pulses using a selected update rule (Fig. 5(c)), allowing the system to implement online learning using standard network training algorithms such as backpropagation [30, 31, 79, 81].

In practice, since RS devices only store positive conductance values, each synaptic weight (w) can be implemented with two RS devices representing a positive and a negative weight, G_{ij}^{+} and G_{ij}^{-}, respectively (Fig. 6(b)) [82]. That is, $w_{ij} = G_{ij}^{+} - G_{ij}^{-}$, where w_{ij} is the desired synaptic weight at row i and column j in the neural network.

If implemented using a passive crossbar array, RS devices with high nonlinear IV characteristics are desirable to minimize non-idealize effects from the sneak path current during training and inference operations. These effects will be more pronounced in larger arrays, when combined with other parasitic effects such as the finite line resistance. Introduction of a selector device in series with the RS device will be helpful to suppress the sneak current and increase the NN size. Besides CMOS transistors, various two-terminal thin film based selectors based on mechanisms such as Schottky barriers [43, 83–85], tunneling junctions [44, 45, 86, 87], ovonic threshold switches (OTS) [46, 88, 89], metal–insulator transitions (MIT) [90, 91], and field assisted superlinear threshold (FAST) [92, 93] effects have been proposed. These selector devices show either a

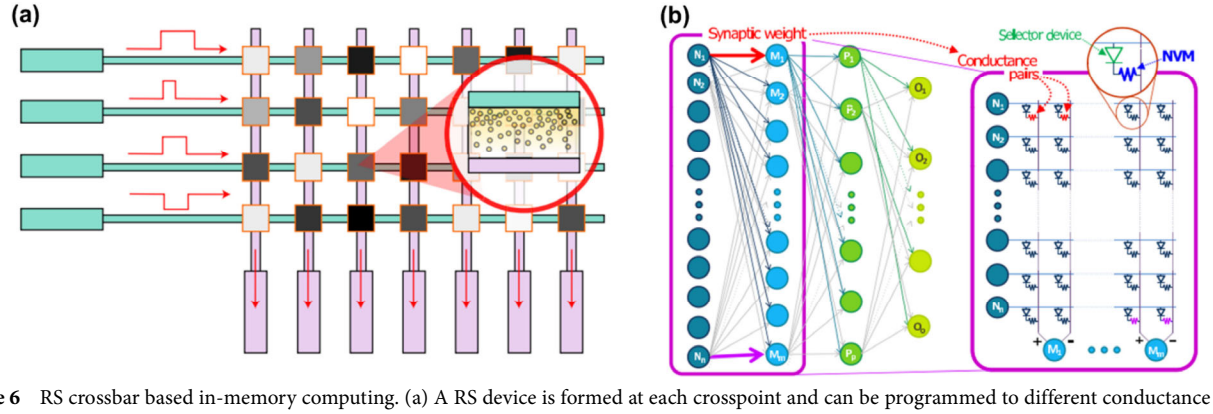


Figure 6 RS crossbar based in-memory computing. (a) A RS device is formed at each crosspoint and can be programmed to different conductance states (represented as grayscale color). The MAC operation can be implemented as vector-matrix multiplication (VMM) on a crossbar array through Ohm's law and Kirchhoff's current law, by using voltage pulses as inputs and collecting current or charge as the outputs. Reproduced with permission from Ref. [79], © Springer Nature 2017. (b) RS crossbars can be used to train and test networks. Signed weights can be implemented by two RS. Reproduced with permission from Ref. [82], © IEEE 2015.

rectifying behavior or a very nonlinear *I*–*V* characteristic that suppresses leakage current in the low-bias regime, and forms so-called 1S1R structure when integrated with the RS device.

4.2.3 *Deep neural network implementations*

As noted above, layers of a neural network can be directly mapped onto the crossbar structure. The input neurons and the output neurons are connected to the rows and columns of the crossbar, respectively, where RS devices at each crosspoint act as synapses. Such an implementation allows VMM to be performed in a massively parallel fashion. The ability to modulate the conductance of RS device also enables online training through standard backpropagation or more bio-inspired algorithms.

Several examples of RS device based artificial neural network (ANN) hardware have been recently demonstrated. For example, a single layer perceptron (SLP) was mapped on a 12 × 12 passive crossbar array with Al_2O_3/TiO_{2-x} RS devices without extrinsic selectors (Fig. 7(a)) [29]. Electroforming of each RS device was performed through a floating scheme, followed by Reset to minimize current leakage before the subsequent forming of other devices. 3 × 3 binary patterns of 3 letters (*z*, *v*, and *n*) and 27 noisy images created by flipping one pixel of each original image were used for training and classification. The total number of RS devices used in the study was 10 × 6, including a bias term at the input and the use of differential pair of RS devices for each synapse (Fig. 7(b)). Based on Manhattan update rule, the network was trained *in-situ*, with an effective conductance range 10–100 μS, and successful classification was achieved, on average, after 23 epochs (Fig. 7(c)).

Experimental implementation of a sparse coding algorithm was recently demonstrated in a crossbar with WO_x-based RS devices (Fig. 7(d)) [79]. Sparse coding, which aims at reducing the complexity of the input signals by representing the original data with a small set of features, can allow improved feature extraction and pattern recognition functions, as well as more efficient signal storage and analysis. Sparse coding algorithms

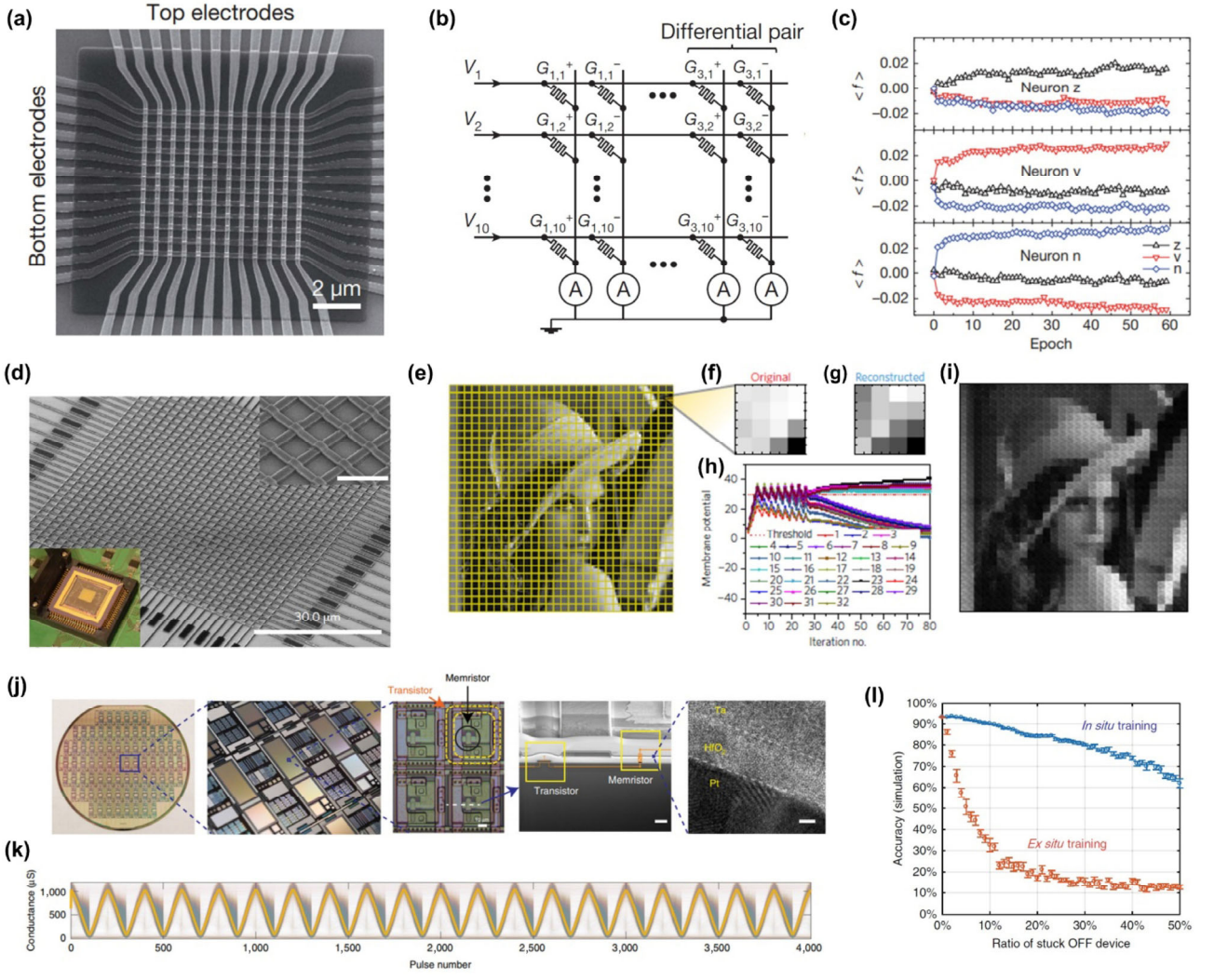


Figure 7 Neuromorphic hardware with RS crossbar arrays. (a) A 12 × 12 crossbar based on Al_2O_3/TiO_{2-x} RS devices. (b) A single layer perceptron (SLP) is implemented using a 10 × 6 sub-array. (c) Experimental pattern classification results during training. Reproduced with permission from Ref. [29], © Springer Nature 2015. (d) SEM image of a 32 × 32 crossbar array based on WO_x RS devices. Upper right inset: magnified SEM image of the crossbar. Scale bar: 3 μm. Lower left inset: the crossbar array chip mounted on the test board. (e) An original 120 × 120 image for sparse coding. The image is divided into non-overlapping 4 × 4 patches for processing on the crossbar-based hardware system. (f) A 4 × 4 patch from the original image. (g) The experimentally reconstructed patch from a 16 × 32 crossbar array. (h) Membrane potentials of the neurons as a function of iteration number during LCA analysis. The red horizontal line marks the threshold parameter λ. (i) Experimentally reconstructed image based on the reconstructed patches. Reproduced with permission from Ref. [79] © Springer Nature 2017. (j) Optical microscope, SEM and TEM images of a 1T1R (Ta/HfO_2/Pt RS device) array structure, from wafer level to individual cell. (k) All responsive devices over 20 potentiation/depression epochs. (l) The impact of non-responsive devices on the inference accuracy with *in-situ* and *ex-situ* training approaches. The experimental accuracy of 91.71% in this work was achieved with 11% devices unresponsive to conductance updates. Reproduced with permission from Ref. [94], © Springer Nature 2018.

can be used to perform natural image processing based on a learned dictionary through pattern matching and neuron lateral inhibition. In this implementation, a locally competitive algorithm (LCA) was performed by the hardware consisting of the memristor crossbar and periphery circuitry on a test board. Key operations, including VMM and matrix transpose operations, were directly and efficiently performed in the analog domain through a 32 × 32 crossbar array without the need to read each stored weight. Specifically, the original 120 × 120 image was divided into non-overlapping 4 × 4 patches for processing (Fig. 7(e)), and the experimentally implemented LCA network correctly reconstructs the input image while minimizing the number of activated neurons during forward-backward iterations for a given sparsity parameter λ (Figs. 7(f)–7(h)). After the network stabilizes, the image was successfully reproduced with the sparse neuron activities and the associated features (Fig. 7(i)). This work highlighted the potential of memristor crossbars to perform not only forward inference, but also backward propagation and online learning functions.

To scale-up the system, 1T1R crossbar arrays have also been extensively studied since selectors based on the mature CMOS technology are more reliable compared to two-terminal selector devices, allowing 1T1R arrays to be implemented in large scale. During forward inference operation, all transistors in the 1T1R crossbar array are turned on to perform VMM, while during weight update, the transistors at unselected devices are turned OFF to eliminate sneak current. Additionally, the transistor at the selected device can also be used to control the programming current that leads to precise tuning of the resistance during weight storage. In one example, a 128 × 64 1T1R crossbar array based on the Ta/HfO_2/Pt RS device structure was used to map a multi-layer neural network with *in-situ* and self-adaptive learning (Fig. 7(j)) [94]. In 1T1R device, the gate voltage on the transistor of the selected cell can be used to control the compliance current to fine tune the conductance of the RS device during the weight storage. Specifically, applying incremental gate voltage with positive voltage to the top electrode, the conductance will be increased (Fig. 7(k)). A multi-layer perceptron with 64 input neurons, 54 hidden neurons, and 10 output neurons was trained on the reduced Modified National Institute of Standards and Technology (MNIST) dataset (using 8 × 8 input images) with a minibatch size 50. Although there are almost 11% of device defects, an accuracy of 91.71% was still achieved, proving that online training can tolerate hardware imperfections by allowing the network to adapt to the defects and update weights accordingly. In contrast, when pre-trained weights are loaded in the network, it does not tolerate the device imperfections, resulting in significant accuracy drop (Fig. 7(l)).

4.2.4 *Fully integrated deep learning platform*

Although the key VMM operations can be performed efficiently with RS crossbar arrays, prior demonstrations of neuromorphic hardware implementations are based on discrete arrays and separate processing elements using test boards. For practical implementation of the RS based hardware, the arrays need to be integrated with all other necessary circuitry including digital to analog converters (DACs) and analog-digital converters (ADCs), which are needed to send the input to and to collect the output from the crossbar array, and controllers to convert the input signals to pulse amplitude or width. To reduce latency and power consumption, all these components need to be integrated together with the crossbar array on a single chip, instead of using discrete components on a board. Integrating a processor on chip will also allow the neuron functions and network structures to be reprogrammed through simple software changes, enabling different models to be mapped on the same hardware platform.

Cai et al. [52] demonstrated the first integrated, programmable chip based on this principle. In this study, a fully-functional, programmable neuromorphic computing chip was fabricated with a passive crossbar array directly integrated with a complete set of analog and digital components and an on-chip processor. Specifically, a 54 × 108 crossbar array was integrated on top of the CMOS periphery and control circuitry in a BEOL process, where each row and column of the crossbar array is connected to a specific landing pad left open during CMOS fabrication process, and then connected to the circuitry underneath through internal CMOS wiring (Figs. 8(a) and 8(b)). The custom CMOS circuitry includes an OpenRISC processor with 64 KB SRAM and a mixed-signal interface with 162 configurable channels. Each channel can be programmed to have either an ADC, or 1 of the 3 DACs connected to a row or column of the crossbar (Fig. 8(c)). During training and inference, the binary program instructions are executed through the OpenRISC processor without the need to access external controllers. The inputs to the RS device array are supplied through the DACs following the instructions, and the VMM results are read as charge values from the ADCs and processed by the OpenRISC processor for batch gradient descent calculations or for running other algorithms. In other words, all operations can be performed on chip during inference and training operations.

The integrated hybrid chip allows different computing tasks efficiently to be mapped on the memristor-based computing platform by using the bi-directional VMM operations in the crossbars and the flexibility of the CMOS interface and control circuitry. For instance, a SLP network was first implemented with 5 × 5 Greek letter binary patterns for classification (Fig. 8(d)) for 5 classes (Ω, M, Π, Σ, and Ψ). The SLP was mapped on the integrated chip using a 26 × 10 subarray in which input data were converted to voltage pulses through the on-chip circuitry, depending on the pixel values (Fig. 8(e)). Online training of the network was achieved using the batch gradient descent rule. After training, the SLP can achieve 100% classification accuracy on a test set not included in training (Fig. 8(f)), verifying successful training and inference of the chip. Using the same chip, a bilayer neural network using two subarrays was also demonstrated to analyze and classify data for breast cancer screening based on principal component analysis (PCA). Specifically, a 9 × 2 network that performs PCA of the original data was used as the first layer of the system, which reduces the 9-dimensional raw input data to a 2-dimensional space based on the learned principal components (PCs). The second layer is a 3 × 1 SLP layer performing classification using the reduced data in the 2-dimensional space for the two classes (benign or malignant) (Figs. 8(g) and 8(h)). After 30 online training epochs, the experimentally implemented bilayer network can achieve 94.6% classification accuracy, comparable to those obtained from software implementation (Figs. 8(i) and 8(j)). When implemented in the 40 nm technology node, the total system power consumption of the system is estimated to be 42.1 mW, corresponding to a power efficiency of 1.37 TOPS/W, with better performance expected if the ADC design can be further improved. These successful demonstrations of RS device-based hardware systems verified the feasibility of highly efficient computing based on these emerging devices and computing architectures, and will stimulate continued interest in device and architecture innovations and potentially lead to applications in edge computing and other AI use scenarios [95, 96].

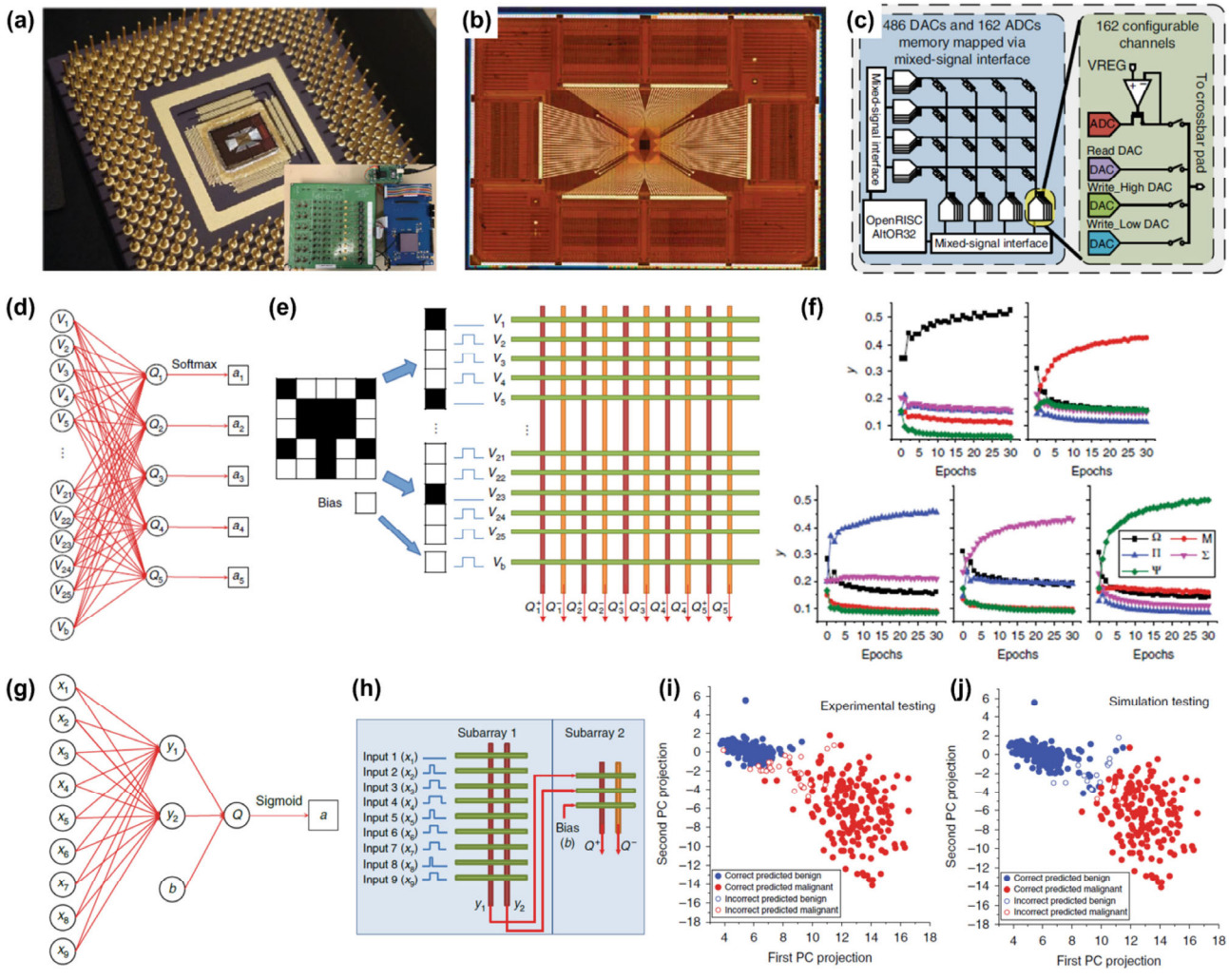


Figure 8 Fully-integrated computing system with a RS crossbar array directly integrated on CMOS circuitry. (a) The integrated chip wire-bonded on a pin grid array (PGA) package. Inset: test board used to power and test the integrated chip. (b) Optical image showing the 54 × 108 WO_x crossbar array integrated on the CMOS circuit. (c) Schematic of the mixed-signal CMOS interface to the 54 × 108 crossbar array, with two write DACs, one read DAC and one ADC for each row and column. (d) Schematic of the single-layer perceptron (SLP) for classification of 5 × 5 images. (e) Implementation of the SLP using a 26 × 10 subarray through the integrated chip. (f) Evolution of the output neuron signals during training, averaged over all training patterns for a specific class. (g) Schematic of the bilayer neural network for PCA analysis and classification. (h) The bilayer network is mapped onto the integrated chip, using a 9 × 2 subarray for the PCA layer and a 3 × 2 subarray for the classification layer. (i) Classification results experimentally obtained from the integrated chip for the testing data. (j) Classification results obtained from simulation for the testing data. Reproduced with permission from Ref. [52], © Springer Nature 2019.

4.3 More bio-faithful implementations

4.3.1 Spiking neural networks (SNNs)

It is generally believed that more bio-realistic implementations could lead to even higher energy efficiency. One such example is SNNs, which encode information in the timing and frequency of spikes. SNNs have been shown to offer extremely high energy efficiency [97, 98]. Unlike DNNs in which all signals are collected from every neuron in the previous layer and the processed information are sent to every neuron in the next layer, neurons in SNNs fire only when the membrane potential reaches above a threshold value. When a neuron fires, the connection strength of the synapses associated with it may also be modulated accordingly. With this approach, data can be represented and processed with a small number of spikes, with the system consuming very little power in between.

With the rich internal ionic dynamic processes, RS devices such as memristors can natively emulate some of the key underlying physical and chemical processes in biological synapses as well as neurons. This allows SNN networks to be efficiently implemented using memristor devices. For example, different synaptic plasticity effects [32, 99, 100] can be natively implemented using a so-called second-order memristor effect that can emulate the internal Ca^{2+} concentration dynamics [101, 102]. Neuron functions such as integrate- and fire function [103, 104] have also been demonstrated using RS devices.

Despite the great potential, SNNs have not been widely implemented as DNNs. One of the reasons is the lack of efficient SNN algorithms, particularly for complex tasks. For instance, spike-time-dependent plasticity (STDP) [76] is an efficient learning rule of synaptic plasticity, but the image classification accuracy of networks based on this rule is generally lower than those achieved using conventional DNNs [105]. In particular, efficient training algorithms for large SNN networks are not well established, making it difficult for SNN systems to compete with conventional DNN implementations which have enjoyed great commercial success recently. Systematic developments of efficient algorithms, along with devices and hardware, are needed to bring the training of SNNs from the current academic research level to large scale commercial implementations.

4.3.2 *Reservoir computing (RC) systems*

Temporal data, including videos, speech, and other signals that evolve with time acrossing different time scales, are of high technological and societal importance but are difficult to process with conventional DNNs. To process temporal data, networks with internal dynamics such as recurrent networks (RNNs) have been developed. However, generic RNNs are expensive to train. To address these challenges, RC systems [106, 107] were proposed, where the original inputs can be non-linearly projected into a high-dimensional feature space through a dynamic reservoir. With this approach, the original features that may not be linearly separable can become linearly separable in the new feature space, and can then be further processed with a simple linear network. RC systems have been shown to outperform classical fully trained RNNs in many tasks [108–110]. To perform the nonlinear transformation of temporal data, a key requirement of the reservoir is to have a "fading memory", or "short-term memory" effect, so that the system can respond to inputs at the near past but not far past. With this approach, the reservoir only needs to be excited (mapping input features to different excited reservoir states) whereas the connectivity structure inside the reservoir remains fixed at all time, and thus does not require training. To further process the transformed data, a second network, called a readout function, is trained and generate the final output. Since the most difficult task of separating the features is implemented in the reservoir, a simple and small readout network is typically sufficient, thus dramatically reducing the training cost of the overall system.

Implementing the reservoirs using conventional systems can however be expensive. Recently, by utilizing the internal short-term ionic dynamics, memristor devices with native short-term memory effects have been successfully used to build RC systems [111]. The internal ionic dynamic processes allow the RS devices to map temporal input patterns into different reservoir states (represented as memristor resistances), which can be further processed via a simple readout function (Fig. 9(a)). For example, with consecutive short pulses, the conductance of a memristor with short-term memory effect is gradually increased, while the conductance state will decay without any stimulation (Fig. 9(b)). As a result, the final device state depends on the temporal pattern of the input, with different patterns leading to different reservoir states. Using the native short-term memory effects allows a reservoir to be built with a small number of devices (including using just one device), significantly reducing hardware implementation complexity. In one experimental implementation, the pre-processed MNIST images (composed of black and white pixels) are first converted into pulse streams and fed to the memristor-based reservoir, where a white pixel corresponds to a write pulse, while no pulse with a black pixel. In this study, 88 memristor devices were used as reservoir, and a 176 × 10 readout network was used for classification, resulting in 88.1% classification accuracy (Fig. 9(c)).

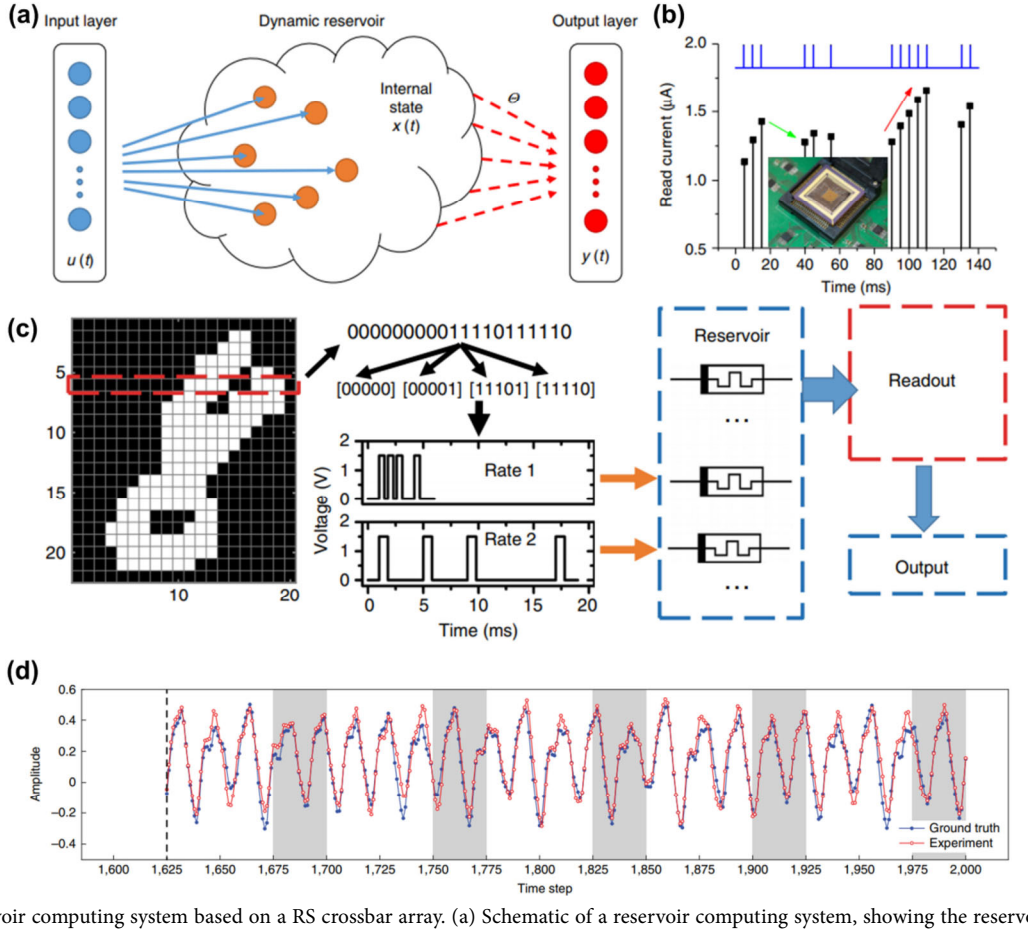


Figure 9 Reservoir computing system based on a RS crossbar array. (a) Schematic of a reservoir computing system, showing the reservoir with internal dynamics and a readout function. Only weights Θ in the readout function, connecting the reservoir state $x(t)$ and the output $y(t)$, need to be trained. (b) Response of a typical WO_x device to a pulse stream with different time intervals between pulses. Inset: image of the crossbar array wired-bonded to a chip carrier and mounted on a test board. (c) The original MNIST image is first converted into pulse streams and fed to the RS device-based reservoir at different rates. The recognition result is generated after feeding the reservoir state to a trained readout function. Reproduced with permission from Ref. [111], © Springer Nature 2017. (d) Long-term forecasting of Mackey–Glass time series using a reservoir system, where the predicted output from the network is fed back to the network as input for the next time step. Reproduced with permission from Ref. [113], © Springer Nature 2019.

It should be noted that in the RC system, the performance strongly depends on the dimension of the reservoir space. However, instead of increasing the number of physical nodes in the reservoir which can increase hardware implementation cost as well as the training cost of the readout layer, the concept of virtual nodes developed in delay systems can be an attractive alternative [112]. The state of virtual nodes depends on the node's own previous state, the current state of adjacent nodes, and the masked input signal, allowing them to be nonlinearly coupled. By using randomly generated masks for input signals, diverse responses can be obtained from the virtual nodes, and these systems have been shown to be able to achieve performance comparable to that of conventional and well-designed reservoirs. In a recent implementation of memristor-based RC system, the reservoir size was increased by using the virtual node concept, and spoken-digit was successfully recognized with accuracy of 99.2%. More interestingly, since the network can capture the temporal features of the input, it was successfully used to perform prediction/forecasting functions. For example, in speech recognition, the speaker's intended word was correctly predicted before the speaker finished it. In another example, the network was able to capture the complex features and make predictions of a chaotic system such as the Mackey–Glass series, a deterministic form of chaotic system but difficult to predict. Periodic updates can be used to bring the reservoir state back to the original dynamics and make it possible to maintain even long-term prediction of the chaotic systems (Fig. 9(d)) [113].

4.4 Logic applications

RS devices have also been proposed for logic operations in electronic circuits, offering advantages such as high compute density and nonvolatility [81, 114–122]. For instance, a fundamental Boolean logic operation, material implication (IMP), was implemented as a logic gate based on two RS devices (e.g. *P* and *Q*) [114]. In this approach, logic values were represented by the resistance of the RS devices (i.e., "0" for the HRS and "1" for the LRS), and the IMP operation was achieved based on the voltage divider effect. With well-designed voltage pulses and series resistance values, the resistance of the output RS device (q) is determined by the input logic state (p), producing the desired truth table for the IMP operation. With iterative IMP operations, all other Boolean logic operations can then be achieved [114, 115].

Beyond logic gate applications, RS devices can be used as nonvolatile switches in field programmable gate arrays (FPGA) systems [81, 121, 122]. Typically, the logic units in the FPGA are connected by volatile switch units (i.e. SRAM) which occupy a relatively large area of more than 100 F^2. Switches based on RS devices can be reprogrammed to reconfigure the connections and the functionality of the FPGA, leading to improved density and power metrics. For instance, FPGA systems using hybrid CMOS/RS-device circuits have been experimentally demonstrated [81]. In this hybrid circuit, the reconfigurable RS devices act as logic elements that define data paths connecting logic gates into digital circuits, enabling FPGA-like functionality.

5 Challenges and future opportunities

RS devices have made remarkable progress over the last ~ 15 years. Beyond already being offered as commercial products for memory applications, RS devices have been extensively studied for neuromorphic computing applications, providing significant benefits for real-time data processing with high throughput and low energy consumption. However, contemporary RS devices are still far from being ideal. First, due to its stochastic switching behavior based on individual ion/cation migration, significant device-to-device and cycle-to-cycle variations exist [123, 124]. Unlike binary memory applications which only need enough read window margin to simply distinguish between "ON" (LRS) from "OFF" (HRS), these variations can be a major factor affecting the accuracy of analog computing, resulting in irreversible computation error. Since ion migration strongly depends on the local electric field in the switching layer, integration of intentionally-designed non-uniform structures in the device might be one possible solution to improve the variability, by enhancing electric field focusing and confining filament formation in a localized area. Inserting nanoclusters [125], formation of nanopores with a graphene layer [126], and confined dislocations in the switching layer [127] have been suggested and demonstrated to improve device performance and uniformity.

Second, the impact of conductance update linearity and symmetry can be critical on the neural network training accuracy [128]. In general, during training, the calculated weight update (Δw) should be directly transformed to the number of programming pulses for the RS device, without having to read out and accounting for the current weight. Ideally, a single programming pulse (LTP or LTD pulse) should change the weight by a constant value, independent of the conductance state. However, due to the non-linear dependence of conductance on filament shape in filament-based RS devices, typical RS devices suffer from nonlinear and asymmetric weight update characteristics. This leads to conductance changes that depend on the present conductance state, making it much more challenging to implement online training.

In addition, a large dynamic range and good resistance stability are also crucial to reduce the error for VMM operation and to maintain high classification accuracy [129]. There is still no perfect device that can meet all these requirements. Continued improvements on the ionic process control and device optimization are essential, along with architecture innovations such as hybrid non-volatile memory (NVM)–CMOS neural-network implementations [130], mixed-precision [131], multi-RS device architectures [132] and other precision extension techniques [133].

It is worth noting that the conductance update nonlinearity and asymmetry mostly affect online training applications, while for inference applications based on models already trained offline, the conductance can be nearly perfectly tuned through iterative programming with the program-and-verify (PNV) technique [134]. These factors, combined with the strong interest and need to bring intelligent inference to the edge, suggest that inference applications may be a good choice for the first generations of AI accelerators based on RS devices.

For practical applications, scaling up the system with larger networks is highly desirable. Rather than simply increasing the crossbar size in which the summed current (VMM results) can be quadratically affected by the line resistance and other nonlinear effects, tiling crossbars together in a modular fashion appears to be a promising approach [52]. For instance, with a tile consisting of a moderate crossbar array (e.g. 128 × 128) with underlying interface CMOS circuits including decoder, multiplexers (MUXs), DACs, and ADCs, tile-to-tile communication can be performed in the CMOS layer in the digital domain. This tiled architecture allows mapping of large models on practical RS hardware components without significantly sacrificing performance [135]. The main remaining challenges include device uniformity and reliability optimizations, and optimization of the ADC design to reduce the power consumption and area, as ADC size and power can now become

dominating factors in these mixed-signal systems. Since the ADC size and power are strongly dependent on the precision and speed requirements, bio-inspired applications that do not require very high precision, and achieve high throughput through parallelism instead of very high speed components, are well suited for these hardware systems. Combined with algorithm advances, such as quantized neural networks that can further reduce precision requirements of the individual components without sacrificing the system performance [136, 137], RS crossbar based hardware appears to have a bright future to enable highly sophisticated and powerful computing systems for applications ranging from IoT devices, autonomous systems, to large scale enterprise applications [138].

Acknowledgements

The authors thank insightful discussions with Dr. M. A. Zidan and J. Moon. This work was supported by in part by the National Science Foundation through awards CCF-1900675 and DMR-1810119. W. D. L. would like to thank Charlie for his tremendous support during his stay as a postdoc in the Lieber group from 2003–2005, and for his advice that led to the conception of the initial concept of metal-ion based RS devices.

Conflict of interest

The authors declare no competing financial interests.

References

[1] Moore, G. E. Cramming more components onto integrated circuits. *Electronics* **1965**, *38*, 114–117.

[2] Moore, G. E. Progress in digital integrated electronics. In *Proceedings of the International Electron Devices Meeting*, Washigton, DC, USA, 1975, pp 11–13.

[3] Sutter, H. The free lunch is over: A fundamental turn toward concurrency in software. *Dr. Dobb's J.* **2005**, *30*, 202–210.

[4] Taur, Y.; Buchanan, D. A.; Chen, W.; Frank, D. J.; Ismail, K. E.; Lo, S. H.; Sai-Halasz, G. A.; Viswanathan, R. G.; Wann, H. J. C.; Wind, S. J. et al. CMOS scaling into the nanometer regime. *Proc. IEEE* **1997**, *85*, 486–504.

[5] Frank, D. J.; Dennard, R. H.; Nowak, E.; Solomon, P. M.; Taur, Y.; Wong, H. S. P. Device scaling limits of Si MOSFETs and their application dependencies. *Proc. IEEE* **2001**, *89*, 259–288.

[6] Khan, H. N.; Hounshell, D. A.; Fuchs, E. R. H. Science and research policy at the end of Moore's law. *Nat. Electron.* **2018**, *1*, 14–21.

[7] Wulf, W. A.; McKee, S. A. Hitting the memory wall: Implications of the obvious. *ACM SIGARCH Comput. Arch. News* **1995**, *23*, 20–24.

[8] Horowitz, M. Computing's energy problem (and what we can do about it). In *Proceedings of 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, San Francisco, CA, USA, 2014, pp 10–14.

[9] Zidan, M. A.; Strachan, J. P.; Lu, W. D. The future of electronics based on memristive systems. *Nat. Electron.* **2018**, *1*, 22–29.

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

[11] Xia, Q. F.; Yang, J. J. Memristive crossbar arrays for brain-inspired computing. *Nat. Mater.* **2019**, *18*, 309–323.

[12] Zhu, X. J.; Lee, S. H.; Lu, W. D. Nanoionic resistive-switching devices. *Adv. Electron. Mater.* **2019**, *5*, 1900184.

[13] Lee, J.; Lu, W. D. On-demand reconfiguration of nanomaterials: When electronics meets ionics. *Adv. Mater.* **2018**, *30*, 1702770.

[14] Waser, R.; Aono, M. Nanoionics-based resistive switching memories. *Nat. Mater.* **2007**, *6*, 833–840.

[15] Strukov, D. B.; Snider, G. S.; Stewart, D. R.; Williams, R. S. The missing memristor found. *Nature* **2008**, *453*, 80–83.

[16] Lam, C. H. Storage class memory. In *Proceedings of 2010 10th IEEE International Conference on Solid-State and Integrated Circuit Technology*, Shanghai, China, 2010, pp 1080–1083.

[17] Burr, G. W.; Kurdi, B. N.; Scott, J. C.; Lam, C. H.; Gopalakrishnan, K.; Shenoy, R. S. Overview of candidate device technologies for storage-class memory. *IBM J. Res. Dev.* **2008**, *52*, 449–464.

[18] Sills, S.; Yasuda, S.; Strand, J.; Calderoni, A.; Aratani, K.; Johnson, A.; Ramaswamy, N. A copper ReRAM cell for storage class memory applications. In *Proceedings of 2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers*, Honolulu, HI, USA, 2014.

[19] Ielmini, D.; Nardi, F.; Cagli, C. Universal reset characteristics of unipolar and bipolar metal-oxide RRAM. *IEEE Trans. Electron Devices* **2011**, *58*, 3246–3253.

[20] Choi, B. J.; Torrezan, A. C.; Strachan, J. P.; Kotula, P. G.; Lohn, A. J.; Marinella, M. J.; Li, Z. Y.; Williams, R. S.; Yang, J. J. High-speed and low-energy nitride memristors. *Adv. Funct. Mater.* **2016**, *26*, 5290–5296.

[21] Pi, S.; Li, C.; Jiang, H.; Xia, W. W.; Xin, H. L.; Yang, J.; Xia, Q. F. Memristor crossbar arrays with 6-nm half-pitch and 2-nm critical dimension. *Nat. Nanotechnol.* **2019**, *14*, 35–39.

[22] Govoreanu, B.; Kar, G. S.; Chen, Y. Y.; Paraschiv, V.; Kubicek, S.; Fantini, A.; Radu, I. P.; Goux, L.; Clima, S.; Degraeve, R. et al. 10 × 10 nm^2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation. In *Proceedings of 2011 International Electron Devices Meeting*, Washington, DC, USA, 2011, pp 31.6.1–31.6.4.

[23] Yang, J. J.; Zhang, M. X.; Strachan, J. P.; Miao, F.; Pickett, M. D.; Kelley, R. D.; Medeiros-Ribeiro, G.; Williams, R. S. High switching endurance in TaO_x memristive devices. *Appl. Phys. Lett.* **2010**, *97*, 232102.

[24] Lee, M. J.; Lee, C. B.; Lee, D.; Lee, S. R.; Chang, M.; Hur, J. H.; Kim, Y. B.; Kim, C. J.; Seo, D. H.; Seo, S. et al. A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta_2O_{5-x}/TaO_{2-x} bilayer structures. *Nat. Mater.* **2011**, *10*, 625–630.

[25] Handy, J. Understanding the intel/micron 3D XPoint memory. In *Proceeding of 2015 Storage Developer Conference*, Santa Clara 2015.

[26] Ielmini, D.; Wong, H. S. P. In-memory computing with resistive switching devices. *Nat. Electron.* **2018**, *1*, 333–343.

[27] Di Ventra, M.; Pershin, Y. V. The parallel approach. *Nat. Phys.* **2013**, *9*, 200–202.

[28] Indiveri, G.; Liu, S. C. Memory and information processing in neuromorphic systems. *Proc. IEEE* **2015**, *103*, 1379–1397.

[29] Prezioso, M.; Merrikh-Bayat, F.; Hoskins, B. D.; Adam, G. C.; Likharev, K. K.; Strukov, D. B. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **2015**, *521*, 61–64.

[30] Li, C.; Hu, M.; Li, Y. N.; Jiang, H.; Ge, N.; Montgomery, E.; Zhang, J. M.; Song, W. H.; Dávila, N.; Graves, C. E. et al. Analogue signal and image processing with large memristor crossbars. *Nat. Electron.* **2018**, *1*, 52–59.

[31] Hu, M.; Graves, C. E.; Li, C.; Li, Y. N.; Ge, N.; Montgomery, E.; Davila, N.; Jiang, H.; Williams, R. S.; Yang, J. J. et al. Memristor-based analog computation and neural network classification with a dot product engine. *Adv. Mater.* **2018**, *30*, 1705914.

[32] Jo, S. H.; Chang, T.; Ebong, I.; Bhadviya, B. B.; Mazumder, P.; Lu, W. Nanoscale memristor device as synapse in neuromorphic systems. *Nano Lett.* **2010**, *10*, 1297–1301.

[33] Wang, Z. R.; Joshi, S.; Savel'ev, S. E.; Jiang, H.; Midya, R.; Lin, P.; Hu, M.; Ge, N.; Strachan, J. P.; Li, Z. Y. et al. Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. *Nat. Mater.* **2017**, *16*, 101–108.

[34] Chua, L. O. Memristor-the missing circuit element. *IEEE Trans. Circuit Theory* **1971**, *18*, 507–519.

[35] Chua, L. O.; Kang, S. M. Memristive devices and systems. *Proc. IEEE* **1976**, *64*, 209–223.

[36] Waser, R.; Dittmann, R.; Staikov, C.; Szot, K. Redox-based resistive switching memories-nanoionic mechanisms, prospects, and challenges. *Adv. Mater.* **2009**, *21*, 2632–2663.

[37] Valov, I.; Lu, W. D. Nanoscale electrochemistry using dielectric thin films as solid electrolytes. *Nanoscale* **2016**, *8*, 13828–13837.

[38] Adam, G. C.; Hoskins, B. D.; Prezioso, M.; Strukov, D. B. Optimized stateful material implication logic for three-dimensional data manipulation. *Nano Res.* **2016**, *9*, 3914–3923.

[39] Russo, U.; Kamalanathan, D.; Ielmini, D.; Lacaita, A. L.; Kozicki, M. N. Study of multilevel programming in programmable metallization cell (PMC) memory. *IEEE Trans. Electron Devices* **2009**, *56*, 1040–1047.
[40] Menzel, S.; Böttger, U.; Waser, R. Simulation of multilevel switching in electrochemical metallization memory cells. *J. Appl. Phys.* **2012**, *111*, 014501.
[41] Balatti, S.; Larentis, S.; Gilmer, D. C.; Ielmini, D. Multiple memory states in resistive switching devices through controlled size and orientation of the conductive filament. *Adv. Mater.* **2013**, *25*, 1474–1478.
[42] Burr, G. W.; Shenoy, R. S.; Virwani, K.; Narayanan, P.; Padilla, A.; Kurdi, B.; Hwang, H. Access devices for 3D crosspoint memory. *J. Vac. Sci. Technol. B* **2014**, *32*, 040802.
[43] Kim, G. H.; Lee, J. H.; Ahn, Y.; Jeon, W.; Song, S. J.; Seok, J. Y.; Yoon, J. H.; Yoon, K. J.; Park, T. J.; Hwang, C. S. 32 × 32 crossbar array resistive memory composed of a stacked schottky diode and unipolar resistive memory. *Adv. Funct. Mater.* **2013**, *23*, 1440–1449.
[44] Choi, B. J.; Zhang, J. M.; Norris, K.; Gibson, G.; Kim, K. M.; Jackson, W.; Zhang, M. X. M.; Li, Z. Y.; Yang, J. J.; Williams, R. S. Trilayer tunnel selectors for memristor memory cells. *Adv. Mater.* **2016**, *28*, 356–362.
[45] Govoreanu, B.; Adelmann, C.; Redolfi, A.; Zhang, L. Q.; Clima, S.; Jurczak, M. High-performance metal-insulator-metal tunnel diode selectors. *IEEE Electron Device Lett.* **2014**, *35*, 63–65.
[46] Kau, D.; Tang, S.; Karpov, I. V.; Dodge, R.; Klehn, B.; Kalb, J. A.; Strand, J.; Diaz, A.; Leung, N.; Wu, J. et al. A stackable cross point phase change memory. In *Proceedings of 2009 IEEE International Electron Devices Meeting*, Baltimore, MD, USA, 2009, pp 1–4.
[47] Zahurak, J.; Miyata, K.; Fischer, M.; Balakrishnan, M.; Chhajed, S.; Wells, D.; Li, H.; Torsi, A.; Lim, J.; Korber, M. et al. Process integration of a 27nm, 16Gb Cu ReRAM. In *Proceedings of 2014 IEEE International Electron Devices Meeting*, San Francisco, CA, USA, 2014, pp 6.2.1–6.2.4.
[48] Baek, I. G.; Park, C. J.; Ju, H.; Seong, D. J.; Ahn, H. S.; Kim, J. H.; Yang, M. K.; Song, S. H.; Kim, E. M.; Park, S. O. et al. Realization of vertical resistive memory (VRRAM) using cost effective 3D process. In *Proceedings of 2011 International Electron Devices Meeting*, Washington, DC, USA, 2011, pp 31.8.1–31.8.4.
[49] Hsu, C. W.; Wan, C. C.; Wang, I. T.; Chen, M. C.; Lo, C. L.; Lee, Y. J.; Jang, W. Y.; Lin, C. H.; Hou, T. H. 3D vertical TaO_x/TiO_2 RRAM with over 10^3 self-rectifying ratio and sub-MA operating current. In *Proceedings of 2013 IEEE International Electron Devices Meeting*, Washington, DC, USA, 2013, pp 10.4.1–10.4.4.
[50] Bai, Y.; Wu, H. Q.; Wu, R. G.; Zhang, Y.; Deng, N.; Yu, Z. P.; Qian, H. Study of multi-level characteristics for 3D vertical resistive switching memory. *Sci. Rep.* **2014**, *4*, 5780.
[51] Liu, T. Y.; Yan, T. H.; Scheuerlein, R.; Chen, Y. C.; Lee, J. K.; Balakrishnan, G.; Yee, G.; Zhang, H.; Yap, A.; Ouyang, J. W. et al. A 130.7mm^2 2-layer 32Gb ReRAM memory device in 24nm technology. In *Proceedings of 2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, San Francisco, CA, USA, 2013, pp 210–211.
[52] Cai, F. X.; Correll, J. M.; Lee, S. H.; Lim, Y.; Bothra, V.; Zhang, Z. Y.; Flynn, M. P.; Lu, W. D. A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations. *Nat. Electron.* **2019**, *2*, 290–299.
[53] Yang, J. J.; Pickett, M. D.; Li, X. M.; Ohlberg, D. A. A.; Stewart, D. R.; Williams, R. S. Memristive switching mechanism for metal/oxide/metal nanodevices. *Nat. Nanotechnol.* **2008**, *3*, 429–433.
[54] Park, G. S.; Kim, Y. B.; Park, S. Y.; Li, X. S.; Heo, S.; Lee, M. J.; Chang, M.; Kwon, J. H.; Kim, M.; Chung, U. I. et al. *In situ* observation of filamentary conducting channels in an asymmetric Ta_2O_{5-x}/TaO_{2-x} bilayer structure. *Nat. Commun.* **2013**, *4*, 2382.
[55] Kim, S.; Choi, S.; Lu, W. Comprehensive physical model of dynamic resistive switching in an oxide memristor. *ACS Nano* **2014**, *8*, 2369–2376.
[56] Nardi, F.; Balatti, S.; Larentis, S.; Ielmini, D. Complementary switching in metal oxides: Toward diode-less crossbar RRAMs. In *Proceedings of 2011 International Electron Devices Meeting*, Washington, DC, USA, 2011, pp 31.1.1–31.1.4.
[57] Larentis, S.; Nardi, F.; Balatti, S.; Gilmer, D. C.; Ielmini, D. Resistive switching by voltage-driven ion migration in bipolar RRAM—Part II: Modeling. *IEEE Trans. Electron Devices* **2012**, *59*, 2468–2475.
[58] Kim, S.; Kim, S. J.; Kim, K. M.; Lee, S. R.; Chang, M.; Cho, E.; Kim, Y. B.; Kim, C. J.; -In Chung, U.; Yoo, I. K. Physical electro-thermal model of resistive switching in Bi-layered resistance-change memory. *Sci. Rep.* **2013**, *3*, 1680.
[59] Kresse, G.; Hafner, J. *Ab initio* molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B* **1994**, *49*, 14251–14269.
[60] Lee, J.; Schell, W.; Zhu, X. J.; Kioupakis, E.; Lu, W. D. Charge transition of oxygen vacancies during resistive switching in oxide-based RRAM. *ACS Appl. Mater. Interfaces* **2019**, *11*, 11579–11586.
[61] Valov, I.; Waser, R.; Jameson, J. R.; Kozicki, M. N. Electrochemical metallization memories—fundamentals, applications, prospects. *Nanotechnology* **2011**, *22*, 254003.
[62] Yang, Y. C.; Gao, P.; Gaba, S.; Chang, T.; Pan, X. Q.; Lu, W. Observation of conducting filament growth in nanoscale resistive memories. *Nat. Commun.* **2012**, *3*, 732.
[63] Tian, X. Z.; Wang, L. F.; Wei, J. K.; Yang, S. Z.; Wang, W. L.; Xu, Z.; Bai, X. D. Filament growth dynamics in solid electrolyte-based resistive memories revealed by *in situ* TEM. *Nano Res.* **2014**, *7*, 1065–1072.
[64] Gaba, S.; Cai, F. X.; Zhou, J. T.; Lu, W. D. Ultralow sub-1-nA operating current resistive memory with intrinsic non-linear characteristics. *IEEE Electron Device Lett.* **2014**, *35*, 1239–1241.
[65] Belmonte, A.; Celano, U.; Chen, Z.; Radhaskrishnan, J.; Redolfi, A.; Clima, S.; Richard, O.; Bender, H.; Kar, G. S.; Vandervorst, W. et al. Voltage-controlled reverse filament growth boosts resistive switching memory. *Nano Res.* **2018**, *11*, 4017–4025.
[66] Jo, S. H.; Kim, K. H.; Lu, W. High-density crossbar arrays based on a Si memristive system. *Nano Lett.* **2009**, *9*, 870–874.
[67] Yang, Y. C.; Gao, P.; Li, L. Z.; Pan, X. Q.; Tappertzhofen, S.; Choi, S.; Waser, R.; Valov, I.; Lu, W. D. Electrochemical dynamics of nanoscale metallic inclusions in dielectrics. *Nat. Commun.* **2014**, *5*, 4232.
[68] Kim, K. H.; Gaba, S.; Wheeler, D.; Cruz-Albrecht, J. M.; Hussain, T.; Srinivasa, N.; Lu, W. A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications. *Nano Lett.* **2012**, *12*, 389–395.
[69] Kawahara, A.; Azuma, R.; Ikeda, Y.; Kawai, K.; Katoh, Y.; Tanabe, K.; Nakamura, T.; Sumimoto, Y.; Yamada, N.; Nakai, N. et al. An 8Mb multi-layered cross-point ReRAM macro with 443MB/s write throughput. In *Proceedings of 2012 IEEE International Solid-State Circuits Conference*, San Francisco, CA, USA, 2012, pp 432–434.
[70] Lee, H. D.; Kim, S. G.; Cho, K.; Hwang, H.; Choi, H.; Lee, J.; Lee, S. H.; Lee, H. J.; Suh, J.; Chung, S. O. et al. Integration of $4F_2$ selector-less crossbar array 2Mb ReRAM based on transition metal oxides for high density memory applications. In *Proceedings of 2012 Symposium on VLSI Technology*, Honolulu, HI, USA, 2012, pp 151–152.
[71] Hsieh, M. C.; Liao, Y. C.; Chin, Y. W.; Lien, C. H.; Chang, T. S.; Chih, Y. D.; Natarajan, S.; Tsai, M. J.; King, Y. C.; Lin, C. J. Ultra high density 3D via RRAM in pure 28nm CMOS process. In *Proceedings of 2013 IEEE International Electron Devices Meeting*, Washington, DC, USA, 2013, pp 10.3.1–10.3.4.
[72] Fackenthal, R.; Kitagawa, M.; Otsuka, W.; Prall, K.; Mills, D.; Tsutsui, K.; Javanifard, J.; Tedrow, K.; Tsushima, T.; Shibahara, Y. et al. A 16Gb ReRAM with 200MB/s write and 1GB/s read in 27nm technology. In *Proceedings of 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, San Francisco, CA, USA, 2014, pp 338–339.
[73] Yao, P.; Wu, H. Q.; Gao, B.; Eryilmaz, S. B.; Huang, X. Y.; Zhang, W. Q.; Zhang, Q. T.; Deng, N.; Shi, L. P.; Wong, H. S. P. et al. Face classification using electronic synapses. *Nat. Commun.* **2017**, *8*, 15199.
[74] Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M. et al. Mastering the game of go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489.
[75] Markram, H. The human brain project. *Sci. Am.* **2012**, *306*, 50–55.
[76] Markram, H.; Lübke, J.; Frotscher, M.; Sakmann, B. Regulation of

synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **1997**, *275*, 213–215.

[77] Chen, Y. H.; Krishna, T.; Emer, J. S.; Sze, V. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE J. Solid-State Circuits* **2017**, *52*, 127–138.

[78] Jouppi, N. P.; Young, C.; Patil, N.; Patterson, D.; Agrawal, G.; Bajwa, R.; Bates, S.; Bhatia, S.; Boden, N.; Borchers, A. et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, Toronto, ON, Canada, **2017**, pp 1–12.

[79] Sheridan, P. M.; Cai, F. X.; Du, C.; Ma, W.; Zhang, Z. Y.; Lu, W. D. Sparse coding with memristor networks. *Nat. Nanotechnol.* **2017**, *12*, 784–789.

[80] Chen, B.; Cai, F. X.; Zhou, J. T.; Ma, W.; Sheridan, P.; Lu, W. D. Efficient in-memory computing architecture based on crossbar arrays. In *Proceedings of 2015 IEEE International Electron Devices Meeting*, Washington, DC, USA, 2015, pp 17.5.1–17.5.4.

[81] Xia, Q. F.; Robinett, W.; Cumbie, M. W.; Banerjee, N.; Cardinali, T. J.; Yang, J. J.; Wu, W.; Li, X. M.; Tong, W. M.; Strukov, D. B. et al. Memristor-CMOS hybrid integrated circuits for reconfigurable logic. *Nano Lett.* **2009**, *9*, 3640–3645.

[82] Burr, G. W.; Shelby, R. M.; Sidler, S.; Di Nolfo, C.; Jang, J.; Boybat, I.; Shenoy, R. S.; Narayanan, P.; Virwani, K.; Giacometti, E. U. et al. Experimental demonstration and tolerancing of a large-scale neural network (165 000 Synapses) using phase-change memory as the synaptic weight element. *IEEE Trans. Electron Devices* **2015**, *62*, 3498–3507.

[83] Kawahara, A.; Azuma, R.; Ikeda, Y.; Kawai, K.; Katoh, Y.; Hayakawa, Y.; Tsuji, K.; Yoneda, S.; Himeno, A.; Shimakawa, K. et al. An 8 Mb multi-layered cross-point ReRAM macro with 443 MB/s write throughput. *IEEE J. Solid-State Circuits* **2013**, *48*, 178–185.

[84] Huang, J. J.; Tseng, Y. M.; Hsu, C. W.; Hou, T. H. Bipolar nonlinear Ni/TiO_2/Ni selector for 1S1R crossbar array applications. *IEEE Electron Device Lett.* **2011**, *32*, 1427–1429.

[85] Shin, J.; Kim, I.; Biju, K. P.; Jo, M.; Park, J.; Lee, J.; Jung, S.; Lee, W.; Kim, S.; Park, S. et al. TiO_2-based metal-insulator-metal selection device for bipolar resistive random access memory cross-point application. *J. Appl. Phys.* **2011**, *109*, 033712.

[86] Lee, W.; Park, J.; Shin, J.; Woo, J.; Kim, S.; Choi, G.; Jung, S.; Park, S.; Lee, D.; Cha, E. et al. Varistor-type bidirectional switch (J_{MAX} > 10^7 A/cm^2, selectivity ~ 10^4) for 3D bipolar resistive memory arrays. In *Proceedings of 2012 Symposium on VLSI Technology*, Honolulu, HI, USA, 2012, pp 37–38.

[87] Woo, J.; Song, J.; Moon, K.; Lee, J. H.; Cha, E.; Prakash, A.; Lee, D.; Lee, S.; Park, J.; Koo, Y. et al. Electrical and reliability characteristics of a scaled (~ 30 nm) tunnel barrier selector (W/Ta_2O_5/TaO_x/TiO_2/TiN) with excellent performance (J_{MAX} > 10^7 A/cm^2). In *Proceedings of 2014 Symposium on VLSI Technology* (*VLSI-Technology*): *Digest of Technical Papers*, Honolulu, HI, USA, 2014.

[88] Ovshinsky, S. R. Reversible electrical switching phenomena in disordered structures. *Phys. Rev. Lett.* **1968**, *21*, 1450–1453.

[89] Lee, M. J.; Lee, D.; Cho, S. H.; Hur, J. H.; Lee, S. M.; Seo, D. H.; Kim, D. S.; Yang, M. S.; Lee, S.; Hwang, E. et al. A plasma-treated chalcogenide switch device for stackable scalable 3D nanoscale memory. *Nat. Commun.* **2013**, *4*, 2629.

[90] Son, M.; Lee, J.; Park, J.; Shin, J.; Choi, G.; Jung, S.; Lee, W.; Kim, S.; Park, S.; Hwang, H. Excellent selector characteristics of nanoscale VO_2 for high-density bipolar ReRAM applications. *IEEE Electron Device Lett.* **2011**, *32*, 1579–1581.

[91] Kim, W. G.; Lee, H. M.; Kim, B. Y.; Jung, K. H.; Seong, T. G.; Kim, S.; Jung, H. C.; Kim, H. J.; Yoo, J. H.; Lee, H. D. et al. NbO_2-based low power and cost effective 1S1R switching for high density cross point ReRAM application. In *Proceedings of 2014 Symposium on VLSI Technology* (*VLSI-Technology*): *Digest of Technical Papers*, Honolulu, HI, USA, 2014.

[92] Jo, S. H.; Kumar, T.; Narayanan, S.; Lu, W. D.; Nazarian, H. 3D-stackable crossbar resistive memory based on field assisted superlinear threshold (FAST) selector. In *Proceedings of 2014 IEEE International Electron Devices Meeting*, San Francisco, CA, USA, 2015, pp 6.7.1–6.7.4.

[93] Jo, S. H.; Kumar, T.; Narayanan, S.; Nazarian, H. Cross-point resistive RAM based on field-assisted superlinear threshold selector. *IEEE Trans. Electron Devices* **2015**, *62*, 3477–3481.

[94] Li, C.; Belkin, D.; Li, Y. N.; Yan, P.; Hu, M.; Ge, N.; Jiang, H.; Montgomery, E.; Lin, P.; Wang, Z. et al. Efficient and self-adaptive *in-situ* learning in multilayer memristor neural networks. *Nat. Commun.* **2018**, *9*, 2385.

[95] Shafiee, A.; Nag, A.; Muralimanohar, N.; Balasubramonian, R.; Strachan, J. P.; Hu, M.; Williams, R. S.; Srikumar, V. ISAAC: A convolutional neural network accelerator with *in-situ* analog arithmetic in crossbars. In *Proceedings of 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture*, Seoul, South Korea, 2016, pp 14–26.

[96] Gokmen, T.; Vlasov, Y. Acceleration of deep neural network training with resistive cross-point devices: Design considerations. *Front. Neurosci.* **2016**, *10*, 333.

[97] Agarwal, S.; Quach, T. T.; Parekh, O.; Hsia, A. H.; DeBenedictis, E. P.; James, C. D.; Marinella, M. J.; Aimone, J. B. Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding. *Front. Neurosci.* **2016**, *9*, 484.

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

[99] Ohno, T.; Hasegawa, T.; Tsuruoka, T.; Terabe, K.; Gimzewski, J. K.; Aono, M. Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. *Nat. Mater.* **2011**, *10*, 591–595.

[100] Wang, Z. Q.; Xu, H. Y.; Li, X. H.; Yu, H.; Liu, Y. C.; Zhu, X. J. Synaptic learning and memory functions achieved using oxygen ion migration/diffusion in an amorphous InGaZnO memristor. *Adv. Funct. Mater.* **2012**, *22*, 2759–2765.

[101] Kim, S.; Du, C.; Sheridan, P.; Ma, W.; Choi, S.; Lu, W. D. Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity. *Nano Lett.* **2015**, *15*, 2203–2211.

[102] Zidan, M. A.; Jeong, Y. J.; Lu, W. D. Temporal learning using second-order memristors. *IEEE Trans. Nanotechnol.* **2017**, *16*, 721–723.

[103] Pickett, M. D.; Medeiros-Ribeiro, G.; Williams, R. S. A scalable neuristor built with Mott memristors. *Nat. Mater.* **2013**, *12*, 114–117.

[104] Stoliar, P.; Tranchant, J.; Corraze, B.; Janod, E.; Besland, M. P.; Tesler, F.; Rozenberg, M.; Cario, L. A leaky-integrate-and-fire neuron analog realized with a Mott insulator. *Adv. Funct. Mater.* **2017**, *27*, 1604740.

[105] Feng, S.; Zhou, H. Y.; Dong, H. B. Using deep neural network with small dataset to predict material defects. *Mater. Des.* **2019**, *162*, 300–310.

[106] Lukoševičius, M.; Jaeger, H. Reservoir computing approaches to recurrent neural network training. *Comput. Sci. Rev.* **2009**, *3*, 127–149.

[107] Torrejon, J.; Riou, M.; Araujo, F. A.; Tsunegi, S.; Khalsa, G.; Querlioz, D.; Bortolotti, P.; Cros, V.; Yakushiji, K.; Fukushima, A. et al. Neuromorphic computing with nanoscale spintronic oscillators. *Nature* **2017**, *547*, 428–431.

[108] Jaeger, H.; Haas, H. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science* **2004**, *304*, 78–80.

[109] Jaeger, H.; Lukoševičius, M.; Popovici, D.; Siewert, U. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Networks* **2007**, *20*, 335–352.

[110] Verstraeten, D.; Schrauwen, B.; Stroobandt, D. Reservoir-based techniques for speech recognition. In *Proceedings of 2006 IEEE International Joint Conference on Neural Network Proceedings*, Vancouver, BC, Canada, 2006, pp 1050–1053.

[111] Du, C.; Cai, F. X.; Zidan, M. A.; Ma, W.; Lee, S. H.; Lu, W. D. Reservoir computing using dynamic memristors for temporal information processing. *Nat. Commun.* **2017**, *8*, 2204.

[112] Appeltant, L.; Soriano, M. C.; Van Der Sande, G.; Danckaert, J.; Massar, S.; Dambre, J.; Schrauwen, B.; Mirasso, C. R.; Fischer, I. Information processing using a single dynamical node as complex system. *Nat. Commun.* **2011**, *2*, 468.

[113] Moon, J.; Ma, W.; Shin, J. H.; Cai, F. X.; Du, C.; Lee, S. H.; Lu, W. D.

Temporal data classification and forecasting using a memristor-based reservoir computing system. *Nat. Electron.* **2019**, *2*, 480–487.

[114] Borghetti, J.; Snider, G. S.; Kuekes, P. J.; Yang, J. J.; Stewart, D. R.; Williams, R. S. 'Memristive' switches enable 'stateful' logic operations via material implication. *Nature* **2010**, *464*, 873–876.

[115] Linn, E.; Rosezin, R.; Tappertzhofen, S.; Böttger, U.; Waser, R. Beyond von neumann—logic operations in passive crossbar arrays alongside memory operations. *Nanotechnology* **2012**, *23*, 305205.

[116] Gao, L. G.; Alibart, F.; Strukov, D. B. Programmable CMOS/memristor threshold logic. *IEEE Trans. Nanotechnol.* **2013**, *12*, 115–119.

[117] James, A. P.; Francis, L. R. V. J.; Kumar, D. S. Resistive threshold logic. *IEEE Trans. Very Large Scale Integr. Syst.* **2014**, *22*, 190–195.

[118] Jeong, D. S.; Kim, K. M.; Kim, S.; Choi, B. J.; Hwang, C. S. Memristors for energy-efficient new computing paradigms. *Adv. Electron. Mater.* **2016**, *2*, 1600090.

[119] Balatti, S.; Ambrogio, S.; Ielmini, D. Normally-off logic based on resistive switches—Part I: Logic gates. *IEEE Trans. Electron Devices* **2015**, *62*, 1831–1838.

[120] Huang, P.; Kang, J. F.; Zhao, Y. D.; Chen, S. J.; Han, R. Z.; Zhou, Z.; Chen, Z.; Ma, W. J.; Li, M.; Liu, L. F. et al. Reconfigurable nonvolatile logic operations in resistance switching crossbar array for large-scale circuits. *Adv. Mater.* **2016**, *28*, 9758–9764.

[121] Strukov, D. B.; Likharev, K. K. CMOL FPGA: A reconfigurable architecture for hybrid digital circuits with two-terminal nanodevices. *Nanotechnology* **2005**, *16*, 888–900.

[122] Snider, G. S.; Williams, R. S. Nano/CMOS Architectures Using a Field-Programmable Nanowire Interconnect. *Nanotechnology* **2007**, *18*, 035204.

[123] Menzel, S.; Kaupmann, P.; Waser, R. Understanding filamentary growth in electrochemical metallization memory cells using kinetic monte carlo simulations. *Nanoscale* **2015**, *7*, 12673–12681.

[124] Qin, S. J.; Liu, Z.; Zhang, G.; Zhang, J. Y.; Sun, Y. P.; Wu, H. Q.; Qian, H.; Yu, Z. P. Atomistic study of dynamics for metallic filament growth in conductive-bridge random access memory. *Phys. Chem. Chem. Phys.* **2015**, *17*, 8627–8632.

[125] Liu, Q.; Long, S. B.; Lv, H. B.; Wang, W.; Niu, J. B.; Huo, Z. L.; Chen, J. N.; Liu, M. 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.

[126] Lee, J.; Du, C.; Sun, K.; Kioupakis, E.; Lu, W. D. Tuning ionic transport in memristive devices by graphene with engineered nanopores. *ACS Nano* **2016**, *10*, 3571–3579.

[127] Choi, S.; Tan, S. H.; Li, Z. F.; Kim, Y.; Choi, C.; Chen, P. Y.; Yeon, H.; Yu, S. M.; Kim, J. SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations. *Nat. Mater.* **2018**, *17*, 335–340.

[128] Chen, P. Y.; Peng, X. C.; Yu, S. M. NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures. In *Proceedings of 2017 IEEE International Electron Devices Meeting*, San Francisco, CA, USA, 2017, pp 6.1.1–6.1.4.

[129] Sun, X. Y.; Yu, S. M. Impact of non-ideal characteristics of resistive synaptic devices on implementing convolutional neural networks. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2019**, *9*, 570–579.

[130] Ambrogio, S.; Narayanan, P.; Tsai, H.; Shelby, R. M.; Boybat, I.; Di Nolfo, C.; Sidler, S.; Giordano, M.; Bodini, M.; Farinha, N. C. P. et al. Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **2018**, *558*, 60–67.

[131] Le Gallo, M.; Sebastian, A.; Mathis, R.; Manica, M.; Giefers, H.; Tuma, T.; Bekas, C.; Curioni, A.; Eleftheriou, E. Mixed-precision in-memory computing. *Nat. Electron.* **2018**, *1*, 246–253.

[132] Boybat, I.; Le Gallo, M.; Nandakumar, S. R.; Moraitis, T.; Parnell, T.; Tuma, T.; Rajendran, B.; Leblebici, Y.; Sebastian, A.; Eleftheriou, E. Neuromorphic computing with multi-memristive synapses. *Nat. Commun.* **2018**, *9*, 2514.

[133] Zidan, M. A.; Jeong, Y. J.; Lee, J.; Chen, B.; Huang, S.; Kushner, M. J.; Lu, W. D. A general memristor-based partial differential equation solver. *Nat. Electron.* **2018**, *1*, 411–420.

[134] Gao, L. G.; Chen, P. Y.; Yu, S. M. Programming protocol optimization for analog weight tuning in resistive memories. *IEEE Electron Device Lett.* **2015**, *36*, 1157–1159.

[135] Zidan, M. A.; Jeong, Y.; Shin, J. H.; Du, C.; Zhang, Z. Y.; Lu, W. D. Field-Programmable Crossbar Array (FPCA) for reconfigurable computing. *IEEE Trans. Multi-Scale Comput. Syst.* **2018**, *4*, 698–710.

[136] Jacob, B.; Kligys, S.; Chen, B.; Zhu, M. L.; Tang, M.; Howard, A.; Adam, H.; Kalenichenko, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp 2704–2713.

[137] Hubara, I.; Courbariaux, M.; Soudry, D.; El-Yaniv, R.; Bengio, Y. Quantized neural networks: Training neural networks with low precision weights and activations. *J. Mach. Learn. Res.* **2018**, *18*, 1–30.

[138] Xu, X. W.; Ding, Y. K.; Hu, S. X.; Niemier, M.; Cong, J.; Hu, Y.; Shi, Y. Y. Scaling for edge inference of deep neural networks. *Nat. Electron.* **2018**, *1*, 216–222.