Hardware Implementation of Neuromorphic Computing Using Large-Scale Memristor Crossbar Arrays

Yesheng Li and Kah-Wee Ang* 

1. Introduction
The exponential growth of data is increasingly consuming tremendous computing resources, which imposes significant challenges to traditional von Neumann architecture-based digital computers. There are two major obstacles that impede the further improvement of traditional computing. First, with the end of Moore’s law insight and the increased fabrication cost, advances in computing capabilities such as speed and energy consumption can no longer be supported by device downscaling alone. [1] The second obstacle for today’s traditional digital computers originates from the high (energy and speed) costs associated with frequent data movement between the memory and the processor, i.e., the von Neumann bottleneck. [2] With the application becoming more data centric, the performance gap between the memory and the processing unit (refer to “memory wall”) will enlarge and even exacerbate. [3] Therefore, new computing architectures are needed to meet the insatiable need for data and information.

Brain-inspired computing or neuromorphic computing through emulating the structure and/or working principle of the biological brain has emerged as a new computing paradigm to improve computing efficiency. Parallel processing and in-memory computing are the two key features of neuromorphic computing. For in-memory computing, the calculations are conducted in situ where the data are located. This approach radically changes the von Neumann architecture and thus the energy burdens and latency can be removed completely. Many computing tasks can be massively accelerated by combining in-memory computing and parallel processing. To obtain these fantastic advantages, extensive brain-inspired artificial neural networks (ANNs) using CMOS integrated circuits have been developed, for example, IBM’s TrueNorth chip with 1 million integrate-and-fire (IF) silicon neurons and 256 million silicon synapses using conventional 40 nm CMOS technology; [4] the Loihi self-learning chip with 130 thousand IF silicon neurons and 130 million silicon synapses using 14 nm CMOS technology; [5] and the recent report on CMOS-based spiking neural networks (SNNs) which use spike timing-based algorithm to emulate the functionality of the human olfactory bulb with high noise tolerance. [6] These CMOS-based ANNs are usually constructed with complex circuit architectures and expensive CMOS chip technologies, which face the same issues of scaling limit and energy consumption.

Neuromorphic computing systems (NCSs) built using beyond-CMOS devices and technology offer a more attractive solution to overcome the challenges of classical computing architecture. [7–11] Different devices, mainly synaptic transistors and memristors, have been reported with the ability to mimic neuromorphic functions. [7–11] The synaptic transistors, i.e., three-/multi-terminal transistor-based devices, include floating gated, ferroelectric gated, electrolyte gated, and optoelectronic synaptic transistors, in which the synaptic weight (i.e., channel...
conductance) can be tuned by the control terminal (gate electrode) and the transduction terminals (source-drain electrodes).\cite{12,13} In contrast, memristors are two-terminal resistive switching devices that can maintain their internal resistance states depending on the history of applied voltages/currents.\cite{14,15} Compared with synaptic transistors, the memristors, which can offer high-density integration, stand out as a promising candidate for in-memory computing. The brain-inspired algorithms are implemented mainly by vector-matrix multiplications to calculate the neuronal outputs. Naturally, the high-density memristor crossbar arrays (CBAs) can perform the parallel vector-matrix multiplication with ultra-low energy.\cite{16} Moreover, the conductance of the memristor can be flexibly tuned by modulating the parameters of the applied voltage pulses, thus providing great potential to construct adaptive systems with online learning capability. These desirable properties make memristors particularly attractive as neuromorphic devices (i.e., synapses and neurons). Nowadays, a large variety of memristive devices with different mechanisms, such as electrochemical metalization mechanism (ECM),\cite{9,17–19} valence change mechanism (VCM),\cite{20–23} phase change mechanism,\cite{24–27} thermochemical mechanism (TCM),\cite{28,29} Mott transition,\cite{30,31} photonic-induced switching,\cite{32,33} ferroelectric transition,\cite{34,35} and magnetic transition,\cite{36,37} are investigated to mimic synaptic and neuronal functions. The variety of memristive devices facilitate the development of high-performance, large-scale, and energy-efficient NCSs. However, most studies on neuromorphic computing are focused on individual devices or small CBAs, and implementations such as image recognition are mainly based on array-level simulation.\cite{38–48} This is because the implementation of massively parallel and highly efficient NCs would require large-scale networks, i.e., large-scale memristor CBAs. Moreover, building such CBAs imposes stringent requirements both on device performances and on array parameters. For example, low device variation and linear and symmetric weight updating are needed for efficient training.\cite{41} In addition, sneak current and wire resistance issues cannot be ignored and must be addressed in large-scale and high-density memristor CBAs.\cite{42,43} Addressing these challenges is critical to the realization of hardware NCs.

In this Progress Report, we will review the stringent requirements and the possible solutions for building large-scale and high-density memristor CBAs, as well as the recent progress on the hardware implementation of NCs based on memristor CBAs. We will begin with artificial neural systems which consist of different types of ANNs built with artificial memristive synapses and neurons. Following that, we will comprehensively discuss the key requirements on the implementation of large-scale ANNs, as well as the solutions to address the challenges associated with nonideal device performance and large-scale memristor CBAs’ fabrication and integration. Thereafter, we will present the recent progress on the experimental implementation of neuromorphic computing hardware. Finally, a summary and outlook will be discussed to provoke further ongoing research in this emerging field.

2. Artificial Neural Systems

The human nervous system, where intelligence resides, supports various intelligent functions such as memory and forgetting, learning, and decision-making. The comparison of the human nervous system and the artificial neural system is shown in Figure 1,\cite{44} in which both systems are organized at multiple levels. As for the human nervous system, three basic levels can be distinguished (Figure 1a). At the first level, different types of neural networks are organized into a vast network to support different computational functions, such as vision, gustation, and hearing. At the second level, the neuron comprises a soma, dendrites, and an axon forming the basic unit, which is connected by synapses. Different types of ion channels can be viewed as the third level, which is the basis of electrical activities in neurons. The information is transferred and processed by controlling the opening and closing of the ion channel. Similarly, a typical artificial neural system on artificial intelligence (AI) chip also consists of different types of ANNs. These ANNs can be mapped onto CBAs, where neuromorphic devices are used as the building blocks. For example, the filament-type memristors, in which conducting channels are formed by ion movements as driven by the electric field and/or joule heating, can be used to mimic the artificial synapses and neurons, as shown in Figure 1b.

Various network architectures have been proposed, among which deep neural networks (DNNs) and SNNs are the most prevalent ANNs. DNNs, which usually use backpropagation algorithms, show fast speed and intensive computing capacity and have achieved extensive success on NCs.\cite{41} The multiply accumulate operation is the heart of DNNs forward inference and training, which can be performed using memristor CBAs. The outputs of typical DNNs have real values and are processed in synchronous time steps. DNN can include multilayer perceptron (MLP),\cite{36,37–47} convolutional neural network (CNN),\cite{48–51} binarized neural network (BNN),\cite{52,53} quantized neural network (QNN),\cite{54} and recurrent neural network (RNN).\cite{55–57} Among them, MLP and CNN are the most popular for hardware implementation of ANN, BNN with binarized weight, and activation function can offer better tolerance for device imperfections (such as device variation). In contrast to DNNs, SNNs are implemented based on the spike-time-dependent-plasticity (STDP) algorithm and use asynchronous spikes with identical amplitude and duration. SNNs are bioplausible, real time, spatiotemporal, low power, and perceived to be more biorealistic than DNNs. However, SNNs still fall short of DNNs with lower accuracy than conventional DNNs, but SNNs typically require much fewer operations and are better candidates for processing spatiotemporal data.\cite{41} Memristor-based hardware SNNs have been developed but with less progress than DNNs.\cite{58,59} The investigation of hardware SNNs is still at the nascent stage, in which algorithms and datasets, etc. could be further enhanced. These different ANNs can be implemented using memristive devices to carry out high-performance in-memory computing in parallel by directly utilizing physical laws, in which memristors are used as the building blocks to emulate the functions of synapse and neurons.

2.1. Artificial Memristive Synapse

Memristive devices based on various mechanisms, such as ECM,\cite{9,17–19} VCM,\cite{20–23} PCM,\cite{24–27} TCM,\cite{28,29} Mott
transition, photonic-induced switching, ferroelectric transition, and magnetic transition are investigated as emerging devices to mimic synaptic functions. Among them, ECM-based, VCM-based, and phase change-based memristors are most prevalent to fabricate large-scale CBAs for hardware implementation of ANNs.

As shown in Figure 2a, the ECM-based memristor also refers to conductive bridge random access memory (CBRAM) and possesses a two-terminal metal/insulator/metal (MIM) structure, in which active metals (such as Ag, Cu, and Ti) are used as top electrodes (TEs) and inert metals (such as Pt, Pd, Au, and TiN) are used as bottom electrodes (BEs). In the ECM-based memristive device, the change in resistance is caused by the migration of cation, i.e., metal ions, and the formation of metal conductive filaments (CFs) between the electrodes. When a positive voltage is applied to the anode (i.e., active metal), the metal ions are generated via an electrochemical reaction (oxidation reaction) at the surface of the electrode or/and in the insulating layer. The metal ions then move toward the cathode as driven by the electric field and are reduced at the surface of the cathode or inside the insulating layer, giving rise to various filament formation modes. With the continually applied positive voltage, CFs finally form to bridge the two electrodes, leading to device switching from high resistance state (HRS) to low resistance state (LRS). Conversely, if a negative voltage is applied to the anode, the metal ions will move in the opposite direction, leading to the rupture of CFs; thus, the device switches from LRS to HRS. The ECM devices usually possess the advantages of high scalability, ultralow operating voltages, and a large dynamic resistance range.

VCM is also one of the most common mechanisms observed in memristive devices, especially in oxide-based memristors. As shown in Figure 2b, in the VCM device, the resistive switching behavior can also be attributed to the formation/dissolution of the CF. The difference is that it is the migration of anion (mainly oxygen ions) and the subsequent redox reaction that leads to the formation of CFs. Anion movement is driven by external stimuli, such as electric fields or temperature gradients. For example, when a positive voltage is applied to the anode, oxygen vacancies move to the cathode, and oxygen ions move toward the anode, increasing the conductance of the insulating layer, i.e., LRS, and vice versa.

With their mechanism of filament formation and dissolution that are associated with ion migration, ECM- and VCM-based memristors are suitable for brain-inspired computing. However, due to the strong stochastic behaviors that exist in

Figure 1. A high-level comparison of the human nervous system and an artificial neural system built with emerging neuromorphic devices. a) The human nervous system has different types of (I) neural networks whose basic functional elements are (II) neurons and (III) synapses, in which different types of ion channels underlie electrical neuronal activity. b) Similarly, the artificial neural system on AI chip is composed of different types of (I) ANNs that can be mapped with CBAs constructed by artificial (II) neurons and synapses; these neuromorphic devices can be different memristive devices (such as ECM, VCM, and phase change-based memristor); (III) ion migration-based memristive devices whose operation mechanism can be CFs associated with electrically induced ion movements as in the case of ECM- and VCM-based memristors. Reproduced with permission. Copyright 2019, Wiley-VCH.
the formation and rupture of CF, in which the shape and location of the CF vary from cycle to cycle and device to device, they often suffer from inevitable device variations and weight update nonlinearity/asymmetry.

The PCM may be the most mature memristor technology to date and is one of the few types that has been commercialized. The set and reset processes in PCM are based on the crystallization and amorphization of phase change materials (Figure 2c).

Figure 2. Artificial synapse using memristive devices. a–c) Three typical memristors used in the synaptic CBAs: schematic showing a) ECM/CBRAM-based memristor, b) VCM-based memristor, and c) phase change-based memristor. d–i) STDP: overlapping STDP realization through overlapping prespike and postspike in d–e) VCM memristor and f) phase change memristor; g–i) nonoverlapping STDP realization by second-order memristor, in which the size of the conductive channels (the memristor conductance, $w$) can be considered as the first-order state variable and the local temperature of the device can be considered as a second-order state variable (denoted as $T$). j) Frequency-dependent learning rule, SRDP. k–l) LTP and LTD emulated by memristive devices: k) typical nonlinear LTP and LTD in Ag/Zr$_{0.5}$Hf$_{0.5}$O$_2$; graphene oxide quantum dots/Ag memristive device; l) schematic showing the ideal linear and symmetric LTP and LTD. d–e) Reproduced with permission. Copyright 2017, American Chemical Society. f) Reproduced with permission. Copyright 2012, American Chemical Society. g–i) Reproduced with permission. Copyright 2015, American Chemical Society. j) Reproduced with permission. Copyright 2019, Wiley-VCH. k) Reproduced with permission. Copyright 2018, Wiley-VCH.
transformed into a 1T (1T′) phase with a twisted octahedral structure. The MoS₂ memristors have been developed by local 2H-1T′ phase transitions via applying the electric field, controlling the migration of Li⁺ ions with an electric field, or photo-irradiation. A MoTe₂ memristor is also explored by electrically driven 2H to 1T′ phase transformation. The phase changes in TMDs can be induced by different factors (electric field, ion migration, or photo-irradiation) and usually occur between crystal phases, which are different from conventional phase change between the crystal phase and amorphous phase through joule heating. In particular, wafer-scale monolayer MoS₂ has been successfully synthesized by chemical vapor deposition (CVD), which is important for building large-scale memristive CBAs. The investigation of 2D materials-based PCM may open a new avenue to develop 2D synaptic memristors for ANN implementation.

All these memristive devices are successfully used to emulate synaptic functions. Among the multisynaptic functions, the basic learning rules of STDP, spike-rate-dependent plasticity (SRDP), and long-term potentiation (LTP), and long-term depression (LTD) are the functions that matter most for hardware implementation of ANNs. As one of the key learning rules in biosynapses, STDP is the prevalent weight updating rules used in SNNs. STDP can be achieved by two methods: one is overlapping STDP and the other is nonoverlapping STDP. The overlapping STDP, as shown in Figure 2d–f, is implemented through overlapping the manual and well-designed programming pulses. In this type of STDP, the spike timing information is encoded into the amplitude or duration of the overlapped pulses applied on the devices. With this approach, the phenomenological bio-like STDP can be emulated in the nonvolatile memristor, such as ion-migrated memristor (Figure 2e) and phase-change-based memristor (Figure 2f).

Generally, any one of the presynaptic and postsynaptic pulses cannot individually modulate the memristor conductance because they are both below the threshold. By overlapping the presynaptic and postsynaptic pulses at a given time interval (Δt), the potential drop across the memristor can reach the threshold that induces a change in device conductance. The spike time information of this kind of STDP is encoded into the net pulse amplitude, as the change of the net pulse amplitude is a function of Δt. However, the STDP in biological systems is simply achieved by nonoverlapping spikes and controlled by synaptic activity. That is to say, bio-STDP is controlled by the frequency and relative timing of the spikes instead of the amplitude and duration of the spikes. Moreover, in bio-synapse, it usually contains both long-term dynamics and short-term dynamics. Nonoverlapping usually cannot be observed in most nonvolatile memristors, because they lack short-term dynamics. The use of second-order memristive devices with both long-term and short-term dynamics or combining a volatile memristive device with a nonvolatile memristive device is demonstrated with the ability to implement non-overlapping STDP. As shown in Figure 2g–i, Kim et al. demonstrated the successful emulation of nonoverlapping STDP in a second-order memristor with a structure of Pd/Ta₂O₅/Ta₂O₅/Pd. In such a device, the first-order state variable, i.e., long-term dynamic, is the size of the conductive channels; the second-state variable, i.e., short-term dynamic, is the local temperature of the device. Wang et al. also successfully demonstrated nonoverlapping STDP by connecting a diffusive memristor and a nonvolatile memristor in series. In this combined circuit, the diffusive memristor with short-term dynamics can perform the threshold switching, the nonvolatile memristor can conduct the nonvolatile bipolar resistive switching.

SRDP is another important learning rule, in which the potentiation and depression of synaptic plasticity are controlled by presynaptic firing rate or frequency. As shown in Figure 2j, when a high frequency of 2 MHz is applied, the current is potentiation, whereas when using low frequencies of 0.2 and 0.1 MHz, the current is depression. In another work, Li et al. demonstrated the SRDP phenomenon in the Ag/AgInSbTe/Ag memristor. When the frequency is below 50 kHz, the depression of synaptic weight can be observed, whereas when the frequency is higher than 50 kHz, the potentiation process can be found.

LTD and LTP, for which the conductance values representing synaptic weights are enhanced (LTP) or depressed (LTD) according to pulse number, are the basic requirements for synaptic emulation and especially the implementation of DNNs. Figure 2k shows a typical LTD and LTP induced by positive and negative pulse sequences in Ag/Zr₀.₅Hf₀.₅O₂: graphene oxide quantum dots/Ag memristor, respectively. Ideally, memristive devices exhibiting linear and symmetric increase/decrease in conductance over a large dynamic range (see Figure 2l) are the key needs for hardware implementation of DNNs with high training efficiency. However, most memristive synapses do not follow these needs but show nonlinearity or asymmetry in LTD and LTP, as shown in Figure 2k. These nonideal properties could be improved by some measures which will be discussed in the later section.

2.2. Artificial Memristive Neuron

The implementation of artificial neurons via CMOS circuits with integrated transistors is the most common way so far. However, CMOS-based neurons encounter the concerns of low-density integration and high power consumption. Developing scalable neurons, such as memristive device-based neurons, is necessary to construct large-scale NCSs. Unlike the extensively reported memristive synapses, the development of memristor-based artificial neurons is limited. Artificial neurons based on VCM, ECM, phase change, and Mott transition have been investigated, whereas artificial neurons based on other types of memristors, such as ferroelectric transition and magnetic transition-based memristors, are rarely reported. Various synaptic functions can be mimicked by a single memristor via the tuning of synaptic weight represented by its conductance. However, a single memristive device is usually not sufficient to perform the functions of a neuron, but a hybrid circuit comprising the memristor, as well as the transistor or capacitor, is usually required to mimic neurons. There are mainly two types of artificial neurons according to different neuron models. One is bioplasuble neurons including IF and leaky-integrate-and-fire (LIF) neurons based on the threshold process. Another one is biophysical neurons, mainly referring to Hodgkin–Huxley (HH) neurons, which follow the understanding of the biophysical dynamic behaviors of ion channels
embedded in the neuron membrane. Bioplausible neurons usually possess a simpler circuit structure than that of biophysical neurons and are more popular in neuromorphic computing.

The key to the implementation of bioplausible neurons is to realize local graded potential (LGP) and threshold process. There are mainly three approaches to implement neurons based on memristive devices. The first way is the implementation of the neuron using threshold-less memristive devices combined with comparators.\textsuperscript{50,80,83,84} Figure 3a shows an artificial neuron based on the memristor with two comparators (OP1 and OP2) and a pulse generator (P).\textsuperscript{84} In this neuron, memristor 2 is a volatile second-order memristor, which is utilized to realize the LGP. The pulse generator is needed to trigger the fire function when LGP reaches the threshold value, and the comparator is used to set the threshold. Using the volatile memristive device, the short-term dynamics can be introduced in membrane potential; thus, the leaky function can be realized. In this threshold-less type neuron, the comparator as well as the pulse generator are implemented with a number of transistors, leading to a critical concern for high-density integration. This issue can be greatly improved by the second approach that implements neuron with the threshold memristor and capacitor.\textsuperscript{59,85–87} Because the memristor itself can perform the threshold process, thus, a leaky function can be realized without the external comparator and pulse generator, resulting in a simpler circuit. In this case, the memristor conducts the threshold process, a capacitor in parallel to the memristor implements the integration effect via the charging process. As shown in Figure 3b–d, an LIF neuron is developed with a very simple circuit constructed by a vertical MoS\textsubscript{2}/graphene threshold switching memristor and a paralleled

Figure 3. Neuron implementation with memristive devices. a) The LIF neuron implementation in the threshold-less memristive devices combined with comparators and pulse generator. The artificial neuron consisting of dendrites, soma, and axon. The dendrites receive inputs from the synapses that are mimicked by the nonvolatile memristors (Memristor 1). The LGP generated by the summation of the excitatory postsynaptic potential is mimicked by Memristor 2. Memristor 2 is a second-order memristor with volatile resistive switching behavior. b–d) the LIF neuron implementation in the threshold memristive device with parallel capacitor: b) schematic illustration of an artificial LIF neuron composed of c) a vertical MoS\textsubscript{2}/graphene threshold memristive device and a parallel capacitor; d) the corresponding circuit used to realize the artificial neuron (threshold memristor [TSM]). e,f) the LIF neuron implementation with a single-threshold memristive device: e) the I–V curves of the threshold memristive device, showing fast turn-on and volatile switching behavior, the inset shows the cross-section view of the memristive device under HRTEM inspection; f) the demonstration of LIF behavior under a continuous voltage pulse train on the memristive device. g,h) Biophysical neuron with HH model: g) basic circuit topology of a two-channel active memristor neuron to emulate the neuronal dynamics, a voltage-gated Na\textsuperscript{+} (K\textsuperscript{+}) channel is emulated by a negatively (positively) D.C.-biased active memristor device, which is closely coupled to a local membrane capacitor C\textsubscript{1} (C\textsubscript{2}) and a series load resistor R\textsubscript{L1} (R\textsubscript{L2}); h) SEM image and schematic structure of a typical VO\textsubscript{2} active memristor nano-crossbar device (X\textsubscript{1} or X\textsubscript{2} in (g)), scale bar is 100 nm. a) Reproduced with permission.\textsuperscript{84} Copyright 2018, Wiley-VCH. b–d) Reproduced with permission.\textsuperscript{85} Copyright 2019, Springer Nature. e–f) Reproduced with permission.\textsuperscript{79} Copyright 2018, Wiley-VCH. g–h) Reproduced with permission.\textsuperscript{82} Copyright 2018, Springer Nature.
capacitor[85]. The capacitor integrates the charge, and when its voltage reaches the threshold switching value of the MoS$_2$-/graphene memristor, the neuron fires and generates an output spike. If a single device can emulate the functions of a neuron, the neural circuits would be simplified notably. Recently, a neurotransistor based on silicon nanowire was reported with the ability to merge learning and memory functions dynamically in a single element.[88] Similarly, mimicking neuronal functions using a single memristor, which itself can double up as a capacitor, has also been developed, leading to more simplified and compact neuron circuits.[79,89–91] As shown in Figure 3e–f, through the volatile switching effect, the Ag/FeO$_x$/Pt memristor can mimic various neural functions without the assistance of auxiliary circuits.[79] The multiple neural dynamics result from the formation and automatic retrieval of silver filaments. The memristor itself can perform the neuron functions, which can highly simplify the circuits and thus would greatly benefit the large-scale and high-density neural network.

In contrast, the progress of the implementation of biophysical neurons (i.e., HH neurons) is very limited. This may be because HH neurons require more complex circuits to emulate complex ion channel dynamics.[82,92,93] Yi et al. reported vanadium dioxide (VO$_2$) active memristors as intrinsically stochastic HH neurons.[82] As shown in Figure 3g–h, the VO$_2$ active memristor-based HH neuron circuit is composed of two resistance-coupled relaxation oscillators, each has a VO$_2$ memristor ($X_1$ or $X_2$), a parallel capacitor ($C_1$ or $C_2$), and a load resistor ($R_{11}$ or $R_{12}$).[82] The voltage-gated Na$^+$ and K$^+$ ion channels are emulated by the polarized memristors $X_1$ and $X_2$, respectively. This VO$_2$-based neuron circuit can mimic most biological neuron functions. Recently, an artificial quasi-HH neuron with LIF functions was developed by Huang et al.[93] In this quasi-HH neuron, two W/VO$_2$/PEDOT:PSS/Pt memristive devices are used, in which one performs the integration and leaky functions, whereas another one implements the firing function. Except for the memristors, auxiliary circuits including comparator and timer are needed, showing more complex circuits than the IF- and LIF-based neuron. As discussed earlier, memristor-based artificial neurons have shown the ability to emulate the neurons. However, memristive neurons (such as Ag- and Cu-ion-based ECM) would encounter the common problems of random variation between the switching events and devices and uncontrollable operation speed according to the circuit requirement.

3. Key Requirements and Progress for Large-Scale Memristor CBAs

Some performances, such as thermal stability and wire resistance, show less influence on neuromorphic behavior when individual memristors or small CBAs are used to emulate neuronal especially synaptic functions (e.g., STDP, LTD, LTP). When a number of memristors are integrated into large-scale CBAs to implement ANNs, more stringent requirements are imposed both on device performances and on array parameters. In this case, for example, thermal stability and thermal crosstalk should be noticed, and wire resistance should be taken into account. Enhancing the device performances and addressing the challenges associated with large-scale implementations are critical for the practical use of ANNs.

3.1. Device Performances

Generally, the requirements of memristors used in the ANNs are somewhat similar to conventional memory. The biggest advantage of memristor is that it can perform in-memory computing, which can remove the latency and reduce the power consumption. The writing time and memory capacity of the state-of-the-art memristors (at nanosecond level) are comparable with current mainstream memory technologies: DRAM (<10 ns), SRAM (<1 ns), and flash memory (>100 μs).[94,95] The cell area of memristor can be $4F^2$ ($F$: half pitch), which is much smaller than that of $8F^2$ of transistor-based flash memory. There is still a gap in the reliability, such as endurance, retention, and device variability, between memristor and conventional memory. By satisfying the requirements of cost, fast, small, reliable, and efficient, the memristor could be an alternative to current flash memory.[94]

Besides the basic requirements of long retention, high endurance, small device size, low energy consumption, and small device various, analog switching, linear/symmetric weight update, and large numbers of stable resistance states with wide resistance range are in demand for brain-inspired computing using memristors.[2,96] For SNNs, the ability of memristors to perform STDP is needed. For DNNs, analog resistive switching, superior endurance, high on/off ratio, and long retention are required. In particular, a small device variation and linear/symmetric weight updating are necessary for efficient training; low energy consumption and good scalability are required for large-scale high-density neural networks; and good device stability is desirable for high computing accuracy. These have been the major challenges of memristive brain-inspired computing systems.[41]

3.1.1. Device Downscaling

Miniaturization of the memristor is necessary to upscale the size of memristive ANNs. With the ever-increasing demands in high computing capacity and low energy consumption, it is crucial to integrate small memristors into high-density CBAs. The smallest feature size of the memristor is 2 nm in width with a single-layer density up to 4.5 terabits per square inch using a newly developed nanofin electrode technology.[97] The resistive layer thickness in oxide-based memristors can be scaled below 2 nm by atomic layer deposition (ALD) and the thinnest record is found to be 0.5 nm in the Ta$_2$O$_5$-based memristor.[98–100] However, it is actually challenging to achieve such thin oxide films under uniform and controllable quality for most of today’s thin-film deposition technology.[101] 2D materials, which have an atomically thin body, offer a potential strategy to reduce the thickness down to the atomic level. Moreover, due to the advances in synthesis technology, such as CVD or molecular beam epitaxy (MBE), monolayers to few-layer 2D materials have been successfully prepared,[102–104] which would facilitate the development of an atomically thin memristor or atomristor.[105] The monolayer TMDs-based atomristor (i.e., MoS$_2$, MoSe$_2$, WS$_2$, WSe$_2$) has
been developed, indicating that the thickness of the 2D resistive layer can be scaled to the atomic limit. The thinnest memristor is fabricated with ≈0.33 nm (monolayer) h-BN synthesized by CVD. Furthermore, the whole thickness of a memristor can be as thin as 2 nm using graphene as the electrode. However, such a thin electrode may lead to high wire resistance that should be considered. With the highly scaled memristor, super high-density memristor CBAs may be achieved in years to come.

### 3.1.2. Device Variation

Device variation has been one of the key challenges for the implementation of large-scale memristive neural networks. There are two types of device variations: device-to-device variation (or spatial variation) and cycle-to-cycle variation (or pulse-to-pulse variation or temporal variation). Due to device-to-device variation, different devices in the array will cause different nonlinearity baselines. There will be some noise on top of the nonlinearity baseline due to cycle-to-cycle variation. The device variation is especially obvious in filamentary-type memristors (mainly VCM- or ECM-based memristor), due to stochastic CFs' formation and rupture. Both of these two types of device variations may degrade the learning performance of ANNs.

The device-to-device variation is associated with the random electroforming process, resulting in different filament structures between devices. Thus, it is necessary to avoid the electroforming process in large CBAs. Forming-free switching is particularly important for a selector as it usually operates more frequently than the memristor and some properties such as threshold voltage should match with the memristor. If both the selector and memristor experience large variation, it will increase the difficulty and even cause failure in the operation of CBAs. Many strategies for forming-free devices have been developed. Some studies have demonstrated that reducing the thickness of resistive materials could eliminate the electroforming process as well as the effect of breakdown. Meanwhile, controlling the concentration of defects (such as vacancies) can also result in electroforming free behavior.

Instead of forming the conductive path by the electric field during the first operation, the preformed conductive path during the manufacturing process can be used to avoid the electroforming process. The cycle-to-cycle variation is associated with the unstable CFs or stochastic formation of new CFs. Seeking strategies to control the CF formation/dissolution is an essential step to reduce the cycle-to-cycle variation. The bilayer structure is a popular method adopted to confine the formation and rupture of filaments. This lies with the two layers with different thermal conductivities or oxygen ion-migration activities, resulting in the interface barrier. Thus, the CF would completely rupture in one layer and remain unchanged as residual CF in another layer. The retained CF can play a role in confining the formation of CF, resulting in uniform switching. As shown in Figure 4a, the metal filament is completely ruptured in the SiCN layer and remains in the Al2O3 layer which has a higher thermal conductivity than that of SiCN. With the retained CFs, the subsequent filament can be formed where the previous filament is located. As for the single-layer devices without residual CF, the filament can form at random places. Hence, the switching uniformity can be greatly improved in the double-layer devices, as shown in Figure 4b. A similar structure can also be adopted in the VCM-based memristor, in which the location of resistive switching is confined by completely rupturing the CFs in the active switching layer with lower oxygen ion-migration barrier and maintaining the CFs (i.e., residual CF) in the layer with the higher oxygen ion migration barrier.

In addition, preconditioning the conductive path is also an effective strategy to control ion transport. As shown in Figure 4c, graphene film can serve as the barrier layer to hinder ionic transport and thus prevent redox reactions; the oxygen vacancies are only allowed transport though the preengineered graphene nanoperes, giving rise to a confined transport path; thus, uniform switching performance can be obtained. Similarly, threading dislocations in the resistive materials can also restrict CFs in a defined 1D channel, resulting in minimal spatial and temporal variations.

Concentrating the electric field in a limited region has also been proved to enhance the switching reliability. It is found that constructing metal pyramidal electrode, nanocone electrode, or inserting metal nanodots can concentrate the electric field in a confined region and thus guide the growth of CFs. In addition to constructing different structures in the memristor, the uniform resistive switching parameters can also be achieved by maximizing the uniformity of resistive materials. If the interfaces between the resistive materials and the electrodes are not uniform, there should be many local sites at which CFs can nucleate, thus further increasing the randomness of the film, especially its switching parameters. When a resistive switching device incorporates a uniform material with a smooth surface, a CF path perpendicular to the film surface rather than random or zigzag will form and thus cause the generated CFs to be reasonably similar and uniform (Figure 4e). In addition, other device-level measures such as doping and nondevice-level measures such as current program, voltage program, and pulse operation have also been investigated to improve the device uniformity.

### 3.1.3. Energy Consumption

Typically, the power consumption is estimated to be ≈10 pW per synaptic event, and the energy consumption is about 1 pJ per synaptic event with a duration of ≈100 ms. The energy consumption of the synaptic memristor can be estimated by the following equation of $E_{\text{update}} = V_{\text{spike}} I_{\text{spike}} T_{\text{duration}}$, where $V_{\text{spike}}$ is the amplitude of the spike, $I_{\text{spike}}$ is the duration of the applied spike, and $T_{\text{duration}}$ is the response current. Low operation voltage and current and high switching speed with short pulse duration are desirable for low energy consumption.

Typically, the ECM-based memristors possess a much lower set voltage than VCM-based memristors, due to a much smaller electric field required for switching. The operating voltages lower than 1 V are very common and even below 200 mV have been reported. The energy consumption per spike with the sub-pJ level has been widely reported. Furthermore,
compared with oxides, chalcogenides-based ECM memristors show a lower operation voltage, due to their faster ion transport and looser structure.\[129\] It is found that the ZnS-based ECM memristor (forming voltages: 0.4 – 1.5 V, set voltages: 0.03 – 1.1 V) possesses much lower forming and set voltages than those of the ZnO-based ECM memristor (forming voltages: 2.2 – 6.8 V, set voltages: 0.6 – 4.2 V).\[130 – 132\] The low operating voltage in the ECM cell is desirable for low-power applications. However, the ultralow voltage operation may give rise to a critical issue of big variation that often exceeds 50\%.\[133,134\] For example, a recently reported Ag/TaO\(_x\)/Pt memristor shows \(V_{\text{set}}\) distribution in the range from 40 to 90 mV and \(V_{\text{reset}}\) distribution in the range from –10 to –60 mV.\[134\] Stable operation is therefore difficult to achieve with such a wide voltage range. Nevertheless, considering thermal noise and variation, it is desirable to operate with a voltage above 100 mV. Using oxidized 2D materials as the switching layer, the switching variation can be effectively improved while maintaining a similar low operation voltage.\[131,135\] For example, by introducing a thin oxide layer via annealing in the ZnS-based memristor, a more controllable memristive switching can be obtained. As discussed earlier, this improved variation is likely to originate from the bilayer structure of the ZnS thin film (ZnS/ZnO), where the formation and rupture of the CFs are restricted in the two-layer interface region due to their different ion transport rates.\[131\] However, due to its high operating current, this oxidized ZnS memristor (\(\approx 36 nW\)) shows higher energy consumption than that of biological synapses (\(\approx 10 pW\)). Therefore, reducing the operating current is also very important for low energy consumption. The reduction of operating current is constrained by the thickness of the resistive layer.\[136\] The thickness of the predominant oxide-based resistive layer is usually above 3 nm. The operation current of these devices is usually at the nA level and the lowest record is found to be 100 pA.\[136,137\] 2D materials with the naturally ultrathin body may provide a new strategy to achieve low operating current. The Ag/h-BNO\(_x\)/Gr devices with thickness down to 0.9 nm show reliable switching behaviors and ultralow set/reset current. The operating current can be reduced to the sub-pA level in the devices with an atomically thin active layer. The energy consumption is reduced to a range from 100 aJ to 1 fJ per spike (set voltage 0.6 – 0.7 V, pulse duration 1 ms), which is much lower than that of the biosynapse.\[136\] However, the state-of-the-art energy consumption is demonstrated to be 4.28 aJ/spike in the oxygen vacancy filament-based Pt/HfAlO\(_x\)/TaN.
memristor,\cite{138} which may be attributed to the much shorter pulse duration of 50 ns, although with a higher operating voltage of 5 V and current of 17.2 pA. Therefore, low operating voltage, low operation current, and short pulse duration are all necessary for low power consumption.

3.1.4. Linear and Symmetric Weight Update

The operation of ANNs would be influenced by the asymmetric and nonlinear conductance response. Thus, linear and symmetric weight updating with the applied pulses is preferable (see Figure 2).\cite{139} However, most filamentary-type memristors suffer from the nonlinear and/or asymmetric weight update issues.\cite{140–144} This is associated with the evolution process of the CFs that are basically divided into two stages. Specifically, the formation process is controlled by the drift process in the first stage, i.e., the CFs are forming, whereas when the CFs have formed, the thickening of CFs is controlled by the diffusion process.\cite{145} Such nonlinearity and asymmetry may result in a significant loss of training efficiency.

The nonlinearity issues can be mitigated by the material modulation\cite{140,142,146–149} and pulse programming method.\cite{60,150–152} Figure 5a shows that the linearity of the LTP and LTD process can be improved by the SiO$_2$ diffusive limiting layer with suitable thickness.\cite{146} The SiO$_2$ layer can effectively decrease the number of oxygen vacancies involved in the initial and abrupt CF formation/rupture process, thus leading to higher switching linearity. The increasing number of oxygen vacancies can also enhance the linearity of some memristors. For example, in a KNbO$_3$ (KN)-based memristor, there are two mechanisms that control the CF evolution: a redox process with a fast rate and an oxygen vacancy diffusion process with a slow rate.\cite{140} Through Cu$^{2+}$ ion doping, the redox process can be the predominant mechanism to improve conductance linearity, as shown in Figure 5b.\cite{140} The results in Figure 5a,b indicate that oxygen vacancies could effectively influence the linearity of the memristors. Figure 5c shows that good linearity and symmetry can also be obtained in the Li-ion-based memristor with a highly lithiated phase of Li$_7$Ti$_5$O$_{12}$ through electrically driven metal–insulator phase separation.\cite{147}

![Figure 5](image-url)

Figure 5. Linear and symmetric weight update. Linear improvement of LTP and LTD a–c) by material modulation: a) through SiO$_2$ thickness control in TaO$_x$-based memristor; b) through increasing the number of oxygen vacancies by Cu$^{2+}$ doping in KNbO$_3$ memristor and c) through introducing the highly lithiated phase of Li$_7$Ti$_5$O$_{12}$; d,e) by pulse programming method: d) using nonidentical spikes; e) by applying pulses in memristor electrodes and transistor gate. a) Reproduced with permission.\cite{146} Copyright 2016, The Royal Society of Chemistry. b) Reproduced with permission.\cite{140} Copyright 2020, American Chemical Society. c) Reproduced with permission.\cite{147} Copyright 2020, Wiley-VCH. d) Reproduced with permission.\cite{60} Copyright 2017, American Chemical Society. e) Reproduced with permission.\cite{152} Copyright 2018, Springer Nature.
Besides material modulation, pulse operation can also be used to tune the linearity and symmetry. As shown in Figure 5d, the gradual potentiation (LTP) and depression (LTD) but with nonlinear and asymmetric behavior are obtained with identical pulse responses. On the contrary, the nonidentical pulses can efficiently tune the variation of conductance. Specifically, the near-linear and symmetric potentiation and depression process can be achieved by applying incremental positive pulses (5–6.15 V with 0.05 V step-1 and 200 ns in width) and decreasing negative pulses (−3.1–5 V with −0.05 V step1 and 100 ns in width), respectively. This is because the rate of conductance change usually shows a gradual decline and saturation trend under identical pulses; thus, nonlinearity can be improved to some extent using progressively enhanced pulse sequences. In the 1T1R architecture, the third terminal (i.e., the gate) of the transistors can also offer controllability in tuning the conductance of the memristor to achieve a linear and symmetric conductance change during the training process, as shown in Figure 5e. It should be noted that using a nonideal pulse scheme usually requires a read-before-write step to first identify the conductance state and then apply the correct pulse scheme to the device, which inevitably increases the complexity of the peripheral circuits as well as the latency and energy, whereas using a lot of the transistors (1T1R structure) would increase the concern of the chip area. Therefore, considering the simplification of peripheral circuits and the requirements of area efficiency, it is of priority to address the linear and symmetric concerns from the device level.

3.1.5. Device Stability

The ability of device conductance states to resist the changes of time or/and environment, i.e., device stability, is critical to computing accuracy. Among different environmental factors (such as moisture and strain), thermal tolerance may be the most important issue, as the programming processes in predominant (such as moisture and strain), thermal tolerance may be the most important issue, as the programming processes in predominant filament-type memristors mostly rely on ion movement. Some heat may benefit the formation of CFs, whereas excessive heat would have a detrimental effect. The thermal instability at elevated temperatures especially in repeating set/reset cycles can result in device failure. With the increasing demand for device miniaturization and high-density integration, i.e., a large number of cells will accommodate in a tiny space, the thermal crosstalk issues will be more serious and thus lead to large fluctuations in switching parameters and cause resistance degradation of neighboring devices. Especially in 3D CBA application, thermal crosstalk is one of the most significant problems that should be noticed. The simulation of the 3D CBAs shows that the thermal crosstalk will cause the degradation of retention performance and even lead to the failure of switching from LRS to HRS in the disturbed memristor. In addition, as the feature size continues to shrink, the degradation of the resistance state will become more serious. Thus, thermal stability issues should be addressed especially in high-density CBAs.

Device engineering has been applied to improve thermal stability, and reliable switching behaviors in some memristors can be observed with temperatures ranging from −170 to 340 °C. Recently, Cheng et al. reported environmentally robust memristors based on lead-free double perovskite Cs2AgBiBr6, which remained robust in harsh environments, including humidity as high as 80%, temperatures up to 180 °C, an alcohol burner flame lasting 10 s, and 60Co γ-ray irradiation for a dosage of 5 × 1016 rad (Si). This robustness stems from the high formation energy and good crystallinity of Cs2AgBiBr6. The memristors based on graphene/MoS2–Ox graphene vdW heterostructures were demonstrated to maintain stable resistive switching at a temperature as high as 340 °C, which is the highest operating temperature reported so far. Such robust high-temperature performance is attributed to the superior high-temperature stability of MoS2–Ox as well as graphene along with the atomically sharp interface between the electrodes and the switching layer. In other work, Yang et al. developed BiFeO3-based ferroelectric memristors in an ultrawide temperature range from −170 to 300 °C due to the robust ferroelectricity of BFO at high temperatures. These imply that the thermal stability of resistive materials is key to high thermal tolerance. Except for the improvement of the thermal stability of the individual device, the thermal effect in the array level should also be considered. It was proposed that reducing the reset current and using the cycle-rehabilitate technique could mitigate the thermal crosstalk problem. Operating in the low current mode would result in a less temperature rise at the neighboring unselected cells, thus alleviating the thermal crosstalk issues in CBAs.

3.2. Fabrication and Integration

3.2.1. 3D Array Stacking

Integrating memristors into planar (2D) CBAs is the first step for practical applications. However, 2D CBAs may not be suitable for ANNs especially DNNs, when carrying out complex human activities that require a wider and deeper network. Moreover, considering that the scaling route of 2D CBAs will be constrained by their physical limitations and fabrication cost, 3D integration is becoming a new paradigm for implementing neuromorphic chip for multilayered neural networks. The scalability of the memristors can be fully explored with 3D integration, thus offering high-density applications. Generally, 3D CBAs can be integrated into two types of architectures: 3D horizontally stacked CBAs and 3D vertical CBAs.

**3D Horizontally Stacked CBAs:** In the 3D horizontal stacked CBA architecture, the density of memristor is increased by vertically stacking 2D CBAs. Besides the vertical stacking, the 3D horizontal stacked structure is also flexible in lateral scaling. Moreover, the peripheral circuits can be placed under the CBAs, leading to high-area efficiency and high compact footprint. In addition, a separate selector or transistor can be integrated with each memristor structure; thus, the performance of the selector (or transistor) and memristor can be optimized individually.

**Figure 6a** shows a typical 3D horizontally stacked CBA by simply stacking 2D CBAs on top of each other, separated by an insulating layer. To reduce the interconnect lines and simplify the fabrication process, one-selector-one-transistor (1S1R) 3D horizontally stacked CBA with a shared middle-electrode structure is often adopted, as shown in Figure 6b. The middle electrode can be a bit line (BL) or word line (WL). The shared BL structure is preferable because the number of sense amplifiers
connected to the BLs can be simultaneously reduced with decreased BLs, leading to higher-area efficiency.\textsuperscript{[43]} The fabrication process of the 3D horizontally stacked CBA (Figure 6b) is shown in Figure 6c.\textsuperscript{[160]} It is known that at least three major lithography and etching steps are needed to fabricate a two-layer 1S1R CBA. With the increase in stacking layers, the lithography and etching steps will also increase, thus imposing burdens for further upscaling the CBAs. Similarly, the interconnecting lines will also increase with the number of stacking layers. Moreover, to place these interconnecting lines for different layers, the staircase-shaped bank will be designed for both BLs and WLs. This will lead to different lengths of interconnecting lines in different layers, thus, resulting in the variation of voltage drops in different layers. Therefore, the placement of interconnecting lines that connect the bottom driving circuits is another concern for the 3D horizontally stacked CBA.\textsuperscript{[160]} Therefore, the 3D horizontally stacked CBA structure may encounter critical challenges as the stacked layers are increased.

Despite the earlier mentioned problems, the 3D horizontally stacked CBA is still thought to be a highly feasible approach using simple planar fabrication technology without complicated processes, which allows for integration with CMOS circuits. Many 3D CBAs are reported based on this kind of structure.\textsuperscript{[51,138,162,163]} In 2015, a two-layer 1S1R CBA with a shared BL was first demonstrated by Intel and Micron, in which the 1R was made of PCM with GeSbTe\textsubscript{4}, and 1S was made of GeSe threshold switching memristor. Adam et al. fabricated 3D horizontally stacked CBAs based on Pt/Al\textsubscript{2}O\textsubscript{3}/TiO\textsubscript{2}-x/TiN/Pt memristors for neuromorphic computing, which is composed of two stacked passive 10×10 CBAs.\textsuperscript{[162]} The fabrication process of this 3D CBA has a low, less than 175 °C temperature budget, which could ensure reliable stacking and compatibility with the CMOS process. In the same year, Yoon et al. fabricated a double-layer-stacked 1-diode-1-resistor (1D1R) CBA using the near-room temperature fabrication process.\textsuperscript{[163]} Wang et al. presented a flexible 3D horizontal CBA, exhibiting the multilevel information transmission functionality with an ultralow power consumption of 4.28 aJ per synaptic event.\textsuperscript{[138]}

3D Vertical CBAs: The development of 3D vertical CBAs may be able to overcome the challenges that 3D horizontal CBAs encounter. The typical structure of 3D vertical CBAs or the word-plane-type CBAs is shown in Figure 6d–f.\textsuperscript{[161]} As shown in Figure 6d, the vertical memristors are formed at the cross points of each pillar electrode and plane electrode (serving as the WL), a vertical MOSFET that serves as the BL selector, whose
gate is controlled by the select line. To enable the random access of each memory cell individually, 3D decoding is needed. Figure 6c,f shows the fabrication of a double-layer-stacked 3D vertical CBA with HfO$_x$-based RRAM and the corresponding cross-section transmission electron microscopy (TEM) and high-resolution transmission electron microscopy (HRTEM) images, respectively. In this vertical CBA, only one critical photolithography step is needed. This is the most advantageous feature of 3D vertical CBA that simplifies the lithography steps compared with horizontally stacked CBA, which is beneficial for reducing the fabrication cost.$^{[161]}$

However, there are two key issues to consider in the 3D vertical CBA. First, with the feature sizes below 20 nm, the fabrication of 3D vertical CBAs requires conformal deposition of all functional layers with accurate thickness control onto a very high-aspect-ratio vertical hole. The etching of deep holes within stacking layers, including the insulating layer and metal layer, has become a critical challenge for dry etching and lithography. This is associated with the difficulty of plasma etching gas reaching the bottom of the deep hole with ultrahigh aspect ratio, resulting in different layers showing different chemical properties. In this case, multiple etching and lithography processes may be necessary. After hole etching, the next step is the deposition of resistive layers and metal electrodes within the narrow deep holes. As for the feature size below 20 nm, the thicknesses of resistive layers and metal electrodes are merely a few nanometers. Such stringent requirements in thickness control could only be satisfied by ALD.$^{[43]}$ Besides the challenges in fabrication, another issue is that integrating the individual selectors or transistors with the corresponding memristors in 3D vertical CBA is impossible. Thus, the selector-less or self-selective (1 R) design is necessary. To fulfill this requirement, memristors with the self-selective property have been explored.$^{[159,164–166]}$ For example, in Figure 6f, a TiON interfacial layer formed during the fabrication process is served to obtain nonlinear $I–V$ curves, which can play the selecting role in the 3D vertical CBA. There are a few reports on fabricating less-layered 3D vertical CBAs,$^{[159,161,167]}$ and the implementation of CNN with vertical two-layer Pt/HfO$_2$/TiN random access memory (ReRAM or RRAM) CBA has been reported very recently.$^{[166]}$ However, the 3D vertical CBA technology is still insufficient and further efforts to address these challenges are needed.

### 3.2.2. Sneak Path Issue

Constructing high-density and large-scale memristor CBAs to enable massive parallel operation is significant for neuromorphic computing. However, accessing the selected memristor(s) correctly from such CBAs is challenging due to the sneak path issue that causes failure in operation and high energy consumption. To address this critical issue, MOS transistor,$^{[47–49,152,168–173]}$ selector,$^{[107,174–177]}$ complementary resistive switching (CRS) memory,$^{[178–184]}$ and self-selective devices$^{[23,163,185–187]}$ have been developed to suppress sneak path current across the unselected devices.

To realize large memristor CBAs, the most practical approach at present is to connect the MOS transistors with memristors in series, i.e., the 1T1R structure, as shown in Figure 7a (I). Typically, the CMOS circuitry can be prefabricated with mature CMOS technology, and the memristors, such as TiN/TaO$_x$/HfAlO$_x$/TiN in Figure 7a (II), are stacked and formed on the drain terminal of the transistors.$^{[45]}$ The 1T1R arrays with the transistor’s gate terminal connected to the WL, the source terminal to the source line (SL), and the TE of the memristor connected to BL, are used to build the ANNs. With the assistance of transistors, the sneak current can be effectively suppressed; thus, only the selected devices can be read and programmed. However, the utilization of the conventional Si-based transistor makes this structure possess a large circuit footprint.

The other typical solution is to fabricate a 1S1R structure, as shown in Figure 7b (I). In this structure, the two-terminal selector is connected to the memristor in series. Because the selector can be stacked on top of the memristor, the 1S1R structure possesses a smaller footprint than the 1T1R architecture. Most reported selectors, such as the ovanic threshold switch (TS),$^{[107,175–176]}$ metal–insulator transition (MIT),$^{[188–190]}$ diodes$^{[163,191–193]}$ are mainly based on nonlinear threshold switching or tunneling mechanism. For example, Figure 7b (II) shows a metal filament based-bidirectional threshold switch, i.e., AND-TS, integrates with a nonvolatile TaO$_x$/Ta$_2$O$_{5,3}$ bilayer RRAM (Figure 7b (II)).$^{[107]}$ The AND-TS is a volatile memristor with threshold switching behavior, which shows ultralow leakage current ($\approx$1 pA) and high on-state current (100 μA). The sneak path current can be addressed through the integration of AND-TS selector with the RRAM to form the 1S1R architecture, as shown in Figure 7b (IV). Selectors based on cation migration have shown over 1 $\times$ 10$^9$ nonlinearity but may suffer from slow speed and uncontrollable randomness. This is because the thermal rupture of the tiny metallic filament induced by the interface energy effect would take tens of ms time, and uncontrollable ion transport would result in stochastic filament formation, which is very problematic for its application as a selector, especially in large arrays. Securing strategies to minimize the randomness and improve the switching speed would be an essential step for its practical application.

1T1R and 1S1R arrays still involve complex processes (particularly, current–voltage matching, and etching fabrication problems) that are not compatible with high-density CBA, area scaling, and especially 3D integration.$^{[185]}$ Avoiding the need for a separate transistor or selector is desirable for building high energy efficiency and integration density CBAs. The use of passive array is an option, where no transistor or selector is connected to the memristor. However, it would encounter the sneak path issue, which hinders the reliable operation of large-scale CBAs.

One approach that may be used to alleviate these issues is CRS memristor. In this structure, two identical memristors are connected back-to-back sharing the middle electrode, as shown in Figure 7c (I). Linn et al. pioneered CRS memristor with two Pt/SiO$_2$/GeSe/Cu memristors connected antisocially by the middle electrode.$^{[180]}$ The typical $I–V$ curve of the CRS memristor is schematically shown in Figure 7c (II). There exist bistable HRS + LRS (binary number “0”) and LRS + HRS (binary number “0”) combined states in the CRS device, which is different from typical memristors that only have LRS and HRS states, as shown in Figure 7c (III). However, the selectivity of the CRS memristor is limited due to the high off current, which is impractical for most ANN applications.
The last important solutions are those self-selective (or self-rectifying) memristors (Figure 7d (I)), which are attractive as they have the same simple MIM trilayer structure and do not need an extra selector connected to the memristor in large CBAs. Such kind of a device holds great promise for high-density CBA and is especially important for 3D vertical CBA. The self-selective behavior generally originates from the existence of nonlinearity. Figure 7d (II–VII) shows a p-Si/SiO2/n-Si memristor with a built-in rectifying selector, which originates from the self-assembled diode within each junction formed by the different doped silicon electrodes. From the I–V curves, it is known that the forward current under positive voltage exhibits a normal loop shape, whereas the reverse current is suppressed without regard of its state; (IV) schematic of the two-layer-stacked memristor CBAs where the two layers of devices are electrically isolated by spin-on glass, the red line is the expected current path, whereas the blue line shows one typical intralayer sneak path being blocked by a reverse-biased cell, and (V) the corresponding the color map of the readout current, the bits are read out correctly which proves the effective blocking of the intralayer sneak path current by the built-in diode; (VI) schematic of the two-layer stacking with shared electrodes with adjacent layers and (VII) the corresponding experimental measurement result in a 2×2 subarray, which shows that, in the worst-case scenario, the only HRS cell in the first layer can be read out correctly although all other cells are in LRS. This result confirms the successful suppression of the interlayer sneak path current in the array. a) Reproduced with permission.© 2017 Springer Nature. b) Reproduced with permission.© 2019, Wiley-VCH. d) Reproduced with permission.© 2017, Springer Nature.

The last important solutions are those self-selective (or self-rectifying) memristors (Figure 7d (I)), which are attractive as they have the same simple MIM trilayer structure and do not need an extra selector connected to the memristor in large CBAs. Such kind of a device holds great promise for high-density CBA and is especially important for 3D vertical CBA. The self-selective behavior generally originates from the existence of nonlinearity. Figure 7d (II–VII) shows a p-Si/SiO2/n-Si memristor with a built-in rectifying selector, which originates from the self-assembled diode within each junction formed by the different doped silicon electrodes. From the I–V curves, it is known that the forward current under positive voltage exhibits a normal loop shape, whereas the reverse current is suppressed without regard of its state; (IV) schematic of the two-layer-stacked memristor CBAs where the two layers of devices are electrically isolated by spin-on glass, the red line is the expected current path, whereas the blue line shows one typical intralayer sneak path being blocked by a reverse-biased cell, and (V) the corresponding the color map of the readout current, the bits are read out correctly which proves the effective blocking of the intralayer sneak path current by the built-in diode; (VI) schematic of the two-layer stacking with shared electrodes with adjacent layers and (VII) the corresponding experimental measurement result in a 2×2 subarray, which shows that, in the worst-case scenario, the only HRS cell in the first layer can be read out correctly although all other cells are in LRS. This result confirms the successful suppression of the interlayer sneak path current in the array. a) Reproduced with permission.© 2017 Springer Nature. b) Reproduced with permission.© 2019, Wiley-VCH. d) Reproduced with permission.© 2017, Springer Nature.

3.2.3. Wire Resistance

When we establish CBAs with a higher integration density and smaller feature size (such as <10 nm), except for the sneak current issue, the wire resistance will also impose challenges. The undesirable wire resistance will distort signal propagation, leading to high power consumption and limited computing precision. Wire resistance will reduce the write/read voltage to the selected memristor(s), depending on its distance from the voltage driver. In small CBAs, the wire resistance is negligible. But as the size of the CBAs increases, the wire resistance issues will become more evident, because the high wire resistance will
prevent the selected memristor from receiving enough voltage, leading to failed operation. Therefore, the maximum CBA size for reliable operation, as well as integration density, is limited by the wire resistance.\cite{201} The impact of wire resistance on the write operation has been studied. It is found that all the three write schemes of floating line, $V/2$, and $V/3$ are reliable for the square array without wire resistance. However, CBAs with high density cannot operate properly by any of these three schemes due to wire resistance.\cite{202} Furthermore, a comprehensive writing margin analysis based on the stacked 1D1R device was reported by Hwang’s group.\cite{202} It is revealed that an excessive voltage drop on the wire resistances of selected WLs and BLs would lead to an increase in voltage for the write operation (programming/erasing), whose degree is associated with the selectivity of the selector, array size, resistance of the selected cell, and wire resistance. Especially, with the increasing ratio of wire resistance/device resistance, the ratio of source voltage/write voltage will increase, which would aggravate when the array size is larger than 1 Mb. This would in turn influence the allowable array size in ANNs. Therefore, at the array and circuit levels, addressing the challenges of wire resistance is necessary for large-scale implementations of memristor CBAs-based neuromorphic computing.

Several solutions may mitigate this issue. First, increasing the metal wire thickness allows high applicability with decreased wire resistance. Second, a possible effective method to mitigate this issue is to adopt a memristor whose resistance is far larger than the wire resistance. Chen et al. reported that the LRS in HfO$_2$-based memristor could be up to 1 M$\Omega$, whereas the HRS could be 10 M$\Omega$.\cite{203} However, the access speed of the memristors would be decreased due to reduced read current as induced by the high resistance. Building 3D networks that can significantly reduce the length and number of wires may be a more effective way to address this issue. As for the same 1 Mb memory, the series resistance of a 100 $\times$ 100 $\times$ 100 3D CBA is about seven times lower than that of a 2D CBA.\cite{204} In addition, optimizing the programming scheme and peripheral circuits may also mitigate the influence of wire resistance.\cite{205,206,207}

4. Hardware Implementation of NCS

With decades of development, remarkable progress in artificial neuromorphic devices such as artificial synapse and neuron has been achieved, which are fundamental to the implementation of ANNs. Most demonstrations of neuromorphic computing, such as pattern recognition, are based on array-level simulation. Integrating memristive neuromorphic devices to construct the ANNs as well as assemble other necessary circuits to implement the hardware-level NCS is critical. A typical hardware NCS primarily consists of memristor synaptic CBAs, neurons, and peripheral circuits. The memristor synaptic CBAs are used to perform vector matrix multiplications for the calculation of neuronal outputs. There are mainly three kinds of synaptic CBAs, including 1T1R arrays, 1S1R arrays, and 1 R arrays. The neurons, i.e., the information processing unit, are usually based on CMOS circuits conducted by operational amplifiers, and the memristor-based neurons are also developed. The peripheral circuits based on transistors include multiplexers to perform switch matrix, digital/analog converters (DACs) or analog/digital converters (ADCs) to convert the data moving in or out of the CBAs, shift & adder, input/output (I/O) register, etc. These components are assembled in single or several printed circuit boards (PCBs) to implement neuromorphic computing. However, experimental implementation of large memristive ANN hardware is still in its nascent stage. The recent hardware implementation of ANNs may give us a preliminary prototype for the development of hardware memristive neuromorphic systems.

4.1. Hardware Implementation with 1T1R Arrays

Due to the maturity of CMOS manufacturing technology, integrating memristor CBAs with MOS transistors, i.e., 1T1R structure, is the most popular approach to realize large memristive ANNs at present. The sneak current can be effectively suppressed by the series-connected transistor. Meanwhile, conductance tuning can be more controllable through the gate control of the transistor, which enables the right reading and programming of the memristor. Although such a structure possesses a larger circuit footprint due to the use of transistors, the manufacturability of large 1T1R arrays using standard CMOS technology allows high yield, thus, hardware implementation of neuromorphic computing with 1T1R arrays are most widely reported. IBM released the first large-scale neural network (165,000 synapses) based on 1T1R technology, using PCMs as the synaptic weight elements and software-based neurons.\cite{208} In the same year, IBM also demonstrated a neuromorphic core with 64k-cell PCM synaptic array (256 axons by 256 dendrites) with in situ learning capability.\cite{209} Since then, many hardware implementations of neuromorphic computing are reported using 1T1R structures to construct different ANNs, such as SLP/MLP,\cite{210,211,212} Deep-Q,\cite{213} and CNN.\cite{214,215}

Figure 8 shows a typical hardware implementation of multiple-layer neural network, i.e., MLP, with 1T1R arrays.\cite{216} As shown in Figure 8a–c, the Ta/HfO$_2$/Pt memristors and the commercial foundry-fabricated transistor arrays are monolithically integrated on a 6 in. wafer. Each Ta/HfO$_2$/Pt memristor is connected to a transistor in series, forming the 1T1R structure. The 1T1R arrays are used to build the multiple-layer neural network, as shown in Figure 8f–g. In this network, each synaptic weight is represented by the difference in conductance between the two memristors, and the weighted sums of the input voltages are calculated at each cross point. Between the 1T1R arrays, there are circuits that perform the functions of reading the current in each cross point, converting the current to a voltage and performing the activation function. Here, the function of nonlinear activation, i.e., the function of the neuron, is conducted in software. The memristive networks are further integrated with peripheral circuits, such as ADC and DAC, to perform neuromorphic computing (Figure 8h). Specifically, the 128 $\times$ 64 arrays are used to build a two-layer perceptron network, including 64 input neurons, 54 hidden neurons, and 10 output neurons. The network is trained with the Modified National Institute of Standards and Technology (MNIST) dataset using stochastic gradient descent, as shown in Figure 8i–q. For each new training
data sample, the network first uses the SoftMax function to conduct inference to obtain the log probability of each output label and then updates the weight of each layer accordingly. After training with the entire dataset, 91.71% of the 10,000 images in the test set can be classified (Figure 8i–n), but many of the misclassified images are also found (Figure 8o–q). By further enlarging the scale of 1T1R arrays (e.g., 1024 × 512), the accuracy could be higher than 97% based on simulation results.

Very recently, Yao et al. reported the implementation of CNN with eight integrated 2048 cell 1T1R arrays.[48] The multiple memristor arrays-based CNN hardware system is built on a customized PCB and a field-programmable gate array evaluation board. Specifically, the system is mainly composed of eight memristor-based processing elements (PEs), in which every PE includes integrated 2048 cell (128 × 16) 1T1R arrays. In these 1T1R arrays, the transistors and main interconnections are pre-fabricated in a CMOS foundry with 130 nm technology node.

Figure 8. Hardware-implemented MLP with 1T1R arrays. a–e) Integrating memristor CBAs with transistors from wafer scale to nanometer scale: a) optical image of a wafer with transistor arrays. b) Close-up of chip image showing transistor arrays of various sizes. c) Microscope image showing the 1T1R structure in CBAs, scale bar, 10 μm. d) Cross-sectional SEM image of an individual 1T1R cell, scale bar, 2 μm. e) Cross-sectional TEM image of the integrated Ta/HfO$_2$/Pt memristor, scale bar, 2 nm. f–h) memristive NCS: f) schematic diagram of a two-layer neural network; g) the implementation of the network with a set of memristor arrays; h) circuit diagram of the measurement system. i–q) In situ online training and inference experiments on MNIST handwritten digit recognition: i) photo of the integrated 128 × 64 arrays during measurement; j) minibatch accuracy increases over the course of training; k) the conductance gate voltage relation extracted from data collected during training; l) the conductance gate voltage relation extracted from data collected during training; l–n) typical correctly classified digit “9”; o–q) misclassified digit “8” after in situ training. Reproduced with permission.[152] Copyright 2018, Springer Nature.
Then the TiN/TaOx/HfOx/TiN memristors are stacked on the drain electrode of the transistors and fabricated in the lab. With those memristor CBAs, a five-layer memristor-based CNN is built to perform MNIST10 image recognition. The five layers are convolutional layer 1 (C1), pooling layer 2 (S2), convolutional layer 3 (C3), pooling layer 4 (S4), and fully connected layer (FC), which are constructed by the memristors. For example, 9 out of 16 memristors in a row are used to realize a 3 × 3 kernel, and the residual cells remain unused. Thus, the 1 × 3 × 3 × 8 (depth × weight × height × batch) kernel weights of the C1 layer required 16 differential rows of memristors (PE1), and the 8 × 3 × 3 × 12 kernel weights of the C3 layer required 192 differential rows of memristors (PE1 and PE3). Meanwhile, a hybrid training method is developed to overcome the nonideal device characteristics. It mainly includes four steps: first, the CNN model is trained ex situ; then the determined weights are transferred to the memristor PE; next, the external input propagates forward through the CNN; finally, the last FC layer is trained in situ afterward to tune the memristor conductance. A high pattern recognition accuracy of more than 96% can be achieved by this memristor-based CNN hardware system with hybrid ex situ and in situ training method.

In most ANN hardware implementation, the memristor chip is integrated with periphery components, such as ADC and DAC on a/ several PCBs. By integrating all the necessary functions on chip, the hardware implementation of memristor-based NCs will be more efficiently realized and the prototypes are allowed to be extended to larger systems.[47] Lu and coworkers have developed a fully integrated, functional, reprogrammable memristor chip, in which the memristor arrays, all the necessary circuitry, digital buses, and an OpenRISC processor are integrated on a chip to form fully on-chip system hardware.[47] The on-chip system is highly flexible and can be programmed to implement different computing models and network structures including the perceptron network, the sparse coding algorithm, and the bilayer principal component analysis system.

4.2. Hardware Implementation with Passive Memristor CBAs

Mapping the ANNs onto large 1S1R or 1R memory array is considered the most practical approach for high-density integration (especially 3D integration). Using two-terminal selectors (e.g., TS selectors) or self-selective devices in a large memristor CBA can minimize the influence of crosstalk current from surrounding memristors, thereby enabling the selected memristor(s) to be accurately accessed.

The reports of neuromorphic functionality based on 1S1R structure have been so far very limited. This may be attributed to the much stringent requirements (such as variation and endurance) on a selector, even more than a memristor. Part of the reason is that the selectors are operated not only during memristor programming but also during more frequent read operations. Moreover, to obtain sufficient I–V nonlinearity in the 1S1R devices, the operating parameters of the selector, e.g., threshold voltage and operating current, should match those of the memristor; thus, a carefully collaborative design between the memristor and the selector is needed. This increases the difficulty of the practical application of the 1S1R structure. Compared with 1S1R structure, CBAs based on self-selective devices, i.e., passive CBAs, are more popular for implementing neuromorphic computing.[46,209,210] A standard hardware neuromorphic system with the 1R structure is shown in Figure 9, which comprises two passive 20 × 20 memristive CBAs with Pt/Al2O3/TiO2/Ti/Pt memristor at each intersection, CMOS neurons, and some other discrete conventional components (Figure 9a).[46] The nonlinearity of the Pt/Al2O3/TiO2/Ti/Pt memristor offers the function of a selector to limit sneak currents in the CBAs and hence reduces interference of half-selected devices during weight tuning. Two passive 20 × 20 CBAs are packaged and integrated with other necessary CMOS components on two separate PCBs to implement the MLP network. The MLP featuring 16 input neurons, 10 hidden-layer neurons, and 4 output neurons is used to conduct black-and-white pattern classification, i.e., “A,” “T,” “V,” and “X,” as shown in Figure 9b–d. When performing pattern classification, the MLP network is first trained offline to obtain the weights through software-implemented network, then the weights are written to the hardware CBAs through the driving circuit, and finally, the inference process is performed in the memristor hardware NCs to recognize the input black-and-white patterns. Here, the typical offline training method is used. Based on this NCS, there are two approaches to perform pattern classification. The first relies on a hardware-oblivious approach in which all the memristors in CBAs are supposed to work perfectly. Using this approach, the classification results for training and test patterns can be 95% and 79.06%, respectively (Figure 9e–f). The second makes use of a hardware-aware approach in which nonideal memristors are considered in the training process. Based on such an approach, the experimental results are much better than the first approach, in which classification accuracy of 100% and 81.4% for the training and test patterns can be achieved, respectively (Figure 9g–h). The demonstrated network has a relatively low complexity, which is beneficial for practical applications. In addition, it contains most typical features, such as multilayer network structure, nonlinear circuit, memristive synapse, and silicon neuron, which are needed to construct a practical large-scale DNN. Nevertheless, without the help of extra transistor or selector, passive CBAs with self-selective properties would be required to deliver more demanding device performance requirements (such as nonlinearity and linear and symmetric weight update) for constructing large-scale ANNs.

4.3. Hardware Implementation with Memristive Neuron

In most ANNs, the neuron functions are either implemented by CMOS circuits composed of 10 or more transistors or in software running on the processors,[46,152,209] which limit its further improvements on scalability, stackability, and energy efficiency. In recent years, memristor-based artificial neurons have been developed. With small feature size, the memristive neurons circuit can be further scaled down, hence improving the area efficiency of future AI chips.

As shown in Figure 10, a memristor neuron is used to construct the CNN for pattern recognition.[46] The artificial spike-based neuron based on HfO2 memristor contains dendrites (inputs), soma, and axon (output), as shown in Figure 10a.
When continuous pulses are applied to the memristive neuron, the electrical potential of the memristor will increase step by step until it exceeds the threshold and results in the firing of a spike, thus realizing the integration function of a neuron. Using the memristive neuron, a hybrid CNN is constructed to demonstrate digit recognition, as shown in Figure 10b–c. At the hardware level, the hybrid CNN system based on memristive neurons is assembled on a PCB, which comprises memristive neurons, transmission gate chips, operational amplifiers, and comparator chips. At the network structure level, there are six layers in the CNN, including two convolutional layers, two maxpool layers, and two fully connected layers (Figure 10b). When performing the inference process, one memristive neuron is used to perform the functions of 784 neurons based on the time division multiplexing access (TDMA) technique in the first layer of the perceptron. For clarity, only one hidden-layer neuron is shown. A complete set of training patterns for the four-class experiment, stylistically representing letters "A," "T," "V," and "X." Perceptron output voltage for e–f) hardware-oblivious and g–h) hardware-aware ex situ training approaches. Reproduced with permission.[46] Copyright 2018, Springer Nature.

When continuous pulses are applied to the memristive neuron, the electrical potential of the memristor will increase step by step until it exceeds the threshold and results in the firing of a spike, thus realizing the integration function of a neuron. Using the memristive neuron, a hybrid CNN is constructed to demonstrate digit recognition, as shown in Figure 10b–c. At the hardware level, the hybrid CNN system based on memristive neurons is assembled on a PCB, which comprises memristive neurons, transmission gate chips, operational amplifiers, and comparator chips. At the network structure level, there are six layers in the CNN, including two convolutional layers, two maxpool layers, and two fully connected layers (Figure 10b). When performing the inference process, one memristive neuron is used to perform the functions of 784 neurons based on the time division multiplexing access (TDMA) technique in the first convolutional layer, whereas neurons in other layers are implemented in software, as shown in Figure 10c. A total 10 000 different images are used to evaluate the performance of this memristive neuron-based CNN and an accuracy of about 97.1% can be achieved, which is close to that of a pure software-implemented network running on GPU. In this CNN, one memristive neuron is used to behave as 784 neurons, which may reduce the number of neurons in ANNs. In contrast, the neurons in other convolutional layers are based on software and the synapses are nonmemristors. The implementation of full memristive hardware ANNs is still limited. Moreover, it is unlike the transistor-type neurons (the so-called neurotransistor) in which multiple gates can be used as extra-terminals to realize various neuronal functions,[88] a single memristor is yet to realize versatile neuron functions at the moment, which hinders the application of full memristive hardware ANNs.

As mentioned earlier, most ANNs are implemented with memristive synapses but nonmemristor neurons or with memristor-based neurons but nonmemristor synapses, which may not fully utilize the scalability of the memristor. Combining memristive synapse with memristor-based neuron to realize full memristive ANNs makes it possible to build more compact and higher-density ANNs. Further progress in integrating diffusive memristive neurons with nonvolatile memristive synapses to build a fully memristive ANN was achieved by Wang et al. The full memristive ANN contains an 8 × 8 1T1R memristive synaptic array that is connected to eight diffusive memristor neurons. The artificial synaptic arrays are constructed by integrating Pd/HfOx/Ta drift memristors with foundry-made transistor arrays using back-end-of-the-line processes and the Pt/Ag/SiOx/Ag/Pt diffusive memristor in parallel with a capacitor is used as LIF neuron. Especially, stochastic LIF dynamics and tunable integration time can be achieved by the diffusive memristor-based artificial neuron. The interaction between the artificial neuron and synapse is built by connecting the drift memristor synapse and the artificial neuron in series. The spiking process in this artificial neuron is a typical threshold process. Specifically, when the memristive synapse is in a low conductance state, the artificial neuron will integrate the inputs but
not fire a spike, as the integrated potential cannot reach the threshold value within the duration of this pulse. When the memristive synapse is in a high conductance state, the integrated potential across the memristor neuron will be larger than the threshold, which will trigger the firing process. In this way, the connection between the memristive synapse and memristor-based neuron will be established. Furthermore, with this integrated network, unsupervised synaptic weight updating and pattern classification are carried out experimentally. This demonstrated full memristive neural network gives a typical example of the integration of memristive synapses and neurons. It would provide an important step forward as it opens up the possibility of implementing ANNs with full memristive devices. With advances in the low-temperature CMOS process, the fabrication of both synaptic and neuronal memristive layers on low-cost, flexible plastic substrates may be possible.

4.4. Hardware Implementation with 3D Memristor CBAs

Existing memristor-based ANNs are mostly built on conventional 2D CBAs due to the mature fabrication process and simple structure. However, their scaling limitation, especially the full connected topology and simplified connections, cannot efficiently implement the complete topology of more complex ANNs. Extending the 2D CBAs into the 3D design, a large number of functional connections can be realized to construct more complex memristor ANNs. Here we take the implementation of CNN as an example to understand the advantage of 3D CBAs over 2D.
For a CNN based on 2D CBAs, each convolutional layer performs 2D convolution operations on the image data by applying spatially copied kernels to the local input window and conducting dot-product computing. A traditional 2D CBA is sufficient to reorganize and perform a single kernel operation. However, due to the fully connected topology, it is challenging to place pixel-level parallel convolutions in the same array. 2D array blocks are required to process all kernel operations in parallel, which is very area expensive for large input. In contrast, the 3D CBAs with multilayers make it possible to copy the replicated kernels into a single array and provide a high area-efficient I/O rather than the conventional perimeter I/O in 2D CBAs, thereby offering a highly compact and efficient implementation of CNNs. With the 3D CBA design, resource-intensive tasks are likely to be transformed into a manageable scale, which provides a substantial increase in speed and energy efficiency when running complex ANNs. Therefore, the development of 3D CBAs is desirable in complex ANNs where a large number of connections and efficient communication are needed.

Chakrabarti et al. first demonstrated a hybrid 3D CMOS/memristor circuit in a CMOL (CMOS + “molecular”) architecture, which combines ultra-high-density 3D memristive CBA with CMOS technology (Figure 11).[211] The 3D CMOL hybrid circuit is proposed as a multiply–add engine to conduct dot-product operations. Figure 11a–d shows the concept of the multiply–add engine using an eight-layer 3D CBAs. In each layer, eight rows of BEs (the thinner red lines) are connected by eight memristors to the TEs (the thinner blue lines), and the TE columns are connected by eight vertically stacked devices. The whole multiply–add operation can be performed in two dimensionalities, i.e., 2D and 3D. In one layer, an array of voltage pulses is applied to the eight BE lines, and weighted input signals (i.e., the conductance of each memristor) from eight BE lines are received by each TE line. Subsequently, the output signals generated by each TE line in each layer are then summed in the TE columns of all layers. The hardware implementation of the 3D CMOL multiply–add engine is performed within the 3D CMOL hybrid circuit, in which the two-layer memristor CBA is monolithically integrated on the prefabricated CMOS substrate.
circuit,[211] as shown in Figure 11e–g. Using the 3D memristive CMOL hybrid circuit, the multiply–add operation in the first and second layers of 3D CBAs can be realized. This work provides an example of performing basic matrix multiplication in ANN using 3D CBAs.

Inspired by the 3D CMOL architecture, Lin et al. demonstrated the first eight-layer 3D memristor circuit to construct ANNs for handwritten digits and edge detection of moving objects in videos very recently.[51] In these 3D CBAs, the input pillar electrodes and output staircase electrodes are arranged in a nonorthogonal alignment, forming high-density but localized connections. The 3D CBAs are monolithically integrated on top of the silicon wafer with I/O contact vias. The 3D CBAs are programmed into parallelly operated kernels to implement a CNN, wherein MNIST handwritten pattern recognition with software-comparable accuracy can be achieved. Meanwhile, the Prewitt filter groups in 3D CBAs are used to process pixels in parallel to perform edge detection of moving objects in video. In this parallel video processing process, only the parallel kernel operation is done in hardware, whereas the remaining network components, including the activation function, pooling layer, and fully connected layer, are implemented in software. Fine edge features of moving objects are extracted successfully by either software-based CNN or hardware-implemented CNN using 3D CBA. This work offers a good example of the implementation of ANNs with 3D CBAs. However, the memristors in these 3D memristor CBAs may fall short of nonlinearity or selectivity and exhibit a certain degree of device variation, which should be further improved for performance optimization. Very recently, Hwang and coworkers demonstrated the implementation of hardware CNN based on a two-layer vertical-stacked CBA (S-CBA) using Pt/HfO$_2$/TiN self-rectifying ReRAM.[166] The two-layer CBA could be configured as having one WL in contact with several layers in a single BL to perform either the single-input multiple-output (SIMO) scheme or multiple-input single-output (MISO) scheme. The SIMO scheme is more suitable to extract the features of a letter with multiple selections of the intended features, whereas the MISO scheme is preferable for the extraction of features of a color image including several component color images. Furthermore, although the Pt/HfO$_{2.0}$/TiN shows pronounced device variation, the suggested adaptive scheme makes the system immune to the variation. The 3D CBAs are deemed to have great potential to build high-density ANNs to conduct complex neuromorphic activities. However, the number of layers in 3D CBAs is limited to eight (mostly the two-layer structure) in the present application of hardware NCs. Extensive effects are needed to fully utilize the merits of 3D CBAs.

5. Summary and Outlook

Neuromorphic computing is remarkably different from conventional computing, which can remove the memory wall of today’s computing systems. Memristors with nonvolatility, high speed, small footprint, scalability, and stackability are promising for neuromorphic computing. Many researchers have demonstrated that memristors are efficient in mimicking multiple synaptic and neuronal functions. To implement massively parallel and energy-efficient NCSs, highly miniaturized memristors are required to be integrated into large-scale CBAs. This gives rise to challenges on both device performances and large-scale CBAs fabrication and integration. At the device level, a large number of resistance states, wide resistance range, stable resistance state, and high endurance for repeated programming/training are the basic requirements for the memristors. Low energy consumption and good scalability are required for large-scale high-density neural networks. Good device stability is desirable for high computing accuracy. Especially, for SNNs, the ability of memristors to implement STDP rule is necessary. For DNNs, analog resistive switching, long retention, small device variation, and linear/symmetric weight updating are in demand. At the array level, 3D stacking is necessary for high-density ANNs to conduct complex neuromorphic activities, but sneak current and finite wire resistance issues should be circumvented in the large-scale CBAs. Possible solutions have been proposed and developed to address these challenges. With the efforts in overcoming these challenges, some progress on the hardware implementation of NCSs has been made, indicating the opportunities of memristive ANNs for practical applications. Despite recent progress, more efforts in other aspects are necessary to overcome for real-life implementation of large-scale NCs.

First, although the memristive devices are promising in neuromorphic computing, it has been challenging to improve the performances of predominant oxide-based memristors in terms of scalability, energy consumption, and flexibility/transparency. To overcome these challenges, the exploitation of new materials is required. One promising approach lies with the use of 2D materials, which exhibit various unique physical properties, such as ultrathin body, flexibility, and transparency. Some progress in the deployment of 2D materials-based memristors in realizing synaptic devices has been achieved.[7–9,213] Meanwhile, organic materials-based memristors also have the potential to offer a functionally promising and cost-effective platform for flexible and wearable neuromorphic computing hardware.[211,214] Furthermore, 0D nanomaterials, such as quantum dots, with superior optical properties are well suited to realize photonic devices to implement neuromorphic computing in photonic systems.[215]

Second, there are also challenges at the peripheral circuit level for neuromorphic computing that remain to be solved. In the memristive NCSs, the read/write process is performed by peripheral circuits, which are usually transistor based, leading to a large area and high energy consumption. Besides the earlier mentioned finite wire resistance, the requirement of DACs or ADCs to transform the data when moving in or out of the CBAs is also one of the critical issues. It typically requires multiplex ADCs across multiple columns, resulting in increased latency.[209] One possible solution is using fully analog peripheral circuits to avoid such conversions at the cost of less flexibility and accuracy.[216] In addition, the peripheral circuits of current memristive ANNs are mostly CMOS based; thus, hybrid integration of memristive CBAs with CMOS circuits is of paramount importance.

The third is algorithm optimization, which involves three aspects of training and testing, overcoming nonideal device properties, and learning algorithms. Training and testing are the routine processes for ANNs, especially DNNs. In these
processes, frequent read/write operations to obtain/tune the parameters of memristive synaptic weights are needed, which require memristors with high speed and endurable programming. As computing accuracy is highly dependent on the complexity of DNNs, increasing the network layers would improve the testing accuracy. In other words, constructing larger-scale systems can enhance the ability of ANNs in solving more complex tasks. Furthermore, some device imperfections can be compensated by algorithms optimization. For example, low precision algorithms and systems, such as fuzzy algorithms, QNNs, and BNNs, can be used to overcome the device variations.[16,217] Instead of deterministic weight states in the common ANNs, the fuzzy learning method allows us to define memristive states dynamically, hence enhancing the adaptation and tolerance in device variations. Another approach to alleviate the influence of device variation is the quantization or binarization of neural network weights, i.e., QNNs or BNNs. With these methods, the ANN topology as well as the learning algorithms of memristive ANNs are still under development. The ANN topology as well as the learning algorithm, such as data mapping, dot product, and STDP, is still requiring optimization for generic applications.

Acknowledgements

This work is supported by A*STAR Science and Engineering Research Council (No. 152-70-00013).

Conflict of Interest

The authors declare no conflict of interest.

Keywords

hardware implementations, large-scale crossbar arrays, memristors, neuromorphic computing

Received: June 18, 2020
Revised: August 27, 2020
Published online: November 4, 2020

[1] M. Horowitz, presented at 2014 IEEE Int. Solid-State Circuits Conf. (ISSCC), San Francisco, CA, February 2014.
[2] Q. Xia, J. J. Yang, Nat. Mater. 2019, 18, 309.
[3] D. Ielmini, H. S. P. Wong, Nat. Electron. 2018, 1, 333.
[4] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esler, R. Appuswamy, B. Tabo, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, D. S. Modha, Science 2014, 345, 668.
[5] M. Mayberry, Intel’s New Self-Learning Chip Promises to Accelerate Artificial Intelligence, Intel, Santa Clara, CA 2017.
[6] N. Imam, T. A. Cleland, Nat. Mach. Intell. 2020, 2, 181.
[7] X. Feng, X. Liu, K.-W. Ang, Nanophotonics 2020, 9, 20190543.
[8] L. Chen, L. Wang, Y. Peng, X. Feng, S. Sarkar, S. Li, B. Li, L. Liu, K. Han, X. Gong, J. Chen, Y. Liu, G. Han, K.-W. Ang, Adv. Electron. Mater. 2020, 6, 2000057.
[9] X. Feng, Y. Li, L. Wang, S. Chen, Z. G. Yu, W. C. Tan, N. Macadam, G. Hu, L. Huang, L. Chen, X. Gong, D. Chi, T. Hasan, A. V. Thean, Y. Zhang, K.-W. Ang, Adv. Electron. Mater. 2019, 5, 1900740.
[10] X. Feng, Y. Li, L. Wang, Z. G. Yu, S. Chen, W. C. Tan, N. Macadam, G. Hu, X. Gong, T. Hasan, Y.-W. Zhang, A. V.-Y. Thean, K.-W. Ang, presented at 2019 Int. Symp. on VLSI Technology, Kyoto, Japan, June 2019.
[11] L. Wang, W. Liao, S. L. Wong, Z. G. Yu, S. Li, Y.-F. Lim, X. Feng, W. C. Tan, X. Huang, L. Chen, L. Liu, J. Chen, X. Gong, C. Zhu, X. Liu, Y.-W. Zhang, D. Chi, K.-W. Ang, Adv. Funct. Mater. 2019, 29, 1901106.
[12] S. Dai, Y. Zhao, Y. Wang, J. Zhang, L. Fang, S. Jin, Y. Shao, J. Huang, Adv. Funct. Mater. 2019, 29, 1903700.
[13] H. Ling, D. A. Koutsouras, S. Kazemzadeh, Y. van de Burgt, F. Yan, P. Gkoupidenis, Appl. Phys. Rev. 2020, 7, 011307.
[14] N. K. Upadhyay, H. Jiang, Z. R. Wang, S. Asapu, Q. F. Xia, J. J. Yang, Adv. Mater. Technol. 2019, 4, 1800589.
[15] M. A. Zidan, J. P. Strachan, W. D. Lu, Nat. Electron. 2018, 1, 22.
[16] T. Zhang, M. H. Yin, X. Y. Lu, Y. M. Cai, Y. C. Yang, R. Huang, Nano Futures 2017, 1, 8.
[17] B. K. You, J. M. Kim, D. J. Joe, K. Yang, Y. Shin, Y. S. Jung, K. J. Lee, ACS Nano 2016, 10, 9478.
[18] Y. Ji, Y. Yang, S.-K. Lee, G. Ruan, T.-W. Kim, H. Fei, S.-H. Lee, O.-Y. Kim, J. Yoon, J. M. Tour, ACS Nano 2016, 10, 7598.
[19] X. H. Wu, R. J. Ge, P. A. Chen, H. Chou, Z. P. Zhang, Y. F. Zhang, S. Banerjee, M. H. Chiang, J. C. Lee, D. Akinwande, Adv. Mater. 2019, 31, 1806790.
[20] S. Roy, G. Niu, Q. Wang, Y. Wang, Y. Zhang, H. Wu, S. Zhai, P. Shi, S. Song, Z. Song, Z.-G. Ye, C. Wenger, T. Schroeder, Y.-H. Xie, X. Meng, W. Luo, W. Ren, ACS Appl. Mater. Interfaces 2020, 12, 10648.
[21] S. Ren, Z. Li, L. Tang, X. Su, H. Zhang, G. Zhang, H. Zhang, G. Cao, J. Zhao, Adv. Electron. Mater. 2020, 6, 2000151.
[22] Y. Chen, Y. Yan, J. Wu, C. Wang, J. Y. Lin, J. S. Zhao, C. S. Hwang, ACS Appl. Mater. Interfaces 2020, 12, 10681.
[23] H. Zhang, X. Ju, Y. Zhou, C. Gu, J. Pan, D. S. Ang, ACS Appl. Mater. Interfaces 2019, 11, 41490.
[24] K. Ding, J. Wang, Y. Zhou, H. Tian, L. Lu, R. Mazzarello, C. Jia, W. Zhang, F. Rao, E. Ma, Science 2019, 366, 210.
[25] F. Zhang, H. R. Zhang, S. Krylyuk, C. A. Milligan, Y. Q. Zhu, D. Y. Zemlyanov, L. A. Bendersky, B. P. Burton, A. V. Dasydov, J. Appenzeller, Nat. Mater. 2019, 18, 35.
[26] K. J. Zhu, D. Li, X. G. Liang, W. D. Lu, Nat. Mater. 2019, 18, 141.
[27] Q. X. Yin, L. Chen, Nanotechnology 2020, 31, 9.
[28] S. Long, L. Perniola, C. Cagli, J. Buckley, X. Lian, E. Miranda, F. Pan, M. Liu, J. Suhé, Sci. Rep. 2013, 3, 2929.
[29] J. J. Wang, F. Wang, L. Yin, M. G. Sendeku, Y. Zhang, R. Q. Cheng, Z. X. Wang, N. N. Li, W. H. Huang, J. He, Nanoscale 2019, 11, 20497.
[30] M. D. Pickett, G. Medeiros-Ribeiro, R. S. Williams, Nat. Mater. 2013, 12, 114.
[31] J. Park, K. Ahn, S. Yu, J. An, T. H. Lee, M.-G. Kim, Thin Solid Films 2018, 648, 69.
[32] F. Ma, Y. Zhu, Z. Xu, Y. Liu, X. Zheng, S. Ju, Q. Li, Z. Ni, H. Hu, Y. Chai, C. Wu, T. W. Kim, F. Li, Adv. Funct. Mater. 2020, 30, 1908901.
[33] Y. Wang, J. Yang, Z. P. Wang, J. R. Chen, Q. Yang, Z. Y. Lv, Y. Zhou, Y. B. Zhai, Z. X. Li, S. T. Han, Small 2019, 15, 1805431.
[34] R. Guo, Y. X. Zhou, L. J. Wu, Z. R. Wang, Z. S. Lim, X. B. Yan, W. N. Lin, H. Wang, H. Y. Yoong, S. H. Chen, Ariando, T. Venkatesan, J. Wang, G. M. Chow, A. Gruverman, X. S. Miao, Y. M. Zhu, J. S. Chen, ACS Appl. Mater. Interfaces 2018, 10, 12862.
[35] V. Mikheev, A. Chouprik, Y. Lebedinskii, S. Zarubin, Y. Matveyev, E. Kondratyuk, M. G. Kozodaev, A. M. Markeev, A. Zenkevich, D. Negrov, ACS Appl. Mater. Interfaces 2019, 11, 32108.
Yesheng Li received her B.S. and Ph.D. in Materials Science and Engineering from Nanchang University, China, in 2012, and Tsinghua University, China, in 2017, respectively. From 2017 to 2018, she was a semiconductor device design engineer working on the development of power devices and their driver chips at Midea Group, China. Since 2019, she has been a research fellow in the Department of Electrical and Computer Engineering at National University of Singapore. Her current research interests include neuromorphic devices and nanoelectronic materials for next-generation memory and neuromorphic computing.

Kah-Wee Ang is currently a faculty member with the Department of Electrical and Computer Engineering, National University of Singapore. His current research interest focuses on nanotechnology, with emphasis on logic, memory, and optoelectronic devices based on 2D materials for neuromorphic/in-memory computing and mid-infrared sensing.