Recent Advancements in Emerging Neuromorphic Device Technologies

Jiyong Woo,* Jeong Hun Kim, Jong-Pil Im, and Seung Eon Moon

The explosive growth of data and information has motivated technological developments in computing systems that utilize them for efficiently discovering patterns and gaining relevant insights. Inspired by the structure and functions of biological synapses and neurons in the brain, neural network algorithms that can realize highly parallel computations have been implemented on conventional silicon transistor-based hardware. However, synapses composed of multiple transistors allow only binary information to be stored, and processing such digital states through complicated silicon neuron circuits makes low-power and low-latency computing difficult. Therefore, the attractiveness of the emerging memories and switches for synaptic and neuronal elements, respectively, in implementing neuromorphic systems, which are suitable for performing energy-efficient cognitive functions and recognition, is discussed herein. Based on a literature survey, recent progress concerning memories shows that novel strategies related to materials and device engineering to mitigate challenges are presented to primarily achieve nonvolatile analog synaptic characteristics. Attempts to emulate the role of the neuron in various ways using compact switches and volatile memories are also discussed. It is hoped that this review will help direct future interdisciplinary research on device, circuit, and architecture levels of neuromorphic systems.

1. Introduction

Artificial intelligence has currently become widespread and has permeated social life. Electronic devices are connected among each other, wirelessly and via other networks, and can constantly communicate. Thus, a substantial amount of data is generated every second worldwide, and the data creation period is shortening. With the unprecedented explosion in data, a new industry has been launched to extract more valuable information and utilize it beyond simply storing and managing data traffic worldwide. For example, driving skills of autonomous vehicles have advanced rapidly by recognizing information about the surrounding environment that is constantly being input to the system in real time and accurately classifying them into specific objects and signals. One of the reasons for the new wave of data-centric paradigms was the development of semiconductor technology in the past few decades. The performance and cost of transistors, a representative semiconductor device, have been improved due to Moore’s law scaling.[1] Consequently, several innovative products have been manufactured at reasonable prices, thereby creating numerous derivative industries. More specifically, increasing the number of tiny transistor elements integrated into a given silicon chip allows more versatile processing and arithmetic operations per clock cycle to be performed promptly. The memory elements based on laterally scaled and vertically stacked structures can also significantly increase memory capacity.[2,3] As we advance into the big-data era, the demand for improved performance of computing systems primarily consisting of these two fundamental components, i.e., central processing units (CPUs) and memories, to handle the exponentially growing amount of data is increasing. However, in the conventional von Neumann computing architecture, data executed at the CPU must be frequently moved back and forth to the memory for storage, which can lead to a memory wall or the von Neumann bottleneck, as shown in Figure 1.[4] Power-constrained computing systems are gaining further importance because all electronic devices should function continuously in always-connected environments. Analyses of the workload of traditional computing systems have clearly indicated that real-time applications, such as hand-tracking services and audio recognition, consume more than half of the total energy when moving and storing their data rather than performing computations.\(^\text{[5]}\) These problems have necessitated the development of new computing systems to overcome power inefficiencies by minimizing the sequential processing.

The implementation of the energy-efficient processor was based on the basic structure of the brain, which comprises biological synapses numbering on the order of \(10^{15}\) connected to neurons on the order of \(10^{11}\).[6] The data in the form of synaptic

---

Dr. J. Woo,[1] Dr. J. H. Kim, Dr. J.-P. Im, Dr. S. E. Moon
ICT Creative Research Laboratory
Electronics and Telecommunications Research Institute
Daejeon 34129, South Korea
E-mail: jiyong.woo@etri.re.kr

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202000111.

[1]Present address: School of Electrical Engineering, Kyungpook National University, 80 Daehakro, Bukgu, Daegu 41566, South Korea

© 2020 The Authors. Published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/aisy.202000111
weight ($w$) is transferred neuron-to-neuron through the synapses in parallel. When the sum of the weights in the neuron exceeds a certain threshold, the neuron responds by generating signals and passing them to other synapses. Because of the parallel-connected synaptic configuration, high-level cognitive functions in the brain can be performed by consuming only tens of watts.\cite{7}

Based on the expectation of attractive low-power benefits, understanding the brain’s structure and essential roles has initiated the development of neuromorphic algorithms through the building of artificial neural networks.\cite{8–12} The input and output neurons in each layer are linked through hidden layer neurons, which is a perceptron structure.\cite{13} The basis of the neural network algorithms is to classify specific outputs by multiplying input vector signals and synaptic weights through forward propagation.

Each neuron plays a role in linear (or binary) classification to determine whether to continuously process the signal based on the sum of the calculations. By inserting more hidden layers to perform additional perceptron processes, the multilayer neural network enables the solution of complex problems and extension of the functionalities to logical functions such as Boolean logic. Thus, such deep neural network (DNN) algorithms outperform conventional methods specifically in case of recognition and classification tasks to determine the desired output from unknown inputs.

The algorithm relies substantially on iterative arithmetic calculations such as vector–matrix multiplication (VMM), or multiply–accumulate (MAC) operation, which runs on graphics processing unit (GPUs)-based platforms\cite{9} that are appropriate for parallel processing or application of specific integrated circuits.\cite{14} The time-consuming computation is architecturally accelerated by using the cross-point array architecture in which synaptic elements are positioned between lines carrying input and output signals crossing each other.\cite{15} Neuromorphic hardware systems are simply described as having multiple synaptic arrays as weight matrix blocks, as shown in Figure 1.\cite{16} Neurons located on the edge of each array convey inputs and outputs to communicate with other segments. The voltage inputs via the word lines (WLs) in parallel reach the synapses and are subsequently multiplied by the stored synaptic weight encoded in the form of the conductance ($G$), according to Ohm’s law. Unlike normal memory operations in the cross-point array that read conductance at a single selected cell, the multiplication takes place at every cross. The weighted sum current as a result of the sum of each output along the bit line (BL), based on Kirchhoff’s current law, is fed to peripheral circuitries (e.g., analog–digital converters and multibit sense amplifiers) serving as the neuronal element.

When the output results differ from the expected values, the signal moves back to the synaptic array, and the synaptic weights are adjusted using a gradient descent method to reduce errors based on the back-propagation algorithm,\cite{8} which is the method used by the neuromorphic system to learn newly acquired information and provide accurate inferences. VMM operations are performed where the weights are physically stored, alleviating memory wall problems.\cite{17} Therefore, for the in-memory computing platform based on the cross-point array architecture,\cite{18} selecting the appropriate devices as the
fundamental building blocks for synaptic and neuronal elements is important for implementing the neuromorphic systems in hardware.

Recently, significant advances in neuromorphic hardware have been successfully reported and demonstrated. Most studies used static random-access memory (SRAM) with eight transistors arranged as the synaptic device.\textsuperscript{[19–21]} However, the SRAM with digital synaptic weights “0” and “1” is unable to satisfy the numerous parameters used in the algorithms.\textsuperscript{[21]} Although the single transistor unit has been significantly reduced to a few nanometers of technology nodes,\textsuperscript{[22]} the large footprint occupied by multiple transistors creates an area overhead. This problem has garnered significant attention to emerging memory technologies for compact and analog weight storage.\textsuperscript{[23–29]} Notably, the newly available memory options are based on resistance changes, in contrast to the conventional storing of charges in a capacitor or floating gate.\textsuperscript{[30,31]} Most resistive memories are thus essentially simple metal–insulator–metal systems, which allow the highest memory capacity in the lowest occupied cell area. The specific denotation of each resistive memory is determined by how the material systems respond to external electrical stimuli. Magnetic random access memory (MRAM)\textsuperscript{[32]} utilizes the orientation of the spin while the rotating objects become dipoles in ferroelectric memory devices.\textsuperscript{[33]} The reversible phase transition between amorphous and crystalline states in chalcogenide materials leads to a difference in resistance, known as phase change memory (PCM).\textsuperscript{[34]} Ion migration in most nonstoichiometric materials, driven locally or globally by an electric field, enables the resistance change as in resistive switching RAM (RRAM)\textsuperscript{[35,36]} or electrochemical RAM (ECRAM).\textsuperscript{[37]} State-of-the-art resistive memory technologies, excluding the ECRAM, have been integrated into \textasciitilde20 nm nodes.\textsuperscript{[38–41]} For a fair and systematic comparison, the latest SRAM is assumed to scale up to a few tens of nanometer nodes.\textsuperscript{[42]} The accelerator performances are benchmarked while considering end-to-end design options from the device- and circuit- to algorithm-level. Unlike the SRAM, the assigned multiple weights are retained even when the power supply is turned off, thereby minimizing standby leakage power.\textsuperscript{[43]} This implies that by using the resistive synapses that function optimally with the cross-point array architecture, the entire system can afford superior throughput and energy efficiency.

The neuron node adjacent to the synaptic array is often neglected in the neuromorphic system study. After the analog computation in the cross-point array, the weighted sum current at the end of each BL should be processed (e.g., converted to voltage spike or digital pulse),\textsuperscript{[44]} which is a vital role of the biological neuron that receives the current from the synapses and thereafter decides whether to activate an action potential to the next neurons in the neural network. Typically, the silicon complementary metal–oxide–semiconductor (CMOS)-based neuronal circuits comprising tens of transistors with a capacitor are used for implementing the integrate-and-fire neuron model.\textsuperscript{[45]} The weighted sum current is first integrated into the capacitor placed at the end of the BL. When the charged voltage exceeds the threshold, digitized output voltage spikes are generated through the circuitry. By counting the number of the output spikes that are designed to be proportional to the amplitude of the read-out current, the neuron node is capable of determining the output firing strength following activation functions such as sigmoid, tanh, softmax, and rectified linear unit.\textsuperscript{[46]} However, the complex neuronal circuit with a capacitor clearly occupies a substantially larger footprint than the BL pitch of the cross-point array. The pitch mismatch problem inevitably causes a single neuron node to be shared with multiple BLs, which implies that the weighted currents computed in parallel from the synaptic arrays have to be sequentially processed.

Herein, we first discuss the advances in the PCM and RRAM, where significant progress has been achieved, to address the requirements of the neuromorphic synaptic devices in Section 2. Recent strategies based on the prominent specific characteristics of other candidates such as ECRAM, ferroelectric memory, and MRAM to overcome relevant challenges have also been explored. In Section 3, we have introduced studies that explored compact neuromorphic neuronal devices based on either two-terminal switches or volatile memories, highlighting the advantages of these devices from an area and energy perspective. Finally, we have concluded the article by indicating future study based on the current status to boost neuromorphic system performances.

2. Neuromorphic Synaptic Devices

2.1. PCM

Emerging resistive memory technologies are well-developed in the order of MRAM, PCM, and RRAM from a typical memory application perspective. However, in neuromorphic applications, the PCM led to the introduction of new analog synaptic weight elements by identifying and defining new important characteristics (e.g., linearity and symmetry) as well as conventional requirements for the memory functions (e.g., endurance and retention). The resistance of the PCM depends on the crystal structure of the chalcogenide materials such as Ge\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5} (GST).\textsuperscript{[47]} In general, it is relatively easy to transmit electrons in a crystalline state, whereas the electrical conductivity is lowered when the structure is transformed to an amorphous state. The two phases can be reversibly changed by first melting the solid-state chalcogenides into a glassy state and subsequently controlling the time required for the ions to be rearranged. To effectively generate heat, a confined electrode serving as a heater is normally used to maximize the current density by reducing the region in which current flows. Applying a pulse that drives a current induces Joule heating, and the phase near the electrode begins to melt, resulting in a mushroom-shaped switching area. Instant cutting off of the pulse satisfies the glassy state of the chalcogenide. It results in a significantly disordered amorphous state, showing a high resistance state (HRS), known as a reset process. Meanwhile, when sufficient time to relocate the ions to a thermodynamically stable position is provided during the molten state, the crystalline state can be formed to obtain a low resistance state (LRS), known as a set process. The analog behavior in the PCM was observed by subdividing and fine-tuning the intermediate pathways that changed from the HRS to LRS, or vice versa. It was possible to experimentally achieve a distinguishable 3-bit state corresponding to the synaptic weight precision by elaborately adjusting the switching current directly related to the volume of the phase transition.\textsuperscript{[48]}

Adv. Intell. Syst. 2020, 2, 2000111

© 2020 The Authors. Published by Wiley-VCH GmbH
Two important stages are performed in the neuromorphic systems implemented with the cross-point PCM synaptic arrays, as shown in Figure 2a,b.[49] In the inference phase, weights predefined from the software or external cloud servers, which is a training (or learning) process, are assigned to each PCM device and mapped to the array to extract the correct value according to input patterns after the VMM execution. The capability of the multiple weights in the PCM allows more numerous and complicated input patterns to be recognized accurately. The accuracy and robustness of the inference are thus related to the state-stability of each state. However, despite the exclusion of the disturbance contributed by accumulative stress induced by the repeated input voltage, the states in the PCM were drifted to the HRS over time due to structural relaxation of the amorphous phase,[48] making it difficult to ensure each state with a reasonable margin of error. To improve state-stability, an additional metallic liner was introduced to mitigate the drift. Consequently, nearly negligible drift and noise reduction were achieved.[50]

In addition to inference accelerators, where the system recognizes and categorizes provided information, there is a demand for the systems to respond in real-time to unknown trends. Because the power consumption is mostly hindered by data movement, the training should be performed within the hardware itself. In the training phase, the synaptic weight within the provided dynamic range of the multilevel states is updated and plays an important role in achieving high recognition accuracy for the training and learning of new information. The weights should be modulated linearly and symmetrically. The multiple conductance states were achieved using an identical pulse for potentiation, while the pulse amplitude needed to be increased for depression. Reproduced with permission.[48] Copyright 2018, AIP Publishing. Therefore, for hardware implementations, the weight is encoded by a pair of two PCM devices. The PCM device for $G^+$ is used to increase the conductance, whereas the other PCM device for $G^-$ is intended to lower the conductance. In other words, the input vectors ($x_i$) in the form of voltage are applied and multiplied by the weight in the form of conductance ($G$) assigned to each PCM synapse. Then, actual weight is represented by subtracting $G^-$ from $G^+$. Reproduced with permission.[49] Copyright 2015, IEEE.

Figure 2. Neuromorphic systems are generally based on two important stages: a) inference and b) training based on forward and backward propagation algorithms, respectively. The input signal vectors ($x^i$) of the input neurons drive analog weights ($w$) to next hidden neurons. The simple sum of each weight multiplied by the vector is performed on the neurons. When the output signal vectors ($x^e$) differ from the expected values ($g$), the signals go back to the synapses and adjust the weights to reduce the error term ($\delta$). c) Here, how the weights are updated plays an important role in achieving high recognition accuracy for the training and learning of new information. The weights should be modulated linearly and symmetrically. d) The multiple conductance states were achieved using an identical pulse for potentiation, while the pulse amplitude needed to be increased for depression. Reproduced with permission.[48] Copyright 2018, AIP Publishing. e) Therefore, for hardware implementations, the weight is encoded by a pair of two PCM devices. The PCM device for $G^+$ is used to increase the conductance, whereas the other PCM device for $G^-$ is intended to lower the conductance. In other words, the input vectors ($x_i$) in the form of voltage are applied and multiplied by the weight in the form of conductance ($G$) assigned to each PCM synapse. Then, actual weight is represented by subtracting $G^-$ from $G^+$. Reproduced with permission.[49] Copyright 2015, IEEE.
system. Moreover, the amount of increase or decrease in any given state of the PCM should be similar because the state of the PCM does not consistently change in a similar direction in the systems. However, due to the weak linearity and symmetry of the PCM, training cannot be effectively conducted.

One of the approaches used to overcome the asymmetric response of the PCM was to only use a potentiation regime that exhibits analogous conductance by periodically resetting (or refreshing) all information to its original state.\[51\] For this technique, a pair of two PCM elements for positive conductance ($G^+$) and negative conductance ($G^-$) comprise a single synaptic device to encode actual weight ($w = G^+ - G^-$) and also represent its negative value (Figure 2e). The weight was increased to a target value by a single step of applying the identical pulses. To lower the weight, depression was performed using a two-step method in which the both PCMs were reset to the initial state. Thereafter, one of the PCMs in the pair, which is responsible for the positive conductance, was only activated again by the pulses while the other PCM representing negative conductance maintained its state. A multilayer perceptron comprising $500 \times 661$ PCM arrays using the technique has been experimentally implemented.\[49\] A recognition accuracy of $\approx 82\%$ was achieved for Modified National Institute of Standards and Technology (MNIST) dataset; however, it was lower than the expected level of $97\%$ due to imperfect PCM device characteristics.

2.2. RRAM

The relevant findings, as detailed in previous sections, regarding the device guidelines for an analog synapse using the PCMs have highlighted the potential of RRAMs. In case of the RRAMs, which generally represent devices that use oxygen vacancies (or oxygen ions) as mobile species, oxygen vacancies are created by breaking the bonds between metal and oxygen either at the bulk oxide or interface.\[55,56\] Alternatively, cations are supplied from electrodes such as Cu or Ag outside the materials, which is known as conductive-bridge RAM (CBRAM).\[52\] Whether the mobile species are anions or cations, the ions driven by the applied set field are clustered, eventually bridging the two separate electrodes. Instantaneously, high current can thus be observed in the RRAM through the formation of a conductive filament. Meanwhile, as the opposite reset voltage spreads the oxygen vacancies from the filament, the filament starts to dissolve through an electrochemical reaction. The current flow is limited as the filament is disconnected. In general, a compliance current that limits excess current over a preset value is applied to the RRAM to prevent permanent breakdown. The magnitude of the compliance current directly determines the amount of current flowing through the RRAM, which implies that the size of the filament is provided. As the filament thickens by increasing the compliance currents, a lower LRS is continuously achieved. In contrast, the higher negative voltage removes more oxygen vacancies from the filament, thereby forming a switching gap between the electrode and the remaining filament. The extended gap can have multiple HRS in the lower direction.

Through using a cross-point array with only a single RRAM\[53,54\] or one-transistor and one-resistor (1T–1R) configuration,\[55–59\] diverse classification and recognition features and functions have been explored and demonstrated experimentally. A two-layer perceptron has been constructed by the building of $128 \times 64$ Ta/HfO$_2$/Pt (from top to bottom) based IT–IR arrays.\[55\] The conductance toward a higher level was precisely tuned by the gate voltage of the monolithically integrated transistor, as shown in Figure 3a. Due to the use of the two pairs of the RRAM as the single synaptic element discussed in case of the PCM, the conductance in the lower direction was achieved by first applying the reset pulse to initialize the state, and the gate voltage was thereafter increased. The tunable linear and symmetric update of the conductance with minimal variation allowed the hardware neural network to be trained properly, experimentally achieving an accuracy of 91.71% of the MNIST dataset.

Although the inference task has been successfully demonstrated using the well-trained analogous conductance states, the neuromorphic hardware system can further be made to be energy efficient by making a device environment, where the weight update can be driven by an identical pulse scheme,\[60\] as discussed earlier. As identical pulses are successively addressed to the HfO$_2$-based RRAM, the asymmetric response of the conductance due to nonlinear potentiation was observed,\[56\] which was an exactly opposite property of the PCM, as shown in Figure 3b. Once abrupt conductance jumped at the initial pulse due to the formation of the filament, no more conductance increase was observed in the potentiation. The conductance was adjusted by the number of negative pulses and the slope of the decrease in conductance was determined by the amplitude and width of the pulse. A microscopic physical description of the RRAM that investigated the link between the filament evolution and the electrical behavior revealed that the formation of a strong filament caused the binary state during the potentiation.\[61\] In contrast, it was discovered that an alternative scenario, where the radial size of the filament is changed, is preferred to have a linear current response. The first attempt was to engineer the filament dynamics from the next cycle as desired because an abruptly grown filament in the initial state was difficult to control in working principle. Introducing an additional barrier layer of AlO$_x$ featuring a slower oxygen mobility compared with that of the HfO$_2$ caused the dissolution process of the filament during the reset to be retarded.\[56\] It resulted in an incompletely disconnected filament. In the subsequent set cycle, the weakest constriction part of the filament, where the bilayer was contacted, was to be a plausible switching region by moving back-and-forth in the vacancy while the filament was still connected between two electrodes. Instead of growing in a vertical direction, the lateral expansion of the filament was discovered to be facilitated to update the conductance linearly, depending on the identical pulses. Other methodologies to manage the generation and migration of the oxygen vacancies in the initial stages, prior to these vacancies making the strong filament, have been proposed. By using a thermal barrier of TaO$_x$ with low thermal conductivity, the heat that is produced during the device operation can be confined into the HfO$_2$ layer.\[62\] The heat spreads the distribution of the vacancies extensively while the vacancies were electrically driven to form the filament as is normal. The laterally expanded filament shape seemingly enabled the analog set transition in the $I-V$ curve and pulse switching. In addition, to exploit the temperature as another kinetic terminology in ion transport, formation
energy of the vacancies was reduced to lower the probability of generating the vacancies using an electric field.\textsuperscript{[63]} It was realized that by incorporating dopants into the HfO\textsubscript{2} matrix, bonding strength was reduced. The uniformly distributed dopants facilitated the broad making of multiple filaments, resulting in analogously updated behaviors in both polarities. Even at a high temperature, the multiple states were distinguishable, ensuring the information at the peripheral sensing circuit was accurate. Thus, hardware systems with eight processing blocks comprising 128/16 TaO\textsubscript{x}/HfO\textsubscript{2}-based 1T–1R analog synapse arrays were successfully integrated to implement a five-layer convolutional neural network to perform MNIST image recognition.\textsuperscript{[58]} The clear distribution of 1024 RRAM devices in 5-bit state within the conductance range of 100–900 μS without any overlap was also achieved by an identical pulse train with a substantially fast speed of tens of ns in the obtained conductance range of the RRAM with TaO\textsubscript{x}/HfO\textsubscript{2} stack was highly reliable. Reproduced with permission.\textsuperscript{[58]} Copyright 2020, Springer Nature.

Similar hardware performance was also verified through mass-produced Ta\textsubscript{2}O\textsubscript{5}/TaO\textsubscript{x}-based 1T–1R synapses.\textsuperscript{[59]} The uniform analog states linearly tuned from 20 to 50 μA with a verification technique that allowed a maximum recognition accuracy of 90% on the MNIST database. The 180 nm Ta\textsubscript{2}O\textsubscript{5}/TaO\textsubscript{x}-based RRAM exhibited a similar number of synapses per unit area when compared with a 65 nm SRAM. However, due to the reduced operational power of the RRAM device, the efficiency in performance and acceleration inference was more than three times, which was sufficiently high to enable a real-time recognition service. Furthermore, due to the local filamentary switching, the RRAM was scaled in a 40 nm test-chip, the efficiency running the neural network workloads can be further improved. In Table 1, reported array-level RRAM-based synapses were compared to identify the normal range of the conductance states and the pulse conditions that were typically required to control the states in most of the HfO\textsubscript{2} device stacks.

Considering the operating power, the maximum conductance of the synaptic device is one of the key governing factors in determining and boosting the neuromorphic hardware performance. This is because if the conductance is significantly high, the size of the transistor of the 1T–1R and peripheral circuit (e.g., multiplexer) should be increased to avoid voltage drop.\textsuperscript{[42]} Significant area overhead occurs and the systems operate slowly, resulting in longer latency and reduced throughput. Accordingly, the noticeable advantage of the RRAM over the PCM is a lower operating current due to non-Joule heating-related switching mechanism, implying synaptic weights in a lower conductance range. However, in practice, non-negligible parasitic components such as line resistance are involved in the cross-point array.\textsuperscript{[66]} The voltage drop due to the line resistance is spontaneously increased when the feature size of the interconnect line is scaled. In the column of the array nearest to the voltage source, most of the
applied read voltages are delivered properly to the synaptic devices without any noticeable loss, accurately executing the multiplication. However, the read voltage decays along the line, and the voltage is significantly lowered in the farthest cell. The weighted sum current is thus lower than expected because the lowered read voltage is multiplied even though the given weight remains unchanged. It has been reported that the operating current of the RRAM can be reduced to \( \approx 1 \) mA.\(^{67}\) Note that the low current operation in the RRAM indicates that the filament weakly comprises a lesser number of vacancies and no longer ensures metallic behavior exhibited by the stronger filament clustered from denser vacancies. Consequently, in the current–voltage (\( I-V \)) curve, the current at the LRS started to get distorted non-linearly with respect to the voltage. It caused the conductance measured at the reduced read voltage to be lowered exponentially, and deviation of the actual computed weighted sum results became pronounceable. Therefore, studies have been conducted to carefully design electrode materials that can modify the conical shape of the filament to dissipate heat appropriately\(^{68}\) or to compensate the nonlinearity with circuits.\(^{69}\) Strengthening the \( I-V \) trace of the RRAM linearly allows the achievement of constant conductance, which can be less affected by the voltage drop.

Nonideal factors such as nonlinearity, asymmetry, and limited conductance range have been intensively studied in device and system-level analysis,\(^{70}\) but reliability concerns such as data retention, cycling endurance, variability, and failure have been less discussed and explored.\(^{71,72}\) The conductance states can be affected in unexpected ways by various reliability issues. For simplicity, the conductance degradation trends were categorized in two major ways by considering whether the weighted sum current was consistently changed toward a certain direction or not. When the external parasitic components such as line resistance or conductance drift were considered at a specific BL, the output current was always changed uniformly due to the lowered input voltage or changed conductance with respect to the time, causing accuracy deterioration. Due to the consideration of the variation of the RRAM as a true stochastic behavior due to an inherent working principle, each weight could either be lower or higher than the criterion (e.g., the median). Thus, the lowest weight was compensated by the highest weight at the selected BL. This result indicated that the total weighted sum at the end of the BL was near the expected value. This explains the reduced effect of the nonuniformity of the individual devices on the accuracy of the recognition. The non-uniformity can introduce advantages that can help overcome the challenge depending on neural networks.\(^{73}\) Learning with a gradient descent scheme allows finding the optimum value defined by the global minimum; however, the learning process can converge to the local minimum level and be stuck while finding a route. The variation in the weights caused by non-uniformity can act as the driving force to escape the minimum area.

To minimize spatial and temporal variations in the filamentary RRAMs that affect accuracy, a novel material engineering technique was proposed. Instead of filaments formed randomly during device operation, dislocations in the material were deliberately threaded to confine the path.\(^{74}\) Thus, ions preferred to travel through the 1D channels, significantly improving uniformity. In the early stage of the RRAM-based synaptic element, an electrical barrier, such as Schottky barrier modulation, which is smoothly controlled by the movement of ions in the entire active area along the electric field, was used as a more uniform switching mechanism.\(^{64,65,75}\) The gradual conductance update that is proportional to the number of identical pulses was available, but the conductance increase (or decrease) was substantial at the very first step from the initial HRS (or LRS), respectively. It resulted in a highly asymmetric conductance response versus the pulse number. A slow speed of a few ms to drive the ions over the entire area was also another critical problem (see Table 1). Therefore, the exploration on the interfacial mechanism has been rarely studied currently in two-terminal device structures, but a similar ion movement mechanism has regained substantial attention and expanded by using a three-terminal structure and new materials, as will be discussed below.

### 2.3. ECRM

The ECRM, which was only designed for ultimate linear and symmetric synaptic characteristics, has been proposed due to the need to improve the controllability of ion transport. By using the traditional three-terminal transistor structure, the channel

---

**Table 1. Considerations for the synaptic characteristics based on array-level RRAMs.**

| Configuration | Device stack | Array size | Dynamic range of conductance | Pulse condition for potentiation/depression\(^{3} \) | Ref. |
|---------------|--------------|------------|-----------------------------|---------------------------------|-----|
| 1T–1R         | Ta/HfO\(_2\)/Pt | 128 × 64   | 100–1100 μS                 | V\(_{E}\) from 0.6 V to 1.6 V (500 μS) | [55] |
| 1T–1R         | TiN/HfO\(_2\)/AlO\(_2\)/TiN | ≈K        | 20–60 μS                   | V\(_{E}\) = 1.2 V, V\(_{SL}\) = 0.9 V (100 μS) | [56] |
| 1T–1R         | TiN/TaO\(_2\)/HfAlO\(_2\)/TaN | 128 × 8    | 10–70 μS                   | V\(_{E}\) = 3.5 V, V\(_{BL}\) = 1.8 V (50 ns)/V\(_{E}\) = 8 V, V\(_{SL}\) = 1.9 V (50 ns) | [57] |
| 1T–1R         | TiN/TaO\(_2\)/HfO\(_2\)/TiN | 128 × 16   | 2–20 μS                    | V\(_{E}\) = 1.8 V, V\(_{BL}\) = 2 V (50 ns)/V\(_{E}\) = 4.7 V, V\(_{SL}\) = 1.8 V (50 ns) | [58] |
| 1T–1R         | Ir/TaO\(_2\)/TaO\(_2\)/TaN | 4 M         | 20–50 μA\(^{6}\)           | Varying V\(_{E}\)                 | [59] |
| 1R            | Pt/Ti/AlO\(_2\)/AlO\(_2\)/Pt | 20 × 20    | 10–500 μS                  | Tuning algorithm               | [53] |
| 1R            | Pt/HfO\(_2\)/Au | 54 × 108   | 2.17–4.67 μS              | 1.8 V (82 μs)/–1.8 V (82 μs)   | [54] |
| 1R            | Pt/AIO\(_2\)/TiN/PCMO/Pt | 32 × 32    | 7–50 ns                    | 3 V (1 ms)/3 V (1 ms)          | [64] |
| 3D 1R         | Ta/TaO\(_2\)/TiO\(_2\)/Ti | 4 × 4       | 80–130 ns                 | 3 V (5 ms)/3 V (5 ms)          | [65] |

\(^{6}\)The value was represented in the form of current; \(^{3}\)In the 1T–1R configuration, bit-line voltage (V\(_{SL}\)) was applied for the potentiation, whereas source-line voltage (V\(_{SL}\)) was used for the depression. The state was tuned by adjusting the gate voltage (V\(_{G}\)).
conductance between the source and drain is precisely tuned by the number of mobile ions provided by the gate. Motivated by the principle of a solid-state ion battery in which mobile Li ions stored at the cathode are transported to the anode, a channel material of LiCoO$_2$, which is capable of providing Li ions due to weak bonding, was used as the channel material as shown in Figure 4a. To promote effective migration of the Li ions, LiPON material was used as an electrolyte. When a negative voltage was applied to the gate and source (V$_{GS}$), the intercalated Li ions in the LiCoO$_2$ channel were pulled to the gate, which was a write operation. In the empty position where the Li ion was released from the LiCoO$_2$, the valence of Co ion was changed from 3$^+$ to 4$^+$ to maintain charge balance, generating positive charge. When the n-type oxide semiconductor MoO$_3$ was used as the channel, due to the formation of the electrons at the positions where Li ions were removed, an increase in conductance was observed by applying a positive gate voltage. The read path was decoupled with the write operation by applying voltage to the drain and grounding it to the source with a zero V$_{GS}$ signal source. The current (I$_{SD}$) flowing between the source and drain separated by a long channel distance of 2 $\mu$m can thus be analogously adjusted by the proportionally modulated quantity of the Li ions moved under the number of applied V$_{GS}$ signals.

As shown in Figure 4a, the conductance continued to increase when V$_{GS}$ was simultaneously provided. A steady and constant current was observed when the gate voltage was removed to identify the altered channel state. In general, the changed conductance lasted for several weeks, and it was expected to continue to last for several months.

Other mobile cations such as H ion that emulates the role of the Li ions have also been examined, as shown in Table 2. Unlike the Li ions embedded in the host material, the gate voltage pushed the cations (e.g., H ion) toward the bottom of the electrolyte of WO$_3$ while attracting the electrons to the top of the channel. It was discovered that the film quality and physical properties of each layer played a crucial role in determining the dynamic range of conductance. Recently, a fully CMOS compatible ECRAM device was reported by exploiting fab-friendly oxygen anions and metal oxides as the mobile source and electrolyte/channel, respectively, as shown in Figure 4b. The ECRAM satisfied the requirements of the basic synaptic characteristics, and it was also experimentally demonstrated in small-sized arrays.

However, the achievable conductance range and operating conditions such as speed and voltage seemed to be strongly and sensitively correlated to the materials and geometry of each layer. Therefore, it is important to consider and design a material...
aimed at specific applications such as defining the required array size and implementing appropriate drive circuitries. Moreover, similar to the challenge of the interface-type RRAM, the ECRAM required a long pulse to drive ions through the entire area. Although the operation was demonstrated in less than 10 ns, the tunable conductance range became very short as a result of a trade-off relation.

2.4. Ferroelectric Memory/MRAM

When the device structure is not limited to the compact two-terminal structure, it is expected that highly uniform and reliable synaptic characteristics will be achieved by exploring domain switching dynamics in ferroelectric (or magnetic) materials instead of the ion migration that accompanies the inevitable switching dynamics in ferroelectric (or magnetic) materials. When the device structure is not limited to the compact two-terminal structure, it is expected that highly uniform and reliable synaptic characteristics will be achieved by exploring domain switching dynamics in ferroelectric (or magnetic) materials instead of the ion migration that accompanies the inevitable switching dynamics in ferroelectric (or magnetic) materials. instead of the ion migration that accompanies the inevitable switching dynamics in ferroelectric (or magnetic) materials. Instead of the ion migration that accompanies the inevitable switching dynamics in ferroelectric (or magnetic) materials.

The FeFET operation is similar to that of the FLASH memory. Applying positive gate voltage ($V_g$) not only induces the channel inversion in the p-type silicon substrate as is normal but also causes spontaneous polarization in the ferroelectric gate dielectric that promotes the accumulation of electrons. Due to the ease of supply of sufficient electrons, the condition of creating an inversion layer is satisfied at a lower threshold voltage ($V_{th}$) than that expected in the nFET. Meanwhile, the negative gate voltage switches the direction of the dipoles in the doped HfO$_2$ and negatively polarized charges induced near the channel, pushing electrons away. As the $V_{th}$ is shifted in the positive direction, the memory window in $V_{th}$ is exhibited. As the polycrystalline-doped HfO$_2$ comprised multiple ferroelectric domains, it was possible to be partially polarized, enabling fine-tuned threshold voltages, as shown in Figure 5a.

Consequently, continuous channel conductance can be extracted from diverse traces of $I_{ds}$ vs $V_{th}$ of the FeFET. For the FeFET-based synapse, three available pulse schemes were evaluated (Figure 5b). The identical pulse showed the gradual potentiation, whereas several states were only achieved due to a significant drop in the depression. By modulating the pulse width, the non-linear response in the depression was improved. This is because the long pulse sufficiently converts the dipoles in the domain. Instead, the amplitude modulation scheme at a given pulse width of 50 ns increased the amount of the switched domains each time the pulse was applied, exhibiting the highest states (5-bit) and its symmetry. Due to the uniform synaptic behavior operated at a high speed, the neuromorphic system’s performance indirectly verified by circuit-level macro simulators was discovered to have better accuracy of $\approx 90\%$ on the MNIST and faster latency than other emerging memory-based synaptic candidates.

It has also been presented that the ferroelectric layer was implemented into more advanced transistor structures such as FinFET and nanowire FET (Figure 5c). Interestingly, both scaled FeFET-based synaptic devices seemed to have analogous conductance controlled by the identical pulse train. However, the linearity deviated from the ideal trajectory, which caused a reduced accuracy of $\approx 80\%$ compared with that of the planar FeFET-based synapse (see Table 3).

Table 2. Comparison of characteristics of ECRAM-based synapses reported to date.

| Mobile ions | Electrolyte | Channel | Linearity (potentiation/depression) | Dynamic range of conductance | Device size (length × width) | Write speed | Ref. |
|-------------|-------------|---------|-----------------------------------|-------------------------------|-----------------------------|-------------|------|
| Li          | LiPO$_4$    | LiCoO$_2$ | 0.1                               | 150–250 µS                   | 2 s                         | [37]        |
| Li          | LiClO$_4$/PEO | MoO$_3$  | 0.31                              | 42–75 nS                     | 10 µs                       | [78]        |
| Li          | LiPO$_4$    | WO$_3$   | 0.34/0.268                        | 1–3 nS                       | 80 × 100 µm$^2$ (smallest size of 0.3 × 0.3 µm$^2$) | 1 s (down to 5 ns) | [79] |
| Li          | LiClO$_4$ dissolved in polyethylene oxide | LiTiO$_2$ | –       | 40–80 µS                                   | 10 µS                       | 10 ms       | [80] |
| Li          | Li$_2$PO$_4$ | LiCoO$_2$ | 1.33/–0.34                       | 2.5–40 nS                    | 20 × 50 µm$^2$              | 1 s         | [81] |
| H           | H: SiO$_2$  | WO$_3$   | 0.44/–1.21                        | 0.3–0.9 µS                   | –                          | 1 s         | [82] |
| H           | PEDOT: PSS  | Nafion   | 50–100 nS                         | 125 × 45 µm$^2$              | 50 µS (down to 200 ns)      | [83]        |
| Oxygen vacancy | Metal oxide/HfO$_2$ | WO$_3$   | 0.6/–0.58                        | 20 × 80 µm$^2$               | 0.5 s (down to < 10 ns)     | [77]        |
The ferroelectric capacitor alternatively denoted as ferroelectric tunnel junction (FTJ) was also used as a stand-alone memory.[92,93,96] Unlike the FeFET-based synapse, where the conductance from the source to drain is adjusted by polarization change occurring in the ferroelectric oxide between the gate and silicon substrate, conductance through the FTJ is directly affected by up or down direction of the dipole. The conductance of the FTJ is transmitted smoothly by the identical pulse. Due to the simple structure, the FTJ-based synapse can be integrated into a 3D vertical NAND structure, where the FTJ is formed on the sidewall (Figure 5d).[92] Inherently low conductance range from 1 to 3 nS of the usual FTJ can slow down the system during the read operation. However, the neuromorphic systems usually sense the weighted sum of the multiple FTJs. Thus, the weight mapping and array size must be carefully designed to calculate the proper output current that does not affect the speed read by peripheral circuitries.

Meanwhile, MRAM utilizes the orientation of the spin rotated by the direct voltage or magnetic field of magnetic metal electrodes placed on either side of a thin tunneling oxide, which is a magnetic tunnel junction (MTJ) structure. Due to the achievement of only two HRS and LRS in the MRAM driven by spin-transfer torque (STT), implementation is expected for limited neuromorphic systems that routinely perform inference on the small-sized input data by adopting binary neural network (BNN) algorithms, where the weights were quantized and binary.[97,98] The digital state can be further extended while the multiple MTJs are stacked.[99] Recently, a new writing
Therefore, to compensate the imperfect synaptic devices, of the hardware level beyond unit device improvement.

To date, several studies have attempted to improve the linear conductance values of the higher significance conductance (HSC) pair, and $g^+$ and $g^−$ represent newly introduced conductance values of the lower significance conductance (LSC) pair. It has been recently discovered that the use of capacitors can result in significantly linear conductance update (Figure 6b). More specifically, the capacitor based synaptic configuration comprised three parts: 1) a readout FET connected to the capacitor, 2) a p-type FET (pFET), and 3) an n-type FET (nFET) for adding and subtracting charge to the capacitor, indicated in 3T–1C configuration. The charge on the capacitor represented the synaptic weight, and it was elaborately varied by the gate voltage for charging and discharging to the capacitor node. However, intrinsic volatile

2.5. Novel Hybrid Synaptic Configuration

To date, several studies have attempted to improve the linear response of the conductance as a function of voltage and identical pulse train in the analog emerging memories for weighted sum and weight update operations, respectively. The absolutely small dynamic conductance step adversely becomes the most significant problem to have considerable effect on the accuracy of the hardware level beyond unit device improvement. Therefore, to compensate the imperfect synaptic devices, a hybrid synaptic configuration has been proposed, as shown in Figure 6a. The purpose of the configuration is to subdivide the role in the training, thus relaxing stringent demands to be satisfied by a single synaptic device. Depending on the numerical importance in the neuromorphic systems, two pairs of conductance elements were newly configured to be a single synaptic element as follows

$$w = F(G^+ - G^-) + (g^+ - g^-)$$

where $F$ defines a significant factor that indicates the numerical significance of the weight, $G^+$ and $G^−$ denote the normal conductance values of the higher significant conductance (HSC) pair, and $g^+$ and $g^−$ represent newly introduced conductance values of the lower significance conductance (LSC) pair. It has been recently discovered that the use of capacitors can result in significantly linear conductance update (Figure 6b). More specifically, the capacitor based synaptic configuration comprised three parts: 1) a readout FET connected to the capacitor, 2) a p-type FET (pFET), and 3) an n-type FET (nFET) for adding and subtracting charge to the capacitor, indicated in 3T–1C configuration. The charge on the capacitor represented the synaptic weight, and it was elaborately varied by the gate voltage for charging and discharging to the capacitor node. However, intrinsic volatile

### Table 3. Comparison of device structures and synaptic parameters for ferroelectric memory-based synapses.

| Configuration       | Materials                  | Dynamic range of conductance | Device size (length/width) | Pulse condition for potentiation/depression | Linearity (potentiation/depression) | Ref. |
|---------------------|----------------------------|------------------------------|---------------------------|---------------------------------------------|-----------------------------------|-----|
| Planar FeFET        | Si:HfO₂/SiON               | 0.4−1.2 V                    | 500 nm/500 nm             | Nonidentical pulse + 3 to −5 V (1 μs)/      | −2 to −4 V (1 μs)                 | [94]|
| Planar FeFET        | HfZrO/SiO₂                 | 0−65 μS                      | 0.6 μm/20 μm              | Nonidentical pulse + 2.85 to +4.5 V (75 ns)/ | 1.75/1.46                        | [88]|
| Planar FeFET        | HfZrO/InGaZnO              | 0.01−0.2 μS                  | 300 μm/50 μm              | Nonidentical pulse +2.7 to +4.3 V (10 ms)/   | −2 to −3.6 V (10 ms)             | [95]|
| Junctionless n-shaped FeFET | HfZrO/SiO₂          | 9−45 nS                      | 120 nm/50 nm              | Identical pulse +3.7 V (100 μs)/            | −1.2 V (100 μs)                  | [90]|
| Nanowire FeFET      | Al₂O₃/HfZrO/GeOₓ          | −                          | 105 nm/32 nm              | Identical pulse 5 V (50 ns)/−5 V (50 ns)    | 1.22/−1.75                      | [91]|
| FTJ                 | HfO₂/SiO₂                  | 1−3 nS                       | −                         | −                                            | −                                 | [92]|
| FTJ                 | HfZrO                      | 15−40 μS                     | −                         | Identical pulse 3 V (100 ns)/−3 V (100 ns)  | −                                 | [93]|
| FTJ                 | HfZrO/Al₂O₃                | 0.25−1.5 nS                  | −                         | Nonidentical pulse +4.5 to +8.85 V (10 μs)/ | −4 to −8.35 V (10 μs)            | [96]|

The value was shown in the form of threshold voltage.

| Writing mechanism | Device stack                          | State      | Dynamic range of conductance | Ref. |
|-------------------|---------------------------------------|------------|------------------------------|-----|
| STT               | CoFeB/MgO/reference layer             | Binary     | 0.33−0.53 mS                 | [98]|
| STT               | Multiple MTJs (CoFeB/MgO/reference layer) | Analog | 25−90 kΩ                  | [99]|
| SOT               | CoFeB/MgO/CoFeB/W                     | Binary     | −                           | [100]|
| SOT               | Ta/Pr/Co/Ta/Co/Al₂O₃                 | Analog     | −60 to 20 mΩ             | [101]|
| SOT               | Ta/MgO/Co/[Co/Ni]/PtMn/Pt/Ta         | Analog     | −0.2 to 0.2 Ω             | [102]|

The value was shown in the form of resistance; The value was shown in the form of Hall resistance.

Table 4. Comparison of synaptic behaviors for magnetic memory-based synapses.

*© 2020 The Authors. Published by Wiley-VCH GmbH*
properties of the capacitor, which take natural decay of the charge, should be refreshed periodically. Through exploiting the benefit of the linearly updated synaptic weight even for a short duration, the volatile component was defined to be LSC \((g^+ \text{ and } g^-)\). In other words, during training, only the LSC pair was updated linearly and bidirectionally. The trained weights

Figure 6. a) A pair of two nonvolatile memories and two volatile capacitors for representing synaptic weight. Reproduced with permission.\(^{[103]}\) Copyright 2018, AIP Publishing. b) An important step in this hybrid configuration is to realize a perfectly linear conductance update through the capacitor and then transfer the updated conductance to the nonvolatile components. Reproduced with permission.\(^{[104]}\) Copyright 2018, IEEE. c) This resulted in obtaining the hardware accuracy equivalent to the software when recognizing from relatively simple MNIST database to a more complex CIFAR-10/100. Reproduced with permission.\(^{[105]}\) Copyright 2018, Springer Nature.
were thereafter transferred to the nonvolatile PCM devices serving as HSC (G− and G+), so that the weights could be stored for a long time. Consequently, the 3T–1C and 2 PCMs represented the weight. This approach enabled software-comparable hardware performance with accuracy of ≈98% and 88% for the MNIST and Canadian Institute for Advanced Research (CIFAR)-10, respectively, as shown in Figure 6c.

As an extended concept, the role of the volatile capacitor component was replaced by a nonvolatile FeFET device, thereby saving the area and power substantially.[106] The pFET and nFET were used for a similar purpose of providing and distracting charges, but the degree of the charge accumulation proportional to the gate automatically affected the polarization in the FeFET. Gate voltage was applied to update the weight induced through the polarization; thus, the update was automatically encoded at the FeFET. This simplified process in the two-source transistors and one-FeFET (2T–1F) can eliminate leakage concern due to the nonvolatility and minimal device area occupied by the capacitors and 2 PCMs. These hybrid synapses were expected to exhibit better training accuracy at the expense of the area.

3. Neuromorphic Neuronal Device

The cross-point array of densely arranged analog synapses used for producing VMM results for the inputs discussed in this article represents one of the layers of the neural networks. Identification of the outcome and communications between the arrays is typically conducted via the silicon-based CMOS neuron circuit by converting the weighted sum, in mechanism.[107] Various binary, ternary, and quaternary chalcogen materials for the volatile memory [113] or by strengthening the selector devices vulnerable to the external stress using glassy feature,[111] the output spike generation adjusted by the spacing and amplitude of the input signals was demonstrated. In addition to the unit neuronal element, a prototype of fully integrated emerging devices based on neuromorphic systems showing the interactions between the nonvolatile RRAM-based synapses and the volatile RRAM-based neurons with capacitors were demonstrated experimentally (Figure 7b). To perform an inference task on letter patterns, the synaptic weights were pre-encoded in the 1T–1R device with the Pd/HfO2/Ta structure as we discussed earlier (Figure 3a). For each pixel of the pattern, different amplitudes of the input voltage were given and fed into the synapse array. The VMM results were concurrently filtered, activating corresponding neuron properties.

3.1.2. Capacitor-Less Neuron Design Exploiting Selector

For accumulation that dynamically tracks history of the addressed input signals, the selector-based neuron seems to inherently require the capacitor. Attempts have been made to get rid of the capacitor, and the idea here is to deliberately make the selector devices vulnerable to the external stress using glassy materials for the volatile memory [113] or by strengthening the Joule heating mechanism on the VO2-based selector.[114] Even when a voltage below \( V_{th} \) was applied to the selector, the input pulses were stimulated to migrate the ions to form the filament in the volatile memory or induce the phase transition in the selector. This continued to steadily lower the \( R_{on} \) and eventually turned on the selector, which implied that a single selector could emulate both integrate and fire behaviors. The degree of sensitivity of the accumulation of the damage under stress in the selector with weak immunity determines the integration and timing of fire in this capacitor-less neuron design.

Meanwhile, the progressive evolution of the HRS and its reach to the LRS in the nonvolatile PCM (Figure 8a)[115] and FeFET (Figure 8b)[116] observed during the potentiation have achieved integration and fire functions. However, at the cost of the nonvolatility of the memories, the reset process to restore the initial state corresponding to the HRS for the next neuronal function should be processed periodically with additional circuitry. Accordingly, the MRAM has been proposed as an alternative.[117] The binary resistance of the MRAM was reversibly changed through spin-torque transfer. However, during normal operation, the LRS unexpectedly returned to the HRS due to a back-hopping oscillation, which was considered as one of the failure mechanisms, as shown in Figure 8c. Therefore, the switching on
Figure 7. a) Threshold-switching behavior in various selector devices such as MIT, OTS, and volatile CBRAM structures connected to a capacitor can emulate the neuron’s integration and fire functions. Reproduced with permission.\textsuperscript{110} Copyright 2019, Wiley-VCH. b) All resistive memory based neuromorphic systems have been experimentally demonstrated by implementing the RRAM-based synaptic arrays and volatile CBRAM-based neurons. Reproduced with permission.\textsuperscript{111} Copyright 2018, Springer Nature.
and off was regularly observed at the specified pulse. Although the obtained frequency of the on/off switching was a stochastic and probabilistic, frequency was discovered to be proportional to the output current intensity. The 4-bit precision that can be distinguished by the MRAM-based neuron without a capacitor and reset circuit achieved an accuracy of 82% to be obtained for the CIFAR-10 image recognition.

3.2. Threshold Switching for Oscillation Neurons

The fired output can be represented in different ways. When the NbO₂-based selector was connected to a load resistor, where the load resistance ($R_{\text{load}}$) is in between $R_{\text{off}}$ and $R_{\text{on}}$ of the selector (i.e., $R_{\text{off}} > R_{\text{load}} > R_{\text{on}}$), in a voltage divider configuration, an oscillation in voltage was monitored in real time. Most of the voltage was initially applied to the selector because the $R_{\text{off}}$ of the selector was greater than the $R_{\text{load}}$. As the charged voltage at the selector exceeded the $V_{\text{th}}$, the off-state of the selector was rapidly switched to the on-state. Because the $R_{\text{on}}$ of the selector was now lowered, the voltage began to discharge until the voltage remaining on the selector reached a hold voltage ($V_{\text{hold}}$), which is the minimum driving force required to maintain the on-state, below which the on-state of the selector was promptly switched to the off-state. The reversible transition of the selector repeatedly induced the back-and-forth voltage charging and discharging, causing an oscillation with a specific frequency in the range of $V_{\text{hold}}$ and $V_{\text{th}}$. Taking one step forward based on the single oscillation neuron with an off-ship discrete load resistor, an 1D 12 x 1 crossbar array that structurally resembles a column of the weight matrix, where one neuron is connected with multiple synapses in parallel for on-chip integration, has been demonstrated, as shown in Figure 9. The single input pulse was delivered to only one of the RRAM-based synapses, and the remaining synapses were floating. The input pulse multiplied by the conductance at the selected RRAM was expected to be observed as a read-out current along the BL at the NbO₂-based neuron, resulting in an oscillation with a slow frequency of 110 kHz. When more input vectors were loaded into the multiple rows of the synaptic array, larger amounts of the weights were summed along the BL, resulting in a larger

Figure 8. While the a) PCM and b) FeFET captured the accumulation and fire behavior, additional circuitry was required to return the nonvolatile memories to its original state. c) Meanwhile, the reset circuitry can be eliminated by deliberately exploiting the disadvantages of the MRAM. The stochastically repeated on/off switching in the MRAM showed that switching frequency was statistically proportional to the input voltage. (a) Reproduced with permission. Copyright 2016, Springer Nature. (b) Reproduced with permission. Copyright 2018, Royal Society of Chemistry. (c) Reproduced with permission. Copyright 2019, IEEE.
read-out current corresponding to the equivalently reduced total resistance. Faster oscillation frequency was measured to be proportional to the analog column current. This compact neuron facilitated the number of synaptic columns shared by one neuron to be reduced, thereby outperforming the conventional silicon neuronal circuit in latency, area, and energy consumption.

4. Conclusion and Outlook

To perform cognitive and recognition workloads in the most efficient manner, hardware systems that implement neural network algorithms are required. A typical performance metric for computing systems, tera operation per second (TOPS), is extended to account for energy efficiency as TOPS per watt (TOPS/W).

Figure 9. The input information can be judged by identifying the frequency of oscillation observed in NbO$_2$-based neuron. The integration of the neuron at the edge of the RRAM synaptic array, which converts the weighted sum to the oscillation frequency, was experimentally demonstrated. Reproduced with permission.[122] Copyright 2019, IEEE.
This study showed that nonvolatile resistive memories and selectors are attractive technologies that not only boost the TOPS/W in the systems up to a few tens of magnitude, which was a sufficient class to be used for recognition in real time but also ensure software-equivalent accuracy on various recognition tasks. Compared to digital SRAM as the binary synapse, the resistive memories stored analog information even in the small cell area. However, for accelerating the neural network computations on the entire neuromorphic system rather than single device, the multiple states of the analog resistive synaptic element needed to be tuned linearly by identical voltage pulses. Therefore, the aim of this review was to address recent progresses and strategies to solve the problem, considering the underlying working principle of each memory candidate. In summary, benchmarking and comparing key performance indicators was shown in Figure 10 to provide design options for building neuromorphic systems. Due to the commercialization of the PCM in the memory field, a solid understanding of physical mechanism and thorough reliability analysis that lead to the development of reliable devices with advanced compensation circuits can continue to expand the possibilities of the PCM for the neuromorphic computing systems. In addition to DNN, another spiking neural network for the next-generation neural network was implemented on PCM-based neuromorphic chips,[123] motivating and highlighting the need and importance of analog synaptic devices. However, the PCM seemed to be far from the ideal synaptic device due to the limited achievable conductance states and its nonlinear and asymmetric response to the consecutive identical pulses; hence the degradation of the recognition accuracy during the training. This is because the phase-change behavior is very sensitive to the compositions of the chalcogenide material. The composition of well-known GST materials was the result of optimizing the trade-off relationship between speed and operating current, making it difficult to modify the composition and materials to improve synaptic properties. Therefore, the studies have primarily attempted to subdivide the synaptic components such as arranging two PCM devices and adding 3T–1C devices to offset the shortcomings of unit PCM. The RRAM, which can operate at a lower operating current than the PCM and can be scaled at 10 nm, has been extensively studied for the synapses and has reasonably satisfied most requirements. In addition to achieving linear and symmetric weight update through innovative material and device engineering, defect-tolerant algorithms and circuits have been developed to evaluate the reliability of each state and various failure modes. For memory applications, the range from 1 to 10 μA was the preferred operating current required for unit RRAM device considering the array size and sensing speed. Meanwhile, for neuromorphic VMM accelerators, most RRAM devices in the column may be required for simultaneous reading depending on the input vector in the worst case. Thus, the number of the RRAM devices placed in the column is related to the quantity of current that the external drive and sense circuits can handle, constraining the maximum allowable array size. In addition, it should be considered that a reduced current level of the RRAM distorted I–V linearity induces an actual weighted sum current that is lower than expected, causing inference error. The most neuromorphic test-chips with peripheral circuits have been demonstrated with the PCM and RRAM synapses with 1T–1R configurations. The three-terminal transistor will eventually be replaced by the two-terminal selector depending on the applications. The area improvement is clear with the introduction of the selector, but conductance linearity as a function of voltage sweep and pulse for weighted sum and weight update, respectively, can be affected.[66] The increase in the operating voltage in the 1S–1R synapse due to the additional selector needs to be optimized while considering the operating power consumptions. The ECRAM that utilizes the ion transport across the entire area, not locally, is still in the early stages of research. The lateral conductance states can be maximally achieved because the ions provided vertically were sophisticatedly controlled from the gate in the ECRAM. However, the dynamic range of the conductance extracted from minimum and maximum levels was low. Even at the expense of the occupied area loss, the nearly perfect synaptic behaviors of the ECRAM was attractive to be used as synaptic elements dedicated for on-chip training. The slow speed to drive the ions and uncertain reliability issues that can be affected by scaling need systematic further investigation through a deep understanding and linking of each role of the selected ions. The use of the ferroelectric polarization mechanism rather than ion-migration-enabled reliable conductance of the FeFET synapses to be controlled symmetrically and promptly. Nevertheless, the conductance related to the number of ferroelectric domains that are rotated in the device and updated by energy- and area-inefficient nonidentical pulse scheme. The variability, which is one of the noticeable reliability issues in the other resistive synaptic devices, is significantly low, but the retention and endurance of the multilevel conductance should be further verified. To date, the synaptic properties have been evaluated in the usual FeFET fabricated for memory applications. Specific engineering methodologies aimed at neuromorphic applications leave design spaces to allow for conductance update in the ferroelectric materials through the identical pulse. Device-level studies on the FeFET-based synapses have been improved in recent years, but it is noteworthy that simulation modeling that accurately describes the physical ferroelectric behavior and matches the experimental results is well-established.[124] Design exploration for kernel operation of convolutional neural networks and DNN accelerators based on simulated FeFET devices has been extensively studied to pioneer more diverse and appropriate options for using the FeFET synaptic elements.[123] For the MRAM with the highest maturity among

Figure 10. Benchmarking overall performance indicators for the analog synaptic candidates.
the emerging memory technologies from manufacturing process and physics perspectives, the analogous behavior beyond reliable binary state has been observed by adopting a new writing mechanism called SOT. However, application flexibility is expected to be low because it is difficult to control the current range that can be obtained and the small on/off ratio ($\approx 2 \times$). Using different types of resistive memories and conventional devices in a hybrid configuration is considered the fastest way to implement fully functional neuromorphic systems compared with developing a single universal memory to perfectly satisfy all the tough criteria. This approach complemented the drawbacks of each memory with other devices, relaxing and alleviating requirements of synaptic devices. It also increases the degree of the freedom to use certain resistive memories that exhibit particularly prominent features such as excellent linearity of the conductance within a very short duration.

For ultimate parallel computing systems, to process what is computed at the synapse in neuron, preference is given for the implementation of the devices with the same size as the BL of the synapse array. Utilizing the capability to provide instantaneous current by the selector-based compact neurons enabled effective classification of the analog weighted sum current based on integrate-and-fire or oscillation frequency modulation technique. By precisely fine-tuning the ion migration and phase transition to have multiple states of the nonvolatile PCM and RRAM for the analog synapse, and intentionally enhancing the volatility of the memory for the neuron, all emerging memory-based neuromorphic systems have been reported.

Several aspects of the implementation and utilization of the neuromorphic hardware have remained unexplored. Hence, important features of the synaptic and neuronal devices may differ from speed, energy, and capacity perspectives depending on the applications ranging from cloud, fog, and edge computing. In particular, unlike the conventional silicon transistors, in which performances have been improved primarily by geometrical scaling and cell design, the synaptic and reliability characteristics of each emerging device are strongly related to the materials used. Further, we believe that unconventional computing platforms are not limited to emerging device technologies, and it can be realized by CMOS and new devices integrated systems. Mixed CMOS-emerging memories hardware can make cognitive tasks more efficient, and will be an intermediate step before ultimately implementing future computing systems implemented entirely with non-CMOS devices. Therefore, it is hoped that the findings and approaches discussed in this article will be a stepping stone toward significant technological advances that can lead to social change beyond building neuromorphic hardware systems.

Acknowledgements

This work was supported in part by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (19ZB1800 and 20ZB1100), and The National Research Council of Science & Technology (NST) grant by the Korea government (MSIP) (No. CAP-14-01-KIST).

Conflict of Interest

The authors declare no conflict of interest.

Keywords

cross-point arrays, emerging resistive memory, neuroorphic systems, neurons, resistive synapses

Received: May 27, 2020
Revised: June 27, 2020
Published online: August 23, 2020

[1] G. E. Moore, IEEE Int. Electron Devices Meeting, IEEE, San Fransisco, CA 1975, pp. 11–13.
[2] A. Spessot, H. Oh, IEEE Trans. Electron Devices 2020, 67, 4.
[3] A. Goda, IEEE Trans. Electron Devices 2020, 67, 4.
[4] J. Von Neumann, The Computer and the Brain, Yale University Press, New Haven, CT 2012.
[5] S. Rabii, E. Beigne, V. Chandra, B. D. Salvo, R. Ho, R. Pendse, Symp. on VLSI Technology, IEEE, Kyoto, Japan 2019, p. C10-1.
[6] M. Bear, B. W. Connors, M. A. Paradiso, Neuroscience: Exploring the Brain, 4th ed. 2015.
[7] E. Kandel, J. Schwartz, T. Jessell, Principles of Neural Science, McGraw-Hill, New York, NY 2000.
[8] Y. LeCun, Y. Bengio, G. Hinton, Nature 2015, 521, 436.
[9] A. Krizhevsky, I. Sutskever, G. E. Hinton, in Proc. of the 25th Int. Conf. on Neural Information Processing Systems, Vol. 1, 2012, pp. 1097–1105.
[10] A. Graves, A.-R. Mohamed, G. Hinton, IEEE Int. Conf. on Acoustics, Speech and Signal Processing, IEEE, Vancouver 2013, pp. 6645–6649.
[11] Y. H. Chen, T. Krishna, J. Emer, V. Sze, IEEE Int. Solid-State Circuits Conf., IEEE, San Francisco, CA 2016, pp. 262–263.
[12] V. Sze, Y. H. Chen, T. J. Yang, J. Emer, Proc. IEEE 2017, 105, 12.
[13] J. Tang, C. Deng, C.-B. Huang, IEEE Trans. Neural Networks Learn. Syst. 2016, 27, 4.
[14] N. P. Jouppe, C. Young, N. Patil, D. Patterson, A. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P.-L. Cantin, C. Chao, C. Clark, J. Coriell, M. Daeley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, et al., Proc. of the 44th Annual Int. Symp. on Computer Architecture 2017.
[15] Q. Xia, W. Robinett, M. W. Cumbie, N. Banerjee, T. J. Cardinali, J. J. Yang, W. Wu, X. Li, W. M. Tong, D. B. Strukov, C. S. Snider, G. M. Ribeiro, R. S. Williams, Nano Lett. 2009, 9, 10.
[16] P. Narayanan, A. Fumarola, L. L. Sanches, K. Hosokawa, S. C. Lewis, R. M. Shelby, G. W. Borr, IBM J. Res. Dev. 2017, 61, 11.
[17] D. Ielmini, H. S. P. Wong, Nat. Electron. 2018, 1, 333.
[18] B. Yan, B. Li, X. Qiao, C.-X. Xue, M.-F. Chang, Y. Chen, H. Li, Adv. Intell. Syst. 2019, 1, 1900668.
[19] S. Schmitt, J. Klaehn, G. Bellec, A. Gruell, M. Guettler, A. Hartel, S. Hartmann, D. Husmann, K. Husmann, V. Karasenko, M. Kleider, C. Koke, C. Mauch, E. Mueller, P. Mueller, J. Partzsch, M. A. Petrovici, S. Schiefer, S. Scholze, B. Vogginger, R. Legenstein, W. Maass, C. Mayr, J. Schermell, K. Meier, Int. Joint Conf. on Neural Networks 2017.
[20] P. A. Merolla, J. V. Arthur, R. A. Iacca, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Tabu, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, D. S. Modha, Science 2014, 345, 6197.
[21] J.-S. Seo, B. Brezzo, Y. Liu, B. D. Parker, S. K. Esser, R. K. Montoye, B. Rajendran, J. A. Tierno, L. Chang, D. S. Modha, D. J. Friedman, IEEE Custom Integrated Circuits Conf. 2011.
[22] G. Yeap, X. Chen, B. Yang, C. J. Lin, F. C. Yang, Y. K. Leung, D. W. Lin, C. P. Chen, K. F. Yu, D. Chen, C.-K. Chang, S.-Y. Lin, H.-T. Chen,
Jiyong Woo received the Ph.D. degree in materials science and engineering from the Pohang University of Science and Technology, Pohang, South Korea, in 2017. He was a postdoctoral research associate with the Arizona State University, Tempe, AZ, and Georgia Institute of Technology, Atlanta, GA, USA. He is currently a senior researcher with the Electronics and Telecommunications Research Institute, Daejeon, South Korea. His current research interests include the fabrication and characterization of nanoelectronic devices for nonvolatile memory and neuromorphic applications to design of the future computing systems.

Jeong Hun Kim received his B.S. degree from the School of Mechanical and Aerospace Engineering at Seoul National University (SNU) in 2009 and obtained his Ph.D. degree from the same department at SNU in 2015. Since 2017, he joined the Electronics and Telecommunications Research Institute (ETRI) as a member of engineering staff, where he worked with Emerging Materials Lab. His research interests include the design, fabrication, and characterization of functional surfaces, and micro-/nanostructured interfaces of the biomimetic systems.

Jong-Pil Im received the B.S. degree in electrical engineering from Hiroshima University, Hiroshima, Japan, in 2007, and the M.S. degree in electrical engineering from Osaka University, Osaka, Japan, in 2009, and the Ph.D. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST) in Daejeon, Korea, in 2013. He joined Electronics and Telecommunications Research Institute (ETRI) in 2013 and is currently a senior researcher in ICT Creative Research Lab. His research interests include design and modeling of switch mode power converters, charge pumps, LDO, and low-power converters for energy harvesting applications.