Memristive Devices for New Computing Paradigms

In Hyuk Im, Seung Ju Kim, and Ho Won Jang

In complementary metal–oxide–semiconductor (CMOS)-based von Neumann architectures, the intrinsic power and speed inefficiencies are worsened by the drastic increase in information with big data. With the potential to store numerous values in I–V pinched hysteresis, memristors (memory resistors) have emerged as alternatives to existing CMOS-based computing systems. Herein, four types of memristive devices, namely, resistive switching, phase-change, spintronics, and ferroelectric tunnel junction memristors, are explored. The application of these devices to a crossbar array (CBA), which is a novel concept of integrated architecture, is a step toward the realization of ultradense electronics. Exploiting the fascinating capabilities of memristive devices, computing systems can be developed with novel computing paradigms, in which large amounts of data can be stored and processed within CBAs. Looking further ahead, the ways in which memristors could be incorporated in neuromorphic computing systems along with various artificial intelligence algorithms are established. Finally, perspectives and challenges that memristor technology should address to provide excellent alternatives to existing computing systems are discussed. The infinite potential of memristors is the key to unlock new computing paradigms, which pave the way for next-generation computing systems.

1. Introduction

Since the first programmable computers were invented, people have no longer been restricted to paper or canvas to process and store information. Due to the significant advances in complementary metal–oxide–semiconductor (CMOS) technology, computing performance has increased drastically based on Moore’s law[1] and Dennard’s law.[2] Currently, computing systems are designed based on von Neumann architecture systems, in which the processor and memory regions are separated and bridged by data buses. With the introduction of cache and improved storage capabilities, and with the development of transistor technology, computing performance has improved significantly. However, the processor and memory performance have improved at different rates. Consequently, the performance gap observed between memory hierarchies causes a delay that is known as a von Neumann bottleneck. According to the 2018 International Roadmap for Devices and Systems (IRDS), conventional computing architectures are expected to reach their physical limits in terms of performance by 2024.[3] However, three significant problems arise, of which the first is volatility. A CMOS operates by reading the capacitance values, which is advantageous in distinguishing on and off states. However, small technical nodes result in high capacity leaks, energy losses, and reliability issues. The second obstacle involves scaling. Because CMOS-based von Neumann computing systems have been developed using a three-terminal structure, which consists of a source, drain, and gate, the device size typically exceeds 6F2 even with smaller feature sizes. As a result, it has become a common trend to fabricate smaller and more integrated devices. The third problem involves speed and energy issues. The incorporation of more sensors and edge computing products has led to data explosion; therefore, data are processed slower due to von Neumann bottlenecks, and an enormous amount of energy is required. Although features, such as bit-cost scalable (BiCS) technology,[4] and a merger of processor and memory have been introduced to overcome these issues,[5] it is still necessary to transition to novel computing systems that are more advanced than CMOS-based von Neumann architectures.

Memristor (memory resistor) technology, which was proposed by Chua,[6] has gained prominence as the most promising novel computing candidate. Unlike a CMOS, memristors, which show pinched I–V hysteresis, identify values through the resistive reading method rather than through the capacitance reading method.[7] Therefore, new computing systems based on memristors would give the opportunity to step toward next computing technologies. Moreover, because memristors can be integrated into crossbar arrays (CBAs), the theoretical density is 4F2, which is expected to overcome the scaling limit of CMOS-based computing.

In this article, we introduce four types of memristor materials and present their operation mechanisms in detail. We describe what memristive CBAs consist of and how they operate, and we demonstrate the unique advantages of integrating memristors into CBAs. Furthermore, we demonstrate two newly proposed computing systems based on memristors. First, we introduce...
memristive computing systems that have enabled in-memory computing and logic design styles using memristors to improve the efficiency of von Neumann computing systems. The concepts and operation principles of memory storage devices are presented as excellent alternatives to NAND (nonvolatile memory) and dynamic random-access memory (DRAM) (volatile memory). Then we examine the tentative stateful logic, which is one type of logic-in-memory that consolidates a latch (storage) and gate (computation). We also introduce neuromorphic computing systems, which are inspired by biological neural networks, and demonstrate how an artificial intelligence (AI) algorithm can be embodied at the device level to process complex data. The output is compared with that of conventional computing systems. Furthermore, we compare the four types of memristors with DRAM and NAND flash memories in terms of various characteristics. Finally, this article explores a wide range of opportunities for applying new computing systems based on memristors and the critical challenges to utilize them as alternatives to CMOS technology.

2. Memristors and CBAs

2.1. Memristors

Memristors have emerged as next-generation nonvolatile memory devices, and they have proven to be excellent alternatives to conventional memory technologies, such as flash, DRAM, and static random-access memory (SRAM). Unlike conventional memory technologies, which program the data as capacitance change, memristors can store data in the form of the difference in the resistance as opposed to the difference in the capacitance, demonstrating nonvolatility with a long retention time of over 10 years.

In 1971, Chua mathematically proved the existence of a fourth fundamental element, memristor (the abbreviation of memory resistor), together with the known fundamental elements (resistor R, capacitor C, and inductor L). The elements are defined as the constitutive correlation among a set of four variables (voltage v, current i, charge q, and flux ϕ). The four variables constitute six different pairs, and five pairs of them \( (v, i), (v, q), (v, ϕ), (i, q), (i, ϕ) \) had already been proven via mathematical relations. Chua derived the mathematical equation, \( dq = M dϕ \), which expresses the missing relationship between \( q \) and \( ϕ \). Chua also demonstrated the pinched hysteresis loop upon controlling electrical stimuli. The concept of the memistor was devised as a two-terminal passive element to express the missing connection between \( q \) and \( ϕ \). At first, Chua’s theory was verified by using active and passive electrical circuit components. Due to its complexity, Chua’s theory was forgotten in the following few decades until its physical validation was demonstrated by HP researchers. They presented a physical model based on a TiO₂ insulator layer sandwiched between two metal electrodes, forming a two-terminal architecture. The demonstration of a memristor by HP researchers in 2008 stimulated the introduction of innovative electronic technologies ranging from storage device technologies to neuromorphic computing. This has galvanized the commercial market to usher in a new age of ultradense nanoscale electronics. Chua’s suggestion in 2011 that all two-terminal nonvolatile memory devices are based on resistance switching, prompted the actualization of the memristor, boosting the studies on various memristive mechanisms: resistive switching, phase-change, spintronics, ferroelectric, and so on.

Memristors are devices that are characterized by their resistance states and are switched between a low-resistance state (LRS) and a high-resistance state (HRS) depending upon the voltage history. When the resistance is varied from HRS to LRS at the threshold voltage \( V_{SET} \), the memristor is “SET.” When it is switched from LRS to HRS at \( V_{RESET} \), the memristor is “RESET.” A certain combination of metal electrodes and insulators organizes the structure of the memristor. When electrical stimuli are applied through the electrodes, the resistance of the insulators undergoes significant change. In a sense, this property can be exploited to store data as a resistance state with a robust retention ability.

The emerging memory technologies include various types of memristors. In this section, four types of memristors are discussed: resistive switching, phase-change, spintronics, and ferroelectric.

2.1.1. Resistive Switching

Resistive switching is one of the most common types of two-terminal memristive devices that have been reported, especially in oxide-based devices, which were introduced with the invention of the TiO₂-based resistive switching device by HP labs. In general, the switching layers are dielectrics, and the fundamental properties of resistive switching are based on the migration of ions, which belong to dielectrics or electrodes. Typically, the ions involved include oxygen vacancies, active metal cations, and anions, such as halides and sulfurs. The mechanism of resistive switching depends on what type of ion migrates in the dielectric, and the typical mechanisms are an electrochemical mechanism (ECM) and a valence-change mechanism (VCM).

The ECM is based on the electrochemical metallization of active metal species, such as Ag or Cu. As shown in Figure 1a, metal cations, which are generated from the electrode, constitute the filament after electro-dissolution, migration, and recrystallization driven by an external electrical bias, mimicking a redox-based nanobattery. An abrupt event in resistance is sparked off following the voltage history. A potential difference promotes the migration of species to achieve the equilibrium condition triggering an electrochemical reaction and causing the hysteresis loop in the \( I-V \) curve. Microscopy has been conducted to observe the growth of Ag nanofilaments in a SiO₂ matrix.

For resetting to the HRS, a reverse electromotive force accompanies the rupture of filaments. Although the principle involved in the RESET process is still unestablished, microscopic observations might be able to clarify the relationship of voltage-activated rupture of filaments. The remaining part of a ruptured filament acts as the preferential site for the filament to grow again upon exposure to the next stimulus.

Anion vacancies, the counterparts of oxygen ions, halide ions, and so on, can be useful to reach an electronic equilibrium by altering their valence states. In general, when a dielectric
containing at least one transition-metal species is sandwiched between inert electrodes, such as Au and Pt, the anion vacancies inherent in the transition-metal compound are preferentially driven to react under a particular potential difference, thereby changing the valence state of the transition-metal cations. These defects are diffused under the electrical field in the dielectric. They act as a medium to satisfy the electrical drift, and they generate a heterogeneous distribution of vacancies. The localized oxygen vacancies lead to the formation of conductive filaments that connect two electrodes. It brings out a change in the resistance of the memristor, showing a hysteresis loop in the $I-V$ curve.

The reset from LRS to HRS is activated by an electrical stimulus in reverse polarity. This allows an opposite migration of vacancies, restoration toward valance states, and the meeting of stoichiometry.

2.1.2. Phase-Change

Phase-change between an amorphous state and a crystalline state, encouraged by Joule heating, enables a device to be characterized as a hysteretic memristor. The memristor based on the phase-change is the most mature technology among all emerging memristor technologies; it is even marketed in the storage class memory (SCM) industry. It can be said that phase-change memory technology has greatly contributed to the development of emerging electronic technologies. The phase-change technology was first reported almost 50 years ago. However, it has been only a few decades since the phase-change technology has come into the spotlight, motivated by research on Ge$_2$Sb$_2$Te$_5$ or Ag- and In-doped Sb$_2$Te$_5$ based phase-change materials. Ge$_2$Sb$_2$Te$_5$ is main material among chalcogenide materials, which are typically researched as phase-change materials. Other examples include GeTe and AgInSbTe. An example of phase-change memory is shown in Figure 1b. When an electrical force is exerted between the top electrode and bottom metal heater, the current saturation through the narrow metal heater stimulates the joule heating process. In sequence, the phase-change material is heated by an internal temperature change. The resistance contrast is derived from phase-change between an amorphous state, acting as an HRS, and a crystalline state, acting as an LRS. The difference in structural disorder between two states is determined by the distribution of internal vacancies. Ordered vacancies give rise to redesign of the band structure, where the localization of charge carriers is formed in a crystalline state causing an increase in the conductivity.

The resistance contrast is determined by how the external voltage is programmed into the metal heater by controlling the magnitude and the length of the pulse. A lower voltage pulse for sufficient time heats the phase-change material to its crystallization temperature ($\geq 500$ K) to melt it, then cools it down slowly, allowing it to be crystallized, and finally sets it (SET). During the programming of pulses, the phase-change material undergoes an incubation and crystallization process, resulting in LRS. To switch the resistance of the device, a higher and shorter pulse is programmed into the phase-change material and melts it at a temperature higher than 1000 K, then rapidly cools it down, and finally resets it (RESET). The melted material would be quenched rapidly and amorphized. Due to a lack of electron density, in consequence, it transitions from LRS to HRS.

2.1.3. Spintronics

Spin-based electronic devices have been developed since the discovery of the giant magnetoresistance in a spin valve structures layered by two ferromagnetic (FM) layers separated by a...
nonmagnetic (NM) metal layer. The relative magnetization alignment can be manipulated by the application of an external magnetic field. The magnetizations of adjacent FM layers, ordered by interlayer exchange coupling, determine the resistivity of these structures. When the magnetizations in adjacent FM layers are oriented in parallel, electrons having the same polarity to the magnetization can flow easily from one FM layer to another, resulting in LRS. However, when they are in antiparallel orientations, electrons flowing through the spin valve structure are very strongly affected by spin-dependent scattering, leading to HRS.

In 1996, Slonczewski first mentioned the spin-transfer torque (STT) effect, in which the magnetization of magnetic structures can be modified by a spin-polarized current without any external magnetic field. This magnetoresistance phenomenon enables the alignment of magnetization to be controlled using the current flowing through a magnetic multilayer.

The magnetic nanostructure consists of a thin NM spacer sandwiched by two FM layers. The FM layers include one FM layer acting as a pinned layer, which maintains its polarization whether or not the current flows through it, and the other acting as a free layer, which is easily modulated by the STT effect of the polarized current. When electrons flow from the pinned layer to the free layer, the electrons are scattered in this pinned layer, oriented to identical polarity with its magnetization. These spin-polarized electrons are partly blocked and partly sent off from the NM spacer. The transmitted spin-polarized electrons exert an angular momentum on the magnetization of the free layer if it has an antiparallel orientation with the pinned layer. This STT effect leads to a reorientation of the magnetization, which is in a parallel alignment, showing decreased resistance. The STT effect can be utilized in spin-transfer torque magnetic random-access memory (STT-MRAM) as a write process.

In addition to the STT effect, the magnetization reorientation can be also induced by spin-orbit torque (SOT). As shown in Figure 1c, the SOT effect in a multilayer of heavy metal (HM), FM, and antiferromagnetic (AFM), which may consist of various other magnetic combinations, originates from a spin-orbit interaction, such as the spin Hall effect (SHE). When an in-plane current flows through the HM layer with a collinear magnetic field, the electrons with antiparallel spins experience spin-scattering to its interface while going through this layer, leading to spin accumulation at the interface. Thus, this accumulation exerts a spin angular momentum on the FM as a free layer, causing reorientation of its magnetization. Through spin-orbit torque magnetic random-access memory (SOT-MRAM), unlike STT-MRAM, a reoriented magnetization is not initiated by thermal effects. Thanks to its advantage, SOT-MRAM has attracted increasing attention.

2.1.4. Ferroelectric Tunnel Junction

Although ferroelectric memory was discovered long ago, it did not attract significant attention due to its large scale for fabrication. Recently, a novel concept of its application, utilizing ferroelectric tunnel junctions (FTJs), was suggested in 2006. It became one of emerging technologies for storage memory and neuromorphic computation.

An FTJ, shown in Figure 1d, consists of two metal electrodes separated by a nanoscale insulator that exhibits the ferroelectric characteristic. It is composed of the ultrathin ferroelectric barrier, and quantum electron tunneling is dominant mechanism in this ferroic nanostructure, in which electrons penetrate through a potential barrier of the ultrathin insulator. The flux of electrons can be modulated by the alignment of ferroelectric polarization in the insulator, generating a giant tunnel electroresistance (TER) effect. The tunneling electrons are repelled or attracted, depending on the polarity of the ferroelectric layer. When the ultrathin ferroelectric layer is sandwiched by two different metals, it results in asymmetry of an energy band profile. With applied voltages, polarization charges are accumulated at the interfaces and affect the electronic potential at the interfaces, which is lowered or raised according to the direction of the polarization. In consequence, electrons flowing through these interfaces are affected by modulated energy potential which is determined by polarization charges. This screening phenomenon produces a large resistance contrast in the FTJ, enabling data storage and processing.

2.2. Integration into CBA

Memristors are based on a two-terminal switching element that can be implemented in a densely packed CBAs. A CBA consists of four components: a word line that is biased, a bit line that is grounded, a selector for rectifying the operation error, and a memristor for storage, as shown in Figure 2a. At every overlap point between each word line and bit line (inset of Figure 2a), the selector (S) and memristor (R) act as one cell (1S-1R). A voltage can be applied through the word line, and the applied voltage exceeds a certain threshold value to exploit the memristor.

The major advantage of this novel architecture is that it overcomes the physical limitations of charge-based memory technologies, such as DRAM and flash memory. The performance of charge-based memory deteriorates with down-scaling. The smaller the technology node gets, the more severe the charge leakage becomes. In contrast, emerging technologies theoretically do not show performance deterioration with size reduction. Furthermore, they can achieve the device dimension of 4F², as shown in Figure 2b which is the smallest size among the existing technologies. Many experts have expected that CBA will be the dominant memory architecture with emerging memristors because of its nanoscale scalability. Due to a facile process and a wide range of material choices, a hybrid memristor combined with a CMOS system through the back end of lines (BEOL) of the CMOS process is also possible.

Compared with the conventional memory that includes a gate, CBA can access each cell individually. A cell can be accessed by biasing into the word line and the bit line, which are connected with the desired cell. The resistance states can be read by sensing the voltage of the reference resistor, which forms a voltage divider with the resistance of the cell. The cells within the architecture can be programmed with a fast speed according to the input bias, which is followed by the immediate output current (Figure 2c).
Massive data from megabytes to terabytes can be stored in an overall CBA composed of numerous crossbar subarrays. Each of these subarrays should be as large as possible and should perform its role without any errors during operation.

### 2.2.1. Inevitable Issue; Sneak Current

The data stored in the memristor correspond to its resistance states. The resistance states, HRS and LRS, are programmed as two Boolean values: “0” and “1.” The programming operation is quite simple. As mentioned earlier, we select the word line and the bit line to program the cell connected with both of them. Ideally, as shown in Figure 2d, only the target cell, which is sandwiched between the selected word and bit lines, should be read, as depicted by its equivalent circuit, where the current only flows through the resistance of the target cell.

Unfortunately, $V_{\text{read}}$ cannot be biased to only the target cell. Due to its architecture in which the metal lines are shared in parallel, many leakage paths flow through the unselected cells sharing the metal lines with the target cell, which is referred to as “sneak current.” This results in parallel biasing to the target cell and an unknown resistor, which is made up of neighboring unselected cells. Sneak current can disturb the operation in a passive CBA due to its contribution to the flowing current through the voltage divider. In Figure 2d, the selected word line is shared among the five cells, and the selected bit line is shared among the same number of cells. In a real case, the current can flow easily through the cells sharing the selected lines if they are in LRS. The worst case is that all the neighboring cells are in LRS. The $V_R$ (applied voltage) is split by the three resistance components, which are half-selected cells sharing the word line, half-selected cells sharing the bit line, and unselected cells sharing the unselected lines.

---

**Figure 2.**

- **a)** A schematic illustration of a memristive CBA. The inset is the cross-sectional SEM view of 1S-1R at a cross-point. Reproduced with permission. Copyright 2013, Wiley-VCH.
- **b)** SEM image of CBA. The cell area is 4F$^2$, which is highly dense architecture. Reproduced with permission. Copyright 2009, American Chemical Society.
- **c)** A plot of write-read-erase cycles measured through current change that is modulated by programmed bias. The top data shows the input bias for write-read-erase sequence operation. The bottom curve shows the output current which shows corresponding resistance state. Reproduced with permission. Copyright 2008, American Chemical Society.
- **d)** Illustration of the operation scheme. Only the selected cell constitutes the voltage divider in ideal case. In real case, sneak current flowing through unselected cell. Reproduced with permission. Copyright 2012, Elsevier.
- **e)** Ideal I–V curve of 1S-1R device. The current at $V_{\text{read}}/2$ decreases when the memristor is connected to a selector. Reproduced with permission.
- **f)** Calculated readout margin ($\Delta V_{\text{out}}/V_{\text{pu}}$) as a function of number of word lines. Reproduced with permission. Copyright 2018, American Chemical Society.
cells not sharing the metal lines. The overall resistance of the sneak current is determined by those three resistances.

The amount of current during the writing and erasing operation is so much larger than that of the sneak current that it is not severe in these cases. However, the sneak current issue is critical in reading operation, which takes place at a lower voltage than that of former operations, resulting in a poor readout margin, as shown in Figure 2f.\(^{[94]}\) Furthermore, it is aggravated with increasing array size; thus, the maximum capacity of the array is limited.\(^{[103]}\) This is one of the significant challenges that should be addressed before memristors can be widely commercialized. Recently, numerous studies have attempted to address this issue,\(^{[93,94,104–109]}\) primarily focusing on the operation schemes to control the voltage of the unselected cells and the selection devices to rectify the sneak currents.

### 2.2.2. Operation Schemes and Selection Devices

Some operation schemes optimize the readout margin by controlling the bias applied to the unselected lines. The all-rounded case is a basic scheme; all unselected lines are simply floated, being left unbiased, and only the selected lines are biased to \(V_{\text{read}}\). Because the voltage that is applied to each unselected LRS cell is unpredictable, this scheme results in the worst readout margin among the aforementioned schemes.\(^{[103]}\) Hence, applying certain voltages to unselected lines, such as the \(V_{\text{read}/2}\), \(V_{\text{read}/3}\), and all-grounded schemes, can help to obtain a reliable readout margin. Typically, the current flowing in an \(N \times M\) CBA can be derived from Kirchhoff’s law, which is used to calculate the current continuity through a numerical method, if the all voltage variation in the CBA can be estimated.\(^{[110]}\) Research is still underway for the optimum operation schemes yielding the best readout margin with the lowest power consumption.\(^{[103,110–112]}\)

As well as the certain voltage being applied to the unselected lines, the introduction of selection devices in series with the cells is another solution to restrict the sneak current. The selection devices should have rectifying characteristics, reducing the current in a low voltage area, which is equal to the applied voltage at an unselected cell, such as \(V_{\text{read}/2}\) or \(V_{\text{read}/3}\). Figure 2e presents a typical \(I–V\) curve in a cross-point composed of one selection device (1S) and one memristor (1R). If the memristor is in LRS, the current flowing through the 1S-1R significantly decreases compared with the current flowing through the standalone 1R, resulting in a negligible contribution of the sneak current to a reading current. Lee et al. reported an oxide barrier using a \(\text{TaO}_x/\text{TiO}_2/\text{TaO}_x\) structure to suppress the sneak current.\(^{[113]}\) When the standalone TiO2 layer, which showed resistive switching, was in LRS, the current reached 1.6 mA at \(V_{\text{read}/2}\). However, the current decreased to 37 nA with application of the oxide multi-barrier as a selection layer. Hua et al. demonstrated the ultra-low leakage current of \(<1 \text{ pA}\) by introducing an Ag nanodots threshold switch which shows a very steep slope of 0.65 mV dec \(^{-1}\). Consequently, an enhanced readout margin was achieved.\(^{[105]}\)

Typically, the selection devices that are used to reduce the leakage are bipolar diodes,\(^{[94,105,107,109,113,114]}\) unipolar diodes,\(^{[90,106,108]}\) and transistors.\(^{[115,116]}\) Lee et al. investigated the 1S-1R structure using an Ag-doped \(\text{SiO}_2\)/\(\text{N}_x\) threshold switch that is a bipolar diodes and achieved an increased maximum value for word lines (N) from 6 to \(3.5 \times 10^5\), satisfying at least a 10% readout margin (Figure 2f).\(^{[104]}\) which should be met for achieving the desired error probability.\(^{[117]}\)

Yoon et al. fabricated a double-layer-stacked 1D-1R CBA using a \(\text{TiO}_2\)-based unipolar diode with a rectification ratio of \(8.4 \times 10^6\).\(^{[108]}\) A 1BT-1R structure was presented by Aluguri et al., who adopted an oxide-based p-n-p bipolar transistor and demonstrated an increased number of word lines for a 10% readout margin, achieving 806-word lines, which can make a 600 kB CBA.\(^{[115]}\)

Diverse approaches have been taken to develop new applications of memristors. The CBA of memristors can be utilized not only as computing components\(^{[115,32,98,118]}\) and photovoltaic cells\(^{[119,120]}\) but also neuromorphic systems.\(^{[121]}\) Although the main focus has been directed on the storage, various novel concepts have been proposed ranging from memristive logic and photo-memory to neuromorphic systems. The concepts of memristive computing and neuromorphic computing, which are the improved or reformed von Neumann architecture, will be discussed in detail in the next sections.

### 3. Memristive Computing

#### 3.1. Storage

As previously mentioned, conventional memory technologies encounter challenges of scaling and performance. Currently, the dominant technologies for the SCM and main memory are NAND and DRAM, respectively. To improve the von Neumann computing, the engineers have been seeking a breakthrough to reorganize the conventional von Neumann structure. Recent studies have demonstrated that the memristive CBA has outstanding performance\(^{[122–124]}\) and scalability;\(^{[12,13,95]}\) therefore, it is emerging as the most likely candidate to replace CMOS technologies. Furthermore, Intel released Optane memory, which is based on phase-change memory technology, has redefined the traditional hierarchy between NAND (SCM) and DRAM (main memory).\(^{[50]}\) This section will discuss the parameters that determine the practical use of memristive storage.

In current semiconductor markets, the main goal of memristive systems is the new generation of storage memory technologies.\(^{[3]}\) To ensure its reliability, endurance, and retention, a distinguished memory window (on/off ratio) is required. Endurance is the maximum number of cycles to reliably switch between on and off; thus, the higher the endurance is, the better it is for frequent programming. Retention is the most critical parameter in nonvolatile memory technology; it indicates the time until stored data is lost. DRAM suffers from a short retention time, requiring repeated refresh steps, and finally inducing power dissipation. Therefore, retention time is strongly related to energy efficiency, which is important as the data increase. The on/off ratio determines the memory window of a device and the memory window affects operation error probability during programming of the memristor. Therefore, the on/off ratio should be as large as possible to ensure reliable storage operation.
Recently, various memristive studies at the single-cell level have shown a high endurance of $>10^{14}$, long retention of $>10$ years, and a large memory window of $>10^8$. Table 1 summarizes the scalability, reliability, and performance compared with conventional CMOS-technologies such as NAND and DRAM. In all aspects, memristors are as competitive as CMOS-technologies, verifying that the memristors will form the basis of next-generation storage technology.

For practical large-scale memristive storage, the devices should guarantee a considerable yield when they are integrated. As shown in Figure 3a, Jo et al. demonstrated a large-scale (2 Gbits cm$^{-2}$) CBA based on a Si-based memristor. Figure 3b shows the yield map of 400 bits within the 1 kb CBA. All crosspoints were written by 200 μS pulses to switch to LRS, resulting in a 92% yield of ON states. The device demonstrated a notable on/off ratio and considerable uniformity.

A CBA can be readily fabricated at low temperatures. In a sense, it can be stacked on top of an underlying CMOS through the BEOL process. Here, the underlying CMOS eliminates the sense, it can be stacked on top of an underlying CMOS through CMOS-technologies, verifying that the memristors will form the basis of next-generation storage technology.

Table 1. Summary and comparison of four types of memristors for memristive computing and neuromorphic computing. All data confirmed at temperature below 85 °C are given.

| Technology | CMOS-based | Memristor-based |
|------------|------------|----------------|
| Device dimension | 4F$^2$ | Memristive computing |
| Feature size (F) | 15 nm | 10$^{-13}$ |
| Density | 512 GB | 32 GB |
| On/off ratio | 10$^{-11}$ | 10$^{-18}$ |
| Retention | $>10$ y | 64 ms |
| Operation energy | 160 pJ | 8 fJ |
| Programming speed | 80 μs | 10 ns |
| Multi-level operation | – | – |
| Variability | – | Medium |
| Linearity | – | Low |
| Research matured | – | 2 Mb array pattern recognition |
| Pros | High density | Gradual conductance change |
| Cons | Individual inaccessible | Nonlinearity |

When the devices consist of flexible dielectrics, such as organic material, and fine-patterned electrodes, they can easily be adapted to flexible memory because of memristors’ simple structure. Consequently, they have potential for application in wearable electronics and flexible radio-frequency identification (RFID) tags, which are intensively studied for utilization in the Internet-of-Things. Figure 3f shows a microscopic image of CBA based on Ag microwire/silk fibroin microwire/Ag microwire, which was demonstrated by Pan et al. This device shows sufficient storage performances under a 5 nm curvature radius bending as shown in Figure 3g.

Due to its versatility, other studies have been conducted on specific applications of CBA storage, such as photo-memory which is a combination of a photodetector and a storage element. Furthermore, multi-state memristors have been intensely reported by which ultradense storage devices can be achieved.

Although most studies on memristors have focused on their use in storage applications, recently, there have been many discussions on the limitation of von Neumann computing, promoting many attempts to apply memristors as other computing elements such as logic and neuromorphic systems.

3.2. Logic

Due to the challenges experienced with von Neumann technology, a novel computation device has been developed to process explosive amounts of data. Many reports have suggested that computation efficiency can be improved by logic-in-memory computing, which does not require data transfer between the gate and latch. Due to fast access speed and nonvolatility, memristors are strong candidates to compose stateful...
logic, a type of logic-in-memory computations. Stateful logic is the computation that performs logic and stores values in the same dimension, mapping the logic states (input and output) as the resistances of the memristor. The logic values of Boolean functions, 0 and 1, are denoted by the resistance states, HRS and LRS, respectively.

Whitehead and Russell suggested in 1910 in Principia Mathematica that the IMP (material implication, “if p, then q”, \( p \rightarrow q \)) is the fourth fundamental logic operation, along with basic logic operations, such as AND, OR, and NOT. Although IMP gained prominence in the electrical engineering and computing fields, it has been investigated insufficiently since 2010 when HP Labs demonstrated a memristor-based IMP. It is the first memristive stateful logic system in which the gate (computation) and latch (storage) are combined, not separated, unlike conventional computation circuits.

Figure 4a shows a schematic of the IMP gate: two memristors, named \( P \) and \( Q \), are connected to a resistor \( R_C \) (\( R_{\text{OFF}} > R_C > R_{\text{ON}} \)), and the logic states of \( P \) and \( Q \) are represented as \( p \) and \( q \). The memristors, \( P \) and \( Q \), are biased to \( V_{\text{COND}} \) and \( V_{\text{SET}} \), respectively, during IMP operation. Here, \( V_{\text{SET}} \) can switch the state of the memristor \( Q \), and \( V_{\text{COND}} \) has a lower magnitude than \( V_{\text{SET}} \), leaving the state of \( P \) unchanged. After IMP operation, the changed state of \( Q \), which is identical to the output, is checked by applying \( V_{\text{read}} \). The initial states of the memristors represent logic inputs, and the final resistance

---

**Figure 3.** a) A Schematic diagram of 20 × 20 CBA. b) The yield map of 20 × 20 CBA. The “not written” indicates the device which was not programmed. Reproduced with permission.[91] Copyright 2009, American Chemical Society. c) A Schematic diagram of the hybrid integrated system. d) An input image data to be stored in the array. e) The output image data obtained by storing the data. Reproduced with permission.[98] Copyright 2012, American Chemical Society. f) Flexible CBA with Ag microwire electrodes and silk fibroin interlayer. g) The I–V result under 5 nm curvature radius bending. Reproduced with permission.[145] Copyright 2019, Wiley-VCH.
The state of memristor $Q$ is an output of this computation. The process of IMP operation is summarized as follows. 1) Reading the inputs: reading out the resistance of the memristors ($p$ and $q$); 2) IMP: applying the voltages ($V_{\text{COND}}$ and $V_{\text{SET}}$) to $P$ and $Q$, respectively; 3) Reading the output: reading out the resistance of the memristor ($q^*$).

The truth table of IMP is shown in Figure 4b. If $P$ and $Q$ are in HRS (case 1: $p = 0$, $q = 0$), the voltage of memristor $P$ is approximately equal to $V_{\text{COND}}$, and the voltage of memristor $Q$ is approximately equal to $V_{\text{SET}}$. Because the voltage divider is composed of the memristors and $R_G$, which is smaller than $R_{\text{OFF}}$, most of the voltages are applied to memristors, resulting in an unchanged state of $P$ and LRS of $Q$, which is switched by $V_{\text{SET}}$. As a consequence, the logic output is 1 ($q^* = 1$). In case 2 ($p = 0$, $q = 1$), $Q$ is already set; it is not affected by $V_{\text{SET}}$ ($q^* = 1$), and $P$ is also left unchanged due to insufficient voltage on it ($|V_{\text{COND}} - V_{\text{SET}}|$). In case 3 ($p = 1$, $q = 0$), the common node undergoes $V_{\text{COND}}$ due to the low resistance of $P$. Therefore, the voltage of $Q$ is $V_{\text{SET}} - V_{\text{COND}}$, which is insufficient to switch the state of $Q$, so that, it remains HRS ($q^* = 0$). When all of the memristors...
are in LRS ($p = 1$, $q = 1$), there is no change in the states of the memristors ($q^* = 1$).

An IMP gate can be integrated within a CBA (Figure 4c). Taking advantage of its scalability, it can remarkably reduce the required size of a logic gate and can be fabricated within a massive CBA. More than two memristors are required to execute a different Boolean function, such as NAND. Through the three memristors ($P$, $Q$, and $S$) which are placed within the same row of the CBA, the NAND computation can be performed, as shown in Figure 4d. Sequential operation of FALSE (write 0) and two IMP results in NAND operation, as shown in Figure 4e. The initial FALSE ($s = 0$) operation should be executed where the $V_{\text{CLEAR}}$ pulse makes the memristor $S$ “open” (HRS). The two memristors ($P$ and $Q$) act as inputs on sequential steps. Subsequently, $P$ and $S$ compute the IMP operation by applying the voltages $V_{\text{COND}}$ and $V_{\text{SET}}$, respectively, producing the output ($s' \leftarrow p$ IMP $q$). The result ($s'$) undergoes another IMP operation with $Q$. This leads to the result ($s'' \leftarrow q$ IMP $s'$). Finally, computation of the NAND operation is completed by total computation sequences ($s'' \leftarrow p$ NAND $q$).

Other general Boolean functions, such as AND, OR, and XOR, can be implemented by various IMP sequences. For example, to perform an AND operation ($p$ AND $q$), the computational step of IMP is ($p$ IMP ($q$ IMP 0)) IMP 0. That is, the execution of multi-input is able to compute multiple implication of Boolean functions within the CBA. This is the major advantage of the IMP stateful logic.

Another approach to utilize the memristor as a logic gate is a MAGIC, which is short for memristor-aided-logic, as shown in Figure 4f. In 2014, Kvatinsky et al. suggested the MAGIC NOR gate which, unlike IMP, has a rather simple structure that does not require a reference register $R_C$, and it has separate inputs and output memristors. As IMP operation, the MAGIC gate is a type of the memristive stateful logic, in which the inputs and output are the resistance of the memristors. For the basic MAGIC NOR computation, three memristors are required, two input memristors ($I_{in1}$, $I_{in2}$), which are connected in parallel, and one output memristor (Out), which is connected in series with the input set. Because the polarities are in the opposite directions between the input set and output, the applied positive voltage tends to reset the inputs and set the output.

The MAGIC operation always includes the initialization step for switching Out to logical 1. The applied voltage $V_0$ should be satisfied with Equation (1) for NOR operation, wherein the output should be logical zero unless all the inputs are logical zero

$$2V_{\text{RESET}} < V_0 < \frac{R_{\text{OFF}}}{2R_{\text{ON}}} \cdot V_{\text{RESET}}$$

Its sequence comprises the following steps: 1) Initialization of the Out to LRS (Out$_{\text{init}} = 1$); 2) Applying the voltage $V_0$ across the MAGIC (the voltage applied to Out depends on the state of the inputs); 3) Reading out the final logical state of the Out. In the case that all the inputs are in HRS ($I_{in1} = 0$, $I_{in2} = 0$), most of the applied voltage $V_0$ is expended on inputs due to their high resistance. Consequently, the magnitude of the voltage on Out is lower than the threshold voltage ($V_{\text{RESET}}$) leaving the logical state of the Out unchanged (Out = 1). If one of the inputs is in LRS, or all the inputs ($I_{in1} = 1$, $I_{in2} = 1$) are in LRS, the voltage on Out has a considerable magnitude, resulting in a change in the logical state of Out (Out = 0). The truth table of MAGIC NOR is shown in Figure 4g.

The MAGIC NOR operation is the stateful logic that preserves the inputs and stores the output, not being observed in IMP operation. It can also be fabricated within a CBA for ultradense logic gates. Furthermore, through MAGIC technologies, other Boolean functions, AND, NAND, NOR, and OR can be performed as standalone gates.

The aforementioned stateful logics have a critical drawback as the sequential computation requires a long time due to its lengthy steps. For example, the IMP operation organized by n-inputs should take $2^{n+1} + 1$ sequential steps, making it much slower with more complicated functions. Furthermore, when cascaded, the signal is decayed due to the passive characteristic of a memristor. Some studies have suggested that integrating the memristive logic gates with CMOSs, such as memristor ratioed logic (MRL), is an effective treatment for the restoration and inversion of the signal to cascade memristive logic gates and construct various Boolean function, such as NAND and NOT.

Although these drawbacks postpone the replacement of conventional CMOS-based logic gates with memristive logic gates, it is a notable point that the stateful logic holds an advantage in terms of power dissipation in comparison to CMOS-based gates. The concept of the memristive processing unit (MPU) was suggested in 2016 by Talati et al., who demonstrated the design algorithms of a MAGIC-based full adder. However, the progress of memristive gate research toward achieving a mature system is ongoing. The proposed memristive logic gates will enable advances toward the era of MPUs that is more advanced than that of CPUs.

4. Neuromorphic Computing Systems with Memristors

Although studies have been conducted to improve the conventional von Neumann architecture, a speed gap problem has been identified that degrades its performance. This is known as a von Neumann bottleneck, and it is caused by the physical separation of the processing unit and the memory. To address this issue, a neuromorphic computing system, which is inspired by biological neural networks, has been introduced for low-power, high-speed, and high-density computing. The new system processes and stores data by elucidating how the information is connected. This theoretical concept was not actualized until transistor-based technologies showed tremendous improvement in performance. Recently, a wide range of neural network technologies has been studied and developed to enable AI applications, such as image recognition, voice recognition, and autonomous driving, supported by advanced semiconductor technology. The neural network technologies exhibited a breakthrough in AI performance, especially in the artificial neural networks (ANNs) and spiking neural networks (SNNs).

Biological neurons consist of dendrites, soma, and axons. Synapses are formed when dendrites come into contact with axon terminals. Biological neural networks store information by controlling the strength of the connections between neurons. Focusing on this, ANNs update the synaptic weights using a
multi-vector matrix and activation functions with hidden layers. Moreover, SNNs learn through spike time-based updates. Figure 5 presents this schematically.

However, such complex algorithms have to be implemented with the conventional CMOS-based von Neumann architecture, which consumes high energy, and use energy inefficiently to operate the cells. The memristor has emerged as a promising alternative to existing computing devices because it can store multi-states at each device level rather than only storing two Boolean values. This novel device that can be utilized for CBA has the potential to overcome the limitation of current computing systems at the hardware level. We provide more details about the operation mechanism of ANNs and SNNs and discuss how the memristor exhibits the synaptic characteristics required for each algorithm. Furthermore, we also elucidate how the new computing system is used for the calculation of AI.

### 4.1. ANNs with Memristors

The first mathematical model of ANNs was proposed by McCulloch and Pitts in 1943\[169\] and was first implemented by Rosenblatt in 1958\[170\]. However, it was inefficient and did not exhibit sufficient computing power to implement the new concept. Thanks to the advances in CMOS technology, the remarkable development of ANNs has significantly contributed to progress in modern computing, allowing the introduction of new applications, such as computer vision\[171\], autonomous mobile robotics\[172\], speech recognition\[173-175\], face recognition\[176,177\], and medical prescription\[178,179\].

ANNs are networks that have layers and mathematical activation functions with a weighted sum of inputs to output values. The purpose of ANNs is to optimize a model to best deduce the output from the input by adjusting the synaptic weight, which is done through back-propagation operations, such as stochastic

---

**Figure 5.** Schematic of the a) biological neuron, b) biological neural networks, c) artificial neuron, d) ANNs, e) spiking neuron, and f) SNNs. In biological neurons, neural signals enter the dendrite and transmit to the axon terminal. They configure synapses to attach. Artificial neurons, which are inspired by biological nervous systems, update synaptic weight with vector multiple matrices and form a network through mathematical activate function. Spiking neurons use time-based encode spike data and store the synaptic weights using the time difference between pre- and post-neuron activity.
gradient descent\textsuperscript{[180]} and chain rules.\textsuperscript{[181]} These networks can be trained by supervised or unsupervised learning, and they can process a large amount of training data in parallel. Furthermore, with the help of accelerators such as graphic processor units (GPUs), field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs) that are designed to perform more efficiently than a CPU would, more complex and more parallel calculations are possible. This has allowed the development of various forms of ANNs, such as multi-level perceptron (MLP), convolution neural networks (CNNs), and deep neural networks (DNNs), and they have led to remarkable progress in AI.

However, these CMOS-based computing systems still meet many challenges with processing ANNs. Because ANNs are derived from the mathematical modeling of biological neural networks, CMOS-based computing systems can perform a very large number of complex and parallel computing operations, resulting in inefficient storage of synaptic weights in bits. They also require high power consumption because they have a complex circuit design and limited scalability. Hence, a new computing system paradigm is required.

A gradual change in the conductance of memristors in terms of synaptic weight change can be implemented in developing a new computing system paradigm. In ANNs, synaptic weight change during back-propagation requires training. Synaptic weight change can be implemented applying a voltage pulse or repeated voltage sweep. Consider, CMOS-memristor hybrid hardware and artificial synapses are proposed to perform CNN and DNN operation.\textsuperscript{[182]}

Numerous memristive materials exhibit synaptic potentiation and depression through electrical stimuli. Xiao and Huang, reported halide perovskite-based resistive switching neuromorphic devices\textsuperscript{[183]} fabricated from a simple metal–insulator–metal structure with polycrystalline MAPbI\textsubscript{3} films. In the two-terminal structure, shown in Figure 6a, Au top and ITO/PEDOT:PSS bottom electrodes act as pre- and post-neurons, respectively, and polycrystalline MAPbI\textsubscript{3} films act as synapse layers. As shown in Figure 6b,c, the device shows six-step positive analog resistive switching and six-step negative switching due to the I\textsuperscript{−} or MA\textsuperscript{+} interstitial ion migration through grain boundaries, which provide ion migration paths. The I–V characteristics correspond to memristive behavior, and this contributes to the gradual change in conductance of an artificial synapse. When positive voltage pulses are applied, charged ions and vacancies migrate to the electrode. In contrast, negative voltage pulses cause ions to move to the counter-electrode. Through these mechanisms, repeated potentiation and depression occur by applied voltage pulses with fixed amplitudes of \textpm 2\textdegree V and 0.75 V read voltage pulses as shown in Figure 6d,f. The synaptic weight changes through potentiation and depression have been repeatedly tuned for more than 500 cycles. Extensive research has demonstrated the linear, parallel, and reproducible synaptic weight changing at the device level. Resistive switching,\textsuperscript{[185,186]} phase-change,\textsuperscript{[187]} magnetic tunnel junction,\textsuperscript{[184]} and ferroelectric\textsuperscript{[180]} memristors form the focus of current research to demonstrate synaptic weight change characteristics.

Several attempts have been made to create new peripheral circuits using memristors in a CBA to control the read and write operations during the synaptic update process.\textsuperscript{[190]} Figure 6f presents the peripheral circuit implementation of DNNs with forward-propagation, back-propagation, and weight update with memristive CBAs.\textsuperscript{[184]} During forward-propagation, the dot products of inputs are calculated in crossbar 1, and the main block (MB1) places them in the activation functions. After sequentially processing the conductance values, the MB3 calculates the synaptic weight with gradient descent and applies update voltage pulses to each CBA. The back-propagation algorithm updates the synaptic weight to minimize error. ANNs integrated with memristors are expected to reduce the architecture complexity and improve AI performance.

4.2. SNNs with Memristors

Natural biological neural networks, such as the brain, are based on spike-event-driven processing. In other words, neurons are naturally in a resting state and only transmit spikes when activated. In addition to ANNs that learn through mathematical derivation, more bio-inspired SNNs have been developed, which mimic the functions of biological neural networks. SNNs encode information in the pulse timing or frequency differences between pre- and post-synaptic spikes, whereas ANNs process information with floating numbers. Moreover, SNNs mimic activation neurons with the leaky integrated and fire (LIF) model, which fires electrical spikes when a threshold voltage is reached. Therefore, SNNs are considered to be the next network architecture model of ANNs because they are more similar to actual neurons and more energy-efficient than ANNs.\textsuperscript{[191]}

Synapses perform calculations by controlling the strength of connections with neuronal activity; this phenomenon is known as synaptic plasticity. Hebb identified the physiological learning rules of synaptic plasticity, known as Hebb’s learning rule,\textsuperscript{[192]} in which the connection strength between neurons changes based on neural activity in pre- and post-neurons. Supported by Hebb’s theory, SNNs learn by spike-timing-dependent plasticity (STDP) implementation. STDP is synaptic plasticity in which a synaptic weight is changed based on the time difference between the neurons’ transmission of a spike signal.\textsuperscript{[193]}

Several phase-changing memristor devices have been demonstrated to be feasible with SNNs.\textsuperscript{[194,195]} Kuzum et al. presented programmable synaptic devices based on phase-change material for SNNs.\textsuperscript{[196]} Figure 7a shows a scheme of spike-based synapses in an SNN CBA compared with biological synapses consisting of pre- and post-neurons. As biological spike transmission, pre-spikes are transmitted through the bottom electrode lines and post-spikes are transmitted based on time differences through the top electrodes. The memristive synapses, which are composed of chalcogenide glass, exhibit a phase-change synaptic behavior between amorphous (shown in red) and crystalline (shown in green) states due to heating. To present the mechanism in detail, cross-sectional transmission electron microscopy (TEM) images are shown in Figure 7b. The greater the number of phase-changes from the set state to reset state by electric spikes is, the larger the volume of the amorphous region is. The amorphous mushroom-like region can be confirmed by the selected area electron diffraction (SAED) patterns, whereas other areas are confirmed to be polycrystalline. The implementation of STDP in the devices is shown in Figure 7c. Compared
with biological STDP measured in hippocampal glutamatergic synapses, the synaptic weight changes at almost the same rate as the change in the time differences in accordance with the asymmetric Hebbian rule. Moreover, numerous types of devices based on memristor materials, such as resistive switching\cite{185,197,198}, magnetic tunnel junctions\cite{199} and FTJs\cite{200,201} have been reported for application to the operation principles of SNNs.

At the architecture level, numerous efforts have been made to implement hardware designs for SNNs using memristors. Figure 7d shows the basic architecture of SNNs\cite{190}. The pre- and post-synaptic neurons are connected in the CBA. This architecture is implemented by applying the LIF model, a winner-take-all bus, and STDP-compatible spike generation.

These studies are expected to lead to the development of memristor hardware systems with a close resemblance to the actual brain.

### 4.3. Machine Learning with Memristors

Conventional programming approaches are implemented by the programmers to determine what patterns are repeated in given datasets and to reduce errors during information processing which is based on algorithms. For relatively simple problems, such approaches can be handled simply, and maintenance is also easy. However, when the amount of data is huge, the practical ability to adjust and debug codes reaches its limits. Machine learning makes the program codes much shorter, easier to debug, and much more accurate.
This is because the neural network algorithm itself learns the patterns that appear for given problems even if programmers do not perform debugging. In other words, network models learn and update the hidden rule on their own, even if programmers do not perform any modification. Such novel algorithms are suitable for complex problems that demand a large number of manual controls or have no existing solutions at all.

Due to the progress of CMOS computing systems, neural network learning produces remarkable results. To effectively calculate AI algorithms, numerous semiconductor companies and research institutes have developed neuromorphic chips that use artificial neurons made of transistors. Several chips focus on complex parallel calculations, including Google TPU, Tesla FSD, and SambaNova. Alternatively, some chips have been designed to produce neural spiking, including SpiNNaker, Neurogrid, and TrueNorth developed by International Business Machines corporation (IBM) and Loihi developed by Intel. In addition, the Tianjic chip, which is hybridized with both of these methods, was also developed. Although there are many applications of machine learning, the performance of computing hardware is limited by current CMOS computing systems when dealing with complex problems. The emerging memristors computing systems enable faster operation with lower power consumption in comparison with the current CMOS-based systems.

Machine learning is divided into supervised, unsupervised, and reinforced learning. All of these are conducted through updating of the synaptic weights in neural networks; therefore, ANNs and SNNs methods can be applied.

Supervised learning refers to the process of determining the most suitable predictive model using data with known outcomes. Because the model handles labeled data, it is suitable for pattern-recognition problems. During the update of synaptic weight in supervised learning, massive parallel calculation and multiplication are involved. Resistive switching memristors with a crossbar structure are appropriate to meet this demand. Li et al. demonstrated 1T1M TaOₓ/HfAlOₓ 1024 cells in a CBA with parallel online learning. Yao et al. also demonstrated that a 128 × 64 1T1M HfOₓ CBA could perform gray-scale face classification with an accuracy of 91.71%. Cai et al. reported a passive selector with WOₓ in 54 × 108 array chips, which achieved an accuracy of 94.6% in Greek letter classification. Phase-change was studied as well using supervised learning. Ambrogi et al. first proposed a mixed hardware–software neural
network implementation with 204 900 phase-changing synapses in a 3-transistor, 1-capacitor system.\textsuperscript{[211]} They showed that the devices achieved an accuracy of 98\% for CFIR-10, 100, and the MNIST database.

In contrast, unsupervised learning deals with unlabeled data which consists of unguided information. Because it learns the relationship between data and finds hidden features to make inferences regarding new data, unsupervised learning is usually used to cluster problems. Choi et al. demonstrated feature extraction and dimensionality reduction with an accuracy of 97.1\% using an 18 × 2 TaO\textsubscript{x} CBA.\textsuperscript{[212]} Jeong et al. also developed a 4 × 3 CBA with TaO\textsubscript{x} and demonstrated successful K-mean clustering with an accuracy of 93.3\% with unlabeled iris flower datasets.\textsuperscript{[213]} Furthermore, Wang et al. developed 8 × 8 fully diffusive memristive neural networks with SiO\textsubscript{x} and Ag.\textsuperscript{[214]} The synaptic weights were determined by silver migration with circuit capacitance. They also used a stochastic LIF model for unsupervised pattern recognition.

The other learning type is reinforcement learning. It differs slightly from supervised learning because the targets correspond to rewards, and the results correspond to actions. The emergence of DNNs has enabled advanced reinforcement learning, which is exemplified by AlphaGo. Using nanoscale FTJs, 5 × 5 analog in-memory reinforcement learning systems were reported by Berdan et al.\textsuperscript{[215]} A 128 × 64 1T1M CBA with HfO\textsubscript{x} was implemented for deep reinforced Q-value learning by Wang et al.\textsuperscript{[216]}

Diverse studies on simulations and chip operations have been conducted to assess the applicability of these learning algorithms in memristor hardware. According to Hu et al.,\textsuperscript{[217]} the analog memristor 128 × 64 CBA achieved a result of high precision in handwritten pattern recognition as shown in Figure 8. TEM and cross-sectional scanning electron microscopy (SEM) images of one-transistor one-memristor (1T1M) with hafnium oxides in a CBA are shown in Figure 8a. The transistor gate acts as a selector for various applied different current compliances for the memristor. The CMOS and memristor layers are integrated by a foundry-compatible BEOL process. The 128 × 64 CBA chip, which was fabricated in the lab, is shown in an optical microscopic image in Figure 8b. They experimentally demonstrated the classification performance of a dot product engine (DPE) using a single-layer neural network for the full set of 10 000 modified National Institute of Standards and Technology (MNIST) handwriting database classification. Figure 8c shows the computing sequences of DPE. A given 19 × 20 pixel image is unwrapped to 380 input vector voltages and classified as ten possible numbers (0–9) by applying the memristor matrix. Figure 8d shows the classification results obtained by a CMOS computing simulation and by the memristor neural network hardware for the entered numbers. Furthermore, Figure 8e shows the accuracy as the result of learning with ten possible numbers. The inference of a neural network yielded a recognition accuracy of 89.9\% for 10 000 MNIST images. These findings confirmed that the memristor neural network can be successfully applied to existing AI systems through learning. As these research fields continue to evolve, it is expected that new computing systems will be able to incorporate various types of learning into the new neural networks.

Figure 8. a) TEM, cross-sectional SEM images, and a schematic of 1T1M memristor device array. b) Optical microscopic image of 128 × 64 CBA integrated on the chip and a schematic of 1-transistor 1-memristor CBA. c) Illustration of the DPE classification. A MNIST 19 × 20 pixel image is converted to 380 input vectors as conductance values and processed by the DPE classification. d) Some example of classification result with simulation (red) and experimental result (blue). e) Recognition accuracy for 10 000 images from the MNIST database. Reproduced with permission.\textsuperscript{[217]} Copyright 2018, Wiley-VCH.
5. Perspective and Challenges

The massive generation of information, which will reach 175 zettabytes in 2025,[218] has necessitated efforts to achieve a seminal transformation toward post-CMOS technology, which is expected to overcome the disadvantages of current CMOS technology, including volatility, scaling limit, and the von Neumann bottleneck. Intensive investigations have verified the outstanding performance of memristors; thus, they can be used as excellent alternatives to conventional computing architectures. However, there are several challenges that hinder the wider adoption and commercialization of memristors. Therefore, further research should be done on materials, devices, architectures, and neural network algorithm levels to address these obstacles.

5.1. Materials and Device Engineering

Many researchers have struggled to understand, develop, and integrate the memristors to develop alternatives to CMOS-based systems. Nevertheless, there are still obstacles to the adoption of memristors to replace CMOS-based computing. For novel computing systems, memristive computing, and neuromorphic computing, several requirements should be satisfied in terms of scalability, reliability, and performance. In Figure 9 and Table 1, parameters related to these requirements are summarized and ranked in relation to the four types of memristors.

Scalability is the key parameter to address the challenge of handling large volumes of data, which is determined by device dimension, feature size, and density. The device dimension refers to the area required for the construction of one cell,
and it is calculated by reciprocating the feature size, which is a manufacturing factor meaning the width of a metal line, to the second power. Density indicates the researched subarray size that can be fabricated. In terms of scalability, resistive switching and phase-change devices are the best candidates due to their facile fabrication process. The spin-dependent scattering operation and epitaxial growth of ferroelectric material deteriorate the scalability of spintronics and FTJs.

For robust operation, memristors should have a reliable on/off ratio as well as reliable endurance and retention. Especially in spintronics and FTJs, a low on/off ratio might be caused by the temperature-dependent disturbance of spin states. SOT-MRAM has the advantage of being less reliant on temperature than STT-MRAM; therefore, it provides a large window memory. Resistive switching and phase-change, which are related to ion migration, seem to be the best candidates for reliable operation.

Programming speed and operation energy determine the performance of the computation elements that are important for eliminating the von Neumann bottleneck problem. Phase-changing requires a large current to elevate the temperature, and a long time is required for crystallization, which results in poor performance. Furthermore, FTJs show poor performance due to the transition of ferroelectric polarization which can be induced at high voltages. Because of the advantages of their nanoscale channels and quantum nature, respectively, resistive switching and spintronic memristors achieve outstanding performance.

To be widely adopted in neuromorphic computing, memristors should show neural plausibility in addition to the three previously discussed features. Neural plausibility is the property of electronic devices that indicates how well they simulate nervous systems. I–V symmetry, linearity, variability, and multi-level states are integral to neural plausibility.

I–V symmetry indicates how symmetrically resistance hysteresis appears when positive and negative voltages are applied. If the devices are asymmetric, when driving on circuits, it makes it difficult to design and apply various circuits.

Inconsistent conductance changes make it difficult to indicate the target value when applying the same electrical signal, which significantly reduces the accuracy of the training process. Linearity indicates how uniformly conductance increases as the weight update pulse number increases. The filaments of a resistive switching device repeatedly disappear, which affect conductance and result in low linearity. In contrast, ferroelectric junction devices show excellent linearity because once the orientation has changed, it becomes consistent.

Stochasticity is the relative level the conductance within the range where it has fluctuated over a certain period of time. In other words, stochasticity indicates the level of fluctuation when the saved conductance is not affected by any other electrical stimuli during a specific time period. Synaptic weight is a time-based fluctuating value; according to the network theory, it increases cognitive efficiency. When stochasticity is low, the recognition performance of ANNs and SNNs shows high accuracy. In the case of a phase-changing device, the conductance drifts more severely than in the other memristive devices because it operates by Joule heating. Due to this mechanism, it has high stochasticity, resulting in poor performance.

In storing the synaptic weight value, the number of bits expressed by memristors greatly affects performance and integration in computing systems. For example, if a cell has a large number of distinguishable conductance states, the number of bits per device increases, giving rise to an exponential rise in performance, compared to digital devices which store information as 0 and 1. Resistive switching can generate a large number of reliable conductance states, while conducting filaments are gradually formed. As a result, it is regarded as the most favorable candidate to implement neuromorphic computing systems. On the other hand, magnetoresistance operates by spin switching, and there are theoretical limits to creating distinguishable states. Developing materials that satisfy all these requirements would accelerate the realization of memristive computing systems.

5.2. Architecture Engineering

As previously mentioned, sneak current is inherent in CBA which interrupts the operation, especially reading operation. Sneak current causes major problems, including reduction in the maximum array size, expansion of the power consumption for reliable operation, and degradation of the readout margin. To limit the sneak current, selectors should be included at each cross-point. Selectors that have nonlinearity, such as bipolar diodes, unipolar diodes, and transistors, can reduce the sneak current generated at low voltages.

With the development of CMOS based on Moore’s law, the transistor has become the most commonly adopted technology for controlling current flow. The transistor is the active element that can modulate the current density by controlling the gate bias. Hence, it is the most powerful current-modulating tool among the selectors. However, three-terminal devices are fabricated by a complex method and should be built in a large space. To achieve an ultradense 4F$^2$ cell dimension, passive selectors should be investigated.

A bipolar diode, which rectifies the reverse current, is a two-terminal device that has the potential to reduce the cell dimension. However, it can be applied only with phase-change and unipolar resistive switching, which are operated under the electrical field with unipolarity.

A unipolar diode is a passive selector characterized by nonlinearity, which rectifies the current flowing at low voltages. Because it can be integrated with all types of memristors with scalability, the unipolar diode seems to be the best candidate for the selector.

In addition to the sneak current, signal degradation is another problem, especially in memristive logic. IMP and MAGIC compute data by resistance contrast. When they are cascaded for various Boolean functions, signals are significantly decreased when they flow through memristors. For signal restoration, the memristive logic system should be designed with amplifiers. MRL is the one example that utilizes CMOSs for signal amplification and inversion.

5.3. Neural Network Algorithms

Although research is ongoing on neuromorphic technologies, a clear understanding of biological nervous systems is still required. Neural network algorithms have shown innovative
achievements in software. Nonetheless, their potential has not been fully exploited. A clear knowledge of neural networks significantly promotes the development of neuromorphic technology and accelerates its progress. Furthermore, ANNs and SNNs still face challenges that should be addressed before they can be fully adapted to new computing systems. In the human brain, the neurons are not all connected. There is a mechanism by which the neural synapses are connected when required. Fully connected ANNs are more accurate, but they are inefficient in stochastic computing systems. Moreover, the computational performance of SNNs has not been demonstrated as much as that of ANNs due to the low efficiency of algorithms in memristive systems.

Therefore, new algorithms for application with memristive computing systems should be developed to match the efficiency of biological neural networks. Recently, various algorithms, such as long short-term memory (LSTM),\textsuperscript{[222]} Hopfield neural networks (HNNs),\textsuperscript{[223–225]} and generative adversarial networks (GANs)\textsuperscript{[226]} have been introduced, which show great potential to overcome the disadvantages of conventional computing systems. Implementation of these algorithms in memristive systems is expected to open new opportunities for the adoption of novel computing systems.

6. Conclusion

The development of new computing systems based on memristors relies on contributions from various research fields. Especially, studies at the material and device levels is critical for the new building blocks that exhibit the essential characteristics in new computing systems and could provide the foundation for circuit and algorithm design. In addition, neuroscience research can provide a clearer understanding and modeling of the cranial nerves, which will further aid in paradigm shift from von Neumann architectures. We believe that this report paves the way for broader exploration of new computing systems.

Acknowledgements

I.H.I. and S.J.K. contributed equally to this work. This work was financially supported by the Future Material Discovery Program (2016M3D1A1027666, 2018M3D1A1058793) through the National Research Foundation of Korea.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

crossbar arrays, memristors, neuromorphic computing, nonvolatile memory, stateful logics

Received: May 15, 2020
Revised: June 18, 2020
Published online: August 2, 2020

[1] G. M. Moore, Electronics 1965, 38, 114.
[2] R. Dennard, F. Gaensslen, W.-N. Yu, L. Rideout, E. Bassous, A. Le Blanc, IEEE J. Solid State Circuits 1974, 9, 257.
[3] IRDS, International Roadmap for Devices and Systems (IRDS) 2018 Edition, 2018, https://irds.ieee.org/editions/2018.
[4] A. N. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, M. Kito, Y. Fukuzumi, M. Sato, Y. Nagata, Y. Matsuoka, Y. Iwata, in IEEE Symp. on VLSI Technology, IEEE, Kyoto 2007, pp. 14–15.
[5] K. Sano, Y. Hatsuda, S. Yamamoto, IEEE Trans. Parallel Distrib. Syst. 2014, 25, 695.
[6] L. Chua, IEEE Trans. Circuit Theory 1971, 18, 507.
[7] L. Chua, Semicond. Sci. Technol. 2014, 29, 104001.
[8] M. Di Ventra, Y. V. Pershin, Nat. Phys. 2013, 9, 200.
[9] D. B. Strukov, G. S. Snider, D. R. Stewart, R. S. Williams, Nature 2008, 453, 80.
[10] L. Chua, Appl. Phys. A 2011, 102, 765.
[11] J. Choi, S. Park, J. Lee, K. Hong, D. H. Kim, C. W. Moon, G. Do Park, J. Suh, J. Hwang, S. Y. Kim, H. S. Jung, N. G. Park, S. Han, K. T. Nam, H. W. Jang, Adv. Mater. 2016, 28, 6562.
[12] K. S. Li, C. Ho, M. T. Lee, M. C. Chen, C. L. Hsu, J. M. Lu, C. H. Lin, C. C. Chen, B. W. Wu, Y. F. Hou, C. Y. Lin, J. Y. Chen, Y. T. Lai, M. Y. Li, I. Yang, C. S. Wu, F. L. Yang, in 2014 Symp. on VLSI Technology Digest of Technical Papers, IEEE, Honolulu, HI 2014, pp. 1–2.
[13] B. Govoreanu, G. S. Kar, Y. J. Y. Chen, V. Paraschiv, S. Kubícek, A. Fantini, I. P. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altinimise, D. J. Wouters, J. A. Kittl, M. Jurczak, in 2011 Int. Electron Devices Meeting (IEDM), IEEE, Washington, DC 2011, pp. 31.6.1–31.6.4.
[14] A. Kawahara, R. Azuma, Y. Ikeda, K. Kawai, Y. Katoh, Y. Hayakawa, K. Tsubi, S. Yoneda, A. Himeno, K. Shimakawa, T. Takagi, T. Mikawa, K. Aono, IEEE J. Solid-State Circuits 2013, 48, 178.
[15] J. Choi, Q. Van Le, K. Hong, C. W. Moon, J. S. Han, K. C. Kwon, P. R. Cha, Y. Kwun, S. Y. Kim, H. W. Jang, ACS Appl. Mater. Interfaces 2017, 9, 30764.
[16] J. Lee, S. Cho, D. Ahn, M. Kang, S. Nam, H. K. Kang, C. Chung, IEEE Electron Device Lett. 2011, 32, 1113.
[17] F. Xiong, A. D. Liao, D. Estrada, E. Pop, Science 2011, 332, 568.
[18] T. Ohyanagi, N. Takaura, M. Tai, M. Kitamura, M. Kinoshita, A. Akita, T. Morikawa, S. Kato, M. Araida, K. Kamiya, T. Yamamoto, K. Shiraishi, in 2013 IEEE Int. Electron Devices Meeting, IEEE, Washington, DC 2013, pp. 30.5.1–30.5.4.
[19] H. Y. Cheng, W. C. Chien, M. BrightSky, Y. H. Ho, Y. Zhu, A. Ray, R. Bruce, W. Kim, C. W. Yeh, H. L. Lung, C. Lam, in 2015 IEEE Int. Electron Devices Meeting (IEDM), IEEE, Washington, DC 2015, pp. 3.3.1–3.3.4.
[20] N. Perrissin, G. Gregoire, S. Lequeux, L. Tillie, N. Strelkov, S. Auffret, L. D. Buda-Preljuban, R. C. Sousa, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altinimise, D. J. Wouters, J. A. Kittl, M. Jurczak, in 2011 Int. Electron Devices Meeting (IEDM), IEEE, Washington, DC 2011, pp. 31.6.1–31.6.4.
[21] K. T. Chen, H. Y. Chen, C. Y. Liao, G. Y. Siang, C. Lo, M. H. Liao, in 2014 Symp. on VLSI Technology Digest of Technical Papers, IEEE, Honolulu, HI 2014, pp. 1–2.
[22] B. Govoreanu, G. S. Kar, Y. J. Y. Chen, V. Paraschiv, S. Kubícek, A. Fantini, I. P. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altinimise, D. J. Wouters, J. A. Kittl, M. Jurczak, in 2011 Int. Electron Devices Meeting (IEDM), IEEE, Washington, DC 2011, pp. 31.6.1–31.6.4.
[23] A. Kawahara, R. Azuma, Y. Ikeda, K. Kawai, Y. Katoh, Y. Hayakawa, K. Tsubi, S. Yoneda, A. Himeno, K. Shimakawa, T. Takagi, T. Mikawa, K. Aono, IEEE J. Solid-State Circuits 2013, 48, 178.
[24] J. Choi, Q. Van Le, K. Hong, C. W. Moon, J. S. Han, K. C. Kwon, P. R. Cha, Y. Kwun, S. Y. Kim, H. W. Jang, ACS Appl. Mater. Interfaces 2017, 9, 30764.
[25] J. Lee, S. Cho, D. Ahn, M. Kang, S. Nam, H. K. Kang, C. Chung, IEEE Electron Device Lett. 2011, 32, 1113.
[26] F. Xiong, A. D. Liao, D. Estrada, E. Pop, Science 2011, 332, 568.
C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang, W. Song, J. P. Strachan, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, Q. Xia, Nat. Commun. 2018, 9, 2385.

In Hyuk Im has been a Ph.D. candidate at the Department of Materials Science and Engineering in Seoul National University since he received B.S. degree at the Department of Materials Science and Engineering in Seoul National University of Science and Technology, in 2019. His research aims at resistive switching device based on halide perovskites and micro-patterning process of them for integrating into crossbar array for their applications such as storage class memory and neuromorphic system.
Seung Ju Kim received his B.S. degree at the Department of Materials Science and Engineering in Seoul National University and he has been a Master’s degree candidate at the same department. His current research focuses on neuromorphic devices based on halide perovskites, both by exploring a new composition of the halide perovskites and optimized fabrication process.

Ho Won Jang is an associate professor of the Department of Materials Science and Engineering in Seoul National University. He earned his Ph.D. from the Department of Materials Science and Engineering in Pohang University of Science and Technology (POSTECH) in 2004. He worked as a research associate at University of Wisconsin–Madison from 2006 to 2009. Before joining Seoul National University in 2012, he worked in Korea Institute of Science and Technology (KIST) as a senior research scientist. His research interests include the synthesis of oxides and 2D materials and their applications for nanoelectronics, solar water splitting, and chemical sensors.