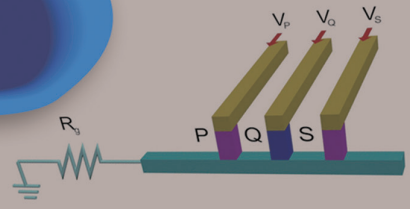
ADVANCED ELECTRONIC MATERIALS

NEUROMORPHIC COMPUTING

In article number 1600090, D. S. Jeong et al. review memristors and their potential application in new, energy saving forms of computation, including stateful logic and neuromorphic computing. Memristors in a cross-bar array format (background image) create a two-terminal voltage- or charge-driven non-volatile memory and logic component, serving as the critical circuit element for mimicking the human brain. Their discussions on stateful logic are based on material implication logic (center image).



WILEY-VCH

ADVANCED ELECTRONIC MATERIALS
www.advelectronicmat.de

Memristors for Energy-Efficient New Computing Paradigms

*Doo Seok Jeong, Kyung Min Kim, Sungho Kim, Byung Joon Choi, and Cheol Seong Hwang**

In this Review, memristors are examined from the frameworks of both von Neumann and neuromorphic computing architectures. For the former, a new logic computational process based on the material implication is discussed. It consists of several memristors which play roles of combined logic processor and memory, called stateful logic circuit. In this circuit configuration, the logic process flows primarily along a time dimension, whereas in current von Neumann computers it occurs along a spatial dimension. In the stateful logic computation scheme, the energy required for the data transfer between the logic and memory chips can be saved. The non-volatile memory in this circuit also saves the energy required for the data refresh. Neuromorphic (cognitive) computing refers to a computing paradigm that mimics the human brain. Currently, the neuromorphic or cognitive computing mainly relies on the software emulation of several brain functionalities, such as image and voice recognition utilizing the recently highlighted deep learning algorithm. However, the human brain typically consumes ≈10–20 Watts for selected "human-like" tasks, which can be currently mimicked by a supercomputer with power consumption of several tens of kilo- to megawatts. Therefore, hardware implementation of such brain functionality must be eventually sought for power-efficient computation. Several fundamental ideas for utilizing the memristors and their recent progresses in these regards are reviewed. Finally, material and processing issues are dealt with, which is followed by the conclusion and outlook of the field. These technical improvements will substantially decrease the energy consumption for futuristic information technology.

1. Introduction

1.1. The Energy Crisis and Information Technology

The amount of digital data worldwide exceeded that of analog data in 1998 due to the explosive growth of personal computers, smartphones and enterprise systems.[1] It is expected that the total amount of digital data in 2040 (only 25 years from now) will be $\approx 10^{28}$ bytes (1 byte = 8 bits), which is approximately one million times greater than the current total.[2] A more interesting (and also sobering) expectation is that the total number of binary operations in 2040 will be $\approx 10^{40}$, which is an astronomically large number.[2] The average energy consumption of binary digital operations, including logic, memory, and input/output operations between logic and memory chips, is currently ≈0.1 picojoules ($\approx 10^{-13}$ Joules). Therefore, if this rate of energy consumption per binary operation is maintained, the total energy expenditure in 2040 for computer operations will reach $\approx 10^{27}$ Joules, which is far higher than the total energy that humans will be able to produce at that time. In 1961, Landauer published a paper on the theoretical aspects of computation based on digital logic and demonstrated that effective computation must be based on an irreversible process (otherwise, the input and output cannot be distinguished), the unit process of which will require minimum energy on the order of $\approx kT$ (Boltzmann constant × temperature), which is $\approx 10^{-21}$ Joules at room temperature.[3] This estimate means that the aforementioned energy consumption in 2040 could be decreased by a factor of $\approx 10^{8}$, which appears to be quite promising. However, what Landauer showed is that this energy is a sort of fundamental limit (energy for one thermodynamic degree of freedom) without consideration of a detailed method of how binary states can be represented by a physical entity and what type

Dr. D. S. Jeong
Center for Electronic Materials
Korea Institute of Science and Technology
5 Hwarang-ro 14-gil, Seongbuk-gu, Seoul 02792, Republic of Korea
Dr. K. M. Kim
Hewlett Packard Laboratories
Hewlett Packard Enterprise
Palo Alto, California 94304, USA
Prof. S. Kim
Department of Electrical Engineering
Sejong University
Neungdong-ro 209, Gwangjin-gu, Seoul 143–747, Republic of Korea
Prof. B. J. Choi
Department of Materials
Science and Engineering
Seoul National University of
Science and Technology
Seoul 01811, Republic of Korea
Prof. C. S. Hwang
Department of Materials Science and Engineering
Inter-university Semiconductor Research Center
College of Engineering
Seoul National University
Seoul 151–744, Republic of Korea
E-mail: cheolsh@snu.ac.kr

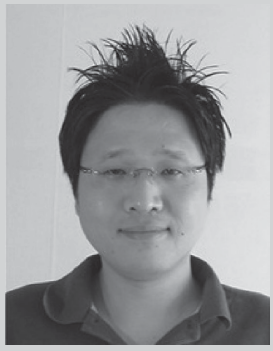
DOI: 10.1002/aelm.201600090

of circuitry can detect and modulate it. A modern electronic circuit always involves thermal noise that can be conveniently expressed as a noise voltage of ≈25 mV at room temperature (kT/q, where q is the elementary charge). Therefore, any practical circuit must have a signal voltage at least several times higher than this value. For example, dynamic random access memory (DRAM) usually has a stored charge of ≈10 femtocoulombs (10^{-14} C), which can induce ≈100 mV of voltage change when this storage charge is transferred to a bit line of which the capacitance is ≈100 femtofarads (10^{-14} C/10^{-13} F = 100 mV). 10 fC of charge storage/dissipation with an operation voltage of 1 V corresponds to 0.01 picojoules (10^{-14} Joules), which far exceeds the Landauer limit. In fact, the input/output of digital data between the logic chips (central processing unit (CPU), and graphic processing unit (GPU)) and memory (DRAM or flash memory) requires approximately 10–100 times more energy than does the logic operation itself or the memory of the data, as long as the data transfer between the logic and memory chips occurs in a form of electrical signal. Recently, Sun et al. reported an optical interconnection technology that can be applied between two identical chips containing both a processor and static random access memory (SRAM) using process technologies that are completely compatible with current complementary metal oxide semiconductor field effect transistors (CMOSFET) to enhance the communication bandwidth.[4] However, these researchers did not report on the possible downsides of such communication method in comparison to its electronic counterpart, as the transmitters, optical amplification, and internal heating that are required to compensate for the resonance deviation might degrade the energy efficiency. These discussions indicate that a radical change must occur in computer architecture to support the feasible growth of information technology (IT). It is notable that the famous Moore's law has survived even though the improvement of its recent performance has been delayed, but that Dennard's law, which indicated performance increase without increasing power consumption, was halted at a much earlier time.

1.2. Strategies to Solve the Problem

Three fundamental approaches can be used to solve this catastrophic energy problem: i) decrease the energy per computation, ii) eliminate the date volatility, and iii) decrease the number of computing steps. In fact, decrease in the energy per computation has been the approach historically used to develop modern computing systems in the form of scaling. However, for electric charge-based computation, minimal opportunity is available for decreasing the energy per computation due to the presence of the aforementioned thermal noise. Other physical parameters, such as spins or photons, can be used to perform logic operations, which might alleviate the noise issue in charge-based computation, but if the final output of the processed data exists in charge form, the problem always remains.

The data volatility of memory is definitely another critical source of energy consumption. Two typical volatile memories in modern computers are SRAM and DRAM. Recently emerging magnetic random access memory based on the spin transfer torque effect (STT-RAM) is considered a strong contender as a possible replacement for SRAM in the form of cache memory because it has a slightly smaller cell size and shows the better device performance than SRAM in addition to its inherent merit of data non-volatility.[5] Although DRAM represents the main memory of modern computers, the fact that it refreshes the data 5–10 times per second even if they are not necessarily retrieved means that significant energy is wasted to simply maintain the data in it.[5] If the main memory can be replaced with high-density nonvolatile memory, the possibility of which does not appear to be high at the current state, much wasted energy could be conserved. Hard disc drive (HDD) has been the



Doo Seok Jeong is a senior scientist at the Korea Institute of Science and Technology (KIST), South Korea. He received his PhD degree in materials science from RWTH Aachen, Germany, in 2008. Since 2008, he has worked for KIST. His research interests focus on spiking neural networks for temporal learning, ranging from building blocks to online learning algorithms.

Kyung Min Kim received his BE and PhD in materials science and engineering from Seoul National University (SNU), Seoul, Korea, in 2003 and 2008, respectively. He is currently a research scientist at Hewlett-Packard Labs in Palo Alto, CA, USA. He is interested in memristor and selector materials and their applications.

Cheol Seong Hwang received his PhD degree from Seoul National University, Seoul, Korea, in 1993. Since 1998, he has been a Professor at the Department of Materials Science and Engineering, Seoul National University. His current research interests include high-k gate oxides, dynamic random access memory capacitors, new memory devices including resistive RAM devices and ferroelectric materials and devices, energy storage capacitors, as well as neuromorphic computing.

main product for the storage memory for a long time, but it has been rapidly replaced by NAND flash memory, especially in hand-held devices. Memristors have been researched to be used as both main and storage memories, but the main focus has been put on the storage side. Hwang recently reviewed recent progresses in semiconductor memory technology.[5] Nonetheless, these are still the topics within the scope of von Neumann computing and, therefore, will not be dealt with in detail in this review. Data non-volatility has important implications for neuromorphic computing, as discussed in the following sections. Therefore, it appears that decreasing the computational step number is the only feasible option, which is the main theme of this review.

1.3. Stateful Logic and Neuromorphic Computing for Energy-Saving Computation

Decreasing the number of computing steps is not a straightforward task because this goal could be achieved by changing the computing paradigm, which means that a well-established computing process based on von Neumann architecture must be reconsidered partially or completely. An immediate but still intermediate solution might be combining the logic and memory chips, which fundamentally eliminates the energy cost incurred by the data input/output step. Such an approach might be duly stated as a "stateful logic" approach.[6] In fact, the current logic chips already contain a large number of memory functions, such as the flip-flop and latch circuits, in addition to the SRAM that enables cache memory. However, the fundamental problems of these conventional memory systems are that they are all volatile memories, and their cell size is too large to be considered high-density memories (the SRAM has a cell size of ≈100– 150F^2, where F is the minimum feature size), whereas DRAM and NAND flash memory have a cell size of ≈6F^2 and ≈5F^2, respectively. Recent new memory architectures (crossbar or cross-point array) that use multi-layer stacking approaches could deliver a cell size < F^2. Merging of the CPU and DRAM has been attempted for a long time in the semiconductor industry but has not been successful due to the rather different processes required for logic transistor fabrication and memory cells, especially the notably tall capacitors in DRAMs.[5]

During the operation of a certain computing program with a modern computer, which can be described as a (universal) Turing machine, all data must be read out and written back using the floating-point arithmetic calculation method even if these data are not modified. However, the human brain works differently and uses ≈20 Watts of power to perform a certain computational (intellectual) task, whereas sometimes a supercomputer needs several tens of kilo- to megawatts of power consumption for the same task. A direct comparison between the energy consumption for the same computational task performed by the human brain and a CMOS logic circuit is quite difficult mostly because many of the details of brain functionality are not yet clearly understood. It is also anticipated that all the computing task will not be performed by human-like computer even if it will be successfully developed in the future. There will be still areas in which the current deterministic computation works better than the human-like computer does. Nonetheless, it will be still elucidating to examine the detailed computational process for a simple exemplary problem for obtaining a better image on how the current von Neumann computer works and how inefficient it can be in some cases in comparison to the human brain. The exemplary problem is to calculate 2 + 3. Of course, a human can calculate its answer from the very basic mathematics, and once the answer is memorized in the brain, the person recalls this knowledge when the same question is asked. Although the details of the process of recall are not precisely known, the specific area of the brain that functions for the recall process can be known from experiments using ca. the functional magnetic resonance image technique. The recall process must involve several spikes of neuronal signals, where one spike typically takes ≈1 pJ (see Section 3.5), and thus, it might be reasonable to assume that the recall process will require several tens of pJ. In contrast, the energy used by the CMOS logic circuit mentioned above can be accurately calculated because the necessary circuits and their functions are precisely known. Interestingly, the conversion of decimal numbers 2 and 3 to binary numbers 10 and 11 takes ≈100 logic steps each, and the addition itself using a full adder circuit takes only ≈20 steps. The back conversion of the binary number 101 to the decimal number 5 takes 40 steps, and thus, the total logic steps needed for 2 + 3 = 5 are ≈260 steps, meaning that ≈26 pJ is necessary for the example computation. This value is not much different from the approximate assumption for the recall process of the brain, suggesting that perhaps each small computation step itself does not create the million times difference in the energy efficiency mentioned above. Therefore, it might be conjectured that the critical difference between the brain and a CMOS computer is the manner in which each of the logic steps is arranged along the space and time dimensions and how efficiently each small computation element is allocated. This aspect of logic computation is discussed again from a slightly different point of view in Section 2 in which the implications of "stateful logic" are covered.

Neuromorphic computing or cognitive computing has been an active research field in computer science in conjunction with the artificial intelligence (AI). Along with the up and down turns of the AI research, it also experienced several different phases, and still the directions of its research are quite diverse.[7] However, it could be discriminated from the current deterministic computation from the following aspects. In this review, the term "cognitive" is regarded as having a similar meaning of "neuromorphic". It finds an "optimal" rather than "correct" solution to a complex problem for which conflicting evidence or factors might be present. The process must be adaptive, interactive, iterative, stateful, and contextual, and it thus more closely resembles the human than current computers. Therefore, neuromorphic computing has also been used to refer to new hardware and/or software that mimic the functioning of the human brain. Once the neuromorphic computing system is built, it will interact with humans in a much more intimate fashion than current computers and will allow users to concentrate more efficiently on creative works. Nevertheless, it is still unclear how such a new computing paradigm can evolve in the real world, although several impressive improvements in this field have occurred, including the recent success of the "deep learning" algorithm[8] and customized processing units in

support of the algorithm (see Section 3). Therefore, at least a certain portion of neuromorphic (cognitive) computing is accessible in the near future as of this moment. Nonetheless, this approach hardly offers energy efficient computation. Therefore, new hardware is necessary. To date, significant improvements have been reported on the hardware side, as represented by the "True North" neuromorphic chip of the IBM labs[9] and the even more recent application processor of Qualcomm (neural processing unit: NPU), a portion of which is based on neuromorphic architecture.[10]

1.4. Comparison Between the Evolutionary Processes of Computer and Human Brain

Another critical aspect that might be learned from nature for composing a new computing machine is the ratio between data processing elements (neuron) and memory (synapse), although the functionality of neuron and synapse in human brain is somehow lumped, and, thus cannot be directly compared with CPU and memory (DRAM and Flash). Nevertheless, such kind of comparison will provide a certain useful insight on the futuristic development of general hardware configuration of a computer. **Figure 1**a shows the ratio for different animals along the evolutionary stage, where the increase rate of the number of neurons is ≈2/3 the power of that of the synapse (blue symbol).[11] The reason for this ≈2/3 power relation might also be derived from the energy consumption and accompanying heat generation in neuronal processes. As the linear dimension of the brain increases, the volume of total synapses increases according to the cube of the linear dimension because these synapses approximately correspond to non-volatile memory and do not consume as much active power. However, neurons work actively and generate heat, and thus, if they are located deep inside the brain, a subset could be thermally damaged. Therefore, these neurons must be preferably located on the surface region of the brain to effectively dissipate the generated heat, meaning that the number of neurons might increase according to the square of the linear dimension of the brain. This comparison can explain the different evolutionary rates of neurons and synapses in mammalian brains. This is in fact a rash oversimplification of evolutional processes of brains, but the outer location of cerebral neocortex in human brains may provide a certain level of justification for such hypothesis, considering that the other parts of brain evolved from the those of reptiles are located relatively deep inside.

Because this is a generic problem related to the energy consumption and heating, a similar trend could be expected for semiconductor chips if they exist in a three-dimensional structure, i.e., the CPU will still take on a planar shape, but the memory will take on a stacking structure, especially for the case of nonvolatile memory. In contrast, the main memory, such as DRAM, might encounter difficulty in pursuing such an evolutionary route because of the effective heat dissipation problem in addition to the difficulty of fabrication in a vertical configuration.

Therefore, it is quite notable that the human-made computer has also displayed a similar evolution between the data processing elements (CPU) and memory (disk density), as shown in Figure 1a (red symbol).[12,13] It is still a tricky task to compare a computer and brain in this aspect, but the increase rate of clock speed of a CPU can be compared with that of the disk density, which shows an even faster rate of increase of memory density than the brain. In fact, to mitigate such a problem, the number of transistors in logic chips has also increased rapidly, thus making the overall rate of increase of memory and logic functionality more comparable, as shown in Figure 1b, where the logic functionality is represented by the CPU clock time × transistor numbers.

The CPU clock and transistor number increased concurrently while clock speed increase has been leveled down already. It might be more reasonable to compare the MOSFET number in CPU chip and neuron number in a brain, but this may give a wrong impression that the number of transistors in CPU and the number of neurons in brain have one-to-one correspondence, which is certainly not the case. The performance or functionality of the two entities are better compared by examining the CPU clock and neuron number considering the most likely sequential and parallel computing processes, respectively. Other comparison could be made, such as the CPU transistor number increase vs. DRAM cell density increase (green symbol in Figure 1a). In this case, the integration density increase rates are much more compatible with each other, which might be due to the fact that both chips consume power constantly during operation, so that they should be made on planar surface. When the functionalities of DRAM and HDD (or NAND flash) are lumped together and compared with synapse in brain, the trend of memory and storage density vs. CPU performance appears to follow natural evolution of mammalian brain. Therefore, it can be conjectured that in future computer, no matter whether it is with conventional von Neumann or new architectures, the importance of higher memory density over the faster CPU will increase.

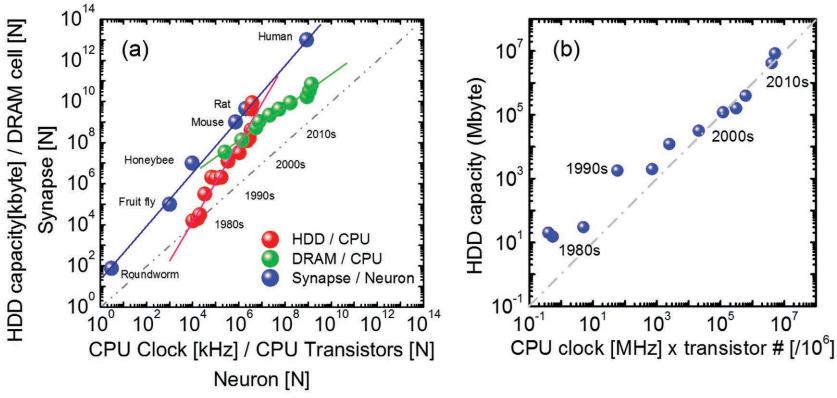


Figure 1. a) Ratio of data processing units (neurons) to memory units (synapses) in animals plotted with evolutionary stage (blue symbols) and in computer systems (red symbols). DRAM cell density increase vs. CPU transistor number increase is represented by green symbol, and b) memory capacity as a function of logic functionality is represented by CPU clock × number of transistors.

As mentioned previously, the memristor is one highly appealing contender for ultra-high density memory, especially in three-dimensional stacking. For nonvolatile memory, three-dimensional stacking is highly feasible, and it is indeed the case as understood from the mass-production of vertical NAND flash memory in 2014.[14]

1.5. The Memristor

The memristor could be a critical ingredient as a stateful logic element and artificial neuron/synapse in von Neumann and neuromorphic computing paradigms. The memristor was suggested by L. Chua in 1971 as the fourth elemental circuit component that correlates the flux and charge.[15] However, although the memristor was claimed to be experimentally demonstrated in 2008 by the Hewlett-Packard (HP) group, it could be more appropriately described as a charge-controlled variable nonvolatile resistor, the precise state of which could be further modified by the applied voltage.[16] Therefore, an immediate application of the memristor might be in the form of the critical component of resistance switching random access memory (ReRAM), according to the suggestion made by the IBM group in the 1960s,[18–20] although it has not been highlighted until 2000, when the IBM Zürich group reported feasible resistance switching (RS) properties from a Cr-doped $SrZrO_3$ film.[17] Since then, ReRAM has become a focus of both academia and industry. Several excellent review articles were published for ReRAM using various materials, and therefore, this review does not focus on that area. Instead, this paper primarily explores new applications of memristors for new energy-efficient computing paradigms.[21–28]

One of the exciting aspects of the memristor is the notably high dynamic variance of its properties in response to external stimuli, such as voltage or charge, which means that the state of a memristor can be drastically changed with a minor change in input, making it appear to show chaotic behaviors at times. A typical example of such behaviors is its negative differential resistance (NDR), which accelerates the response speed up to the chaotic level. L. Chua recently stated that biological neurons are 'poised at the edge of chaos'.[29] This behavior means that the resting states of neurons display near-chaotic behaviors such that even a minute perturbation, such as thermal fluctuation from the environment, can make the neurons fire with apparently chaotic behaviors. It is still unclear how such chaotic behavior can be used or what the possible role of such chaotic behavior could be in the human brain, but the similarity between the near-chaotic state of the neuron and memristor could be a significant ingredient for a futuristic cognitive computing machine.

Another critical feature of the memristor is its threshold switching (TS) behavior. This TS can be used as a critical component to easily build an oscillation circuit that might require several tens of MOSFETs if attempted with conventional Si-based semiconductor devices. The recent demonstration of a "neuristor" by the HP labs is a good example of such performance, which is also related to the high dynamics (and somehow chaotic behavior) of the memristor.[30] Additional details on this device are provided later.

1.6. Other Approaches

There are several other approaches that may offer better energy efficiency than the current von Neumann computer. Quantum computing is one of the appealing contenders for decreasing computing energy consumption because it can process multiple data sets simultaneously, which is beneficial for deceasing energy per unit computation as well as the total number of computing steps. However, this approach still lies within the von Neumann framework (at least up to now), and physical implementation of the computing principle is still rather challenging. Therefore, this topic is not addressed in this review, but the entanglement of quantum states within a certain unit and simultaneous operation of many bits is quite intriguing, especially if compared with brain function. The recent announcement of the D-wave 2X quantum computer operating at 15 mK represents a highly impressive improvement in this field.[31]

Analog computing, which can also be described as "approximate" computing, is another appealing contender for an energy-efficient next-generation computing machine.[32] In fact, the analog computer is much older than the current digital computer, but it has been obsolete since the digital paradigm dominated the field as driven by the enormous progress of general-purpose microprocessors (CPU) and solid-state memory (DRAM and Flash) over the past several decades. As processors become more general, the calculation and energy efficiencies become worse. Single-core CPU is perhaps the worst, and the multi-core CPU, graphic processing unit (GPU), field programmable gate array (FPGA) and application specific integrated circuit (ASIC) all follow. In a sense, the analog computing processor might represent a type of ASIC device, and the fundamental analog computing processor must rely upon a physical (or chemical or even biological) function of a certain material. For example, the differential equation, which is often encountered in many computer simulations, can be alternatively solved by an integration equation, and it is well known that a charge stored in a capacitor is the time integration of the flown-in current. Therefore, measuring the capacitor voltage after a certain number of current pulses applied to a capacitor can represent an integration function. The memristor, for which the physical status can be changed by accumulation of electrical stimuli, can therefore act as a fundamental asset for an analog computing processor. Various limitations and drawbacks exist in analog computing compared with digital computing, and one of them is the inability to achieve an arbitrary level of calculation accuracy. However, even in digital computing, the precise requirement for the level of accuracy is not well known in many cases, and it can be expected that several critical areas could exist in which accuracy can be sacrificed. In other words, if a certain level of "error" is allowed, the calculation efficiency must be enormously enhanced. Nevertheless, this review does not address this interesting area in detail to limit the scope of this paper, but it should be noted that the synaptic behavior of the memristor, as presented in Sections 3 and 4, is essentially an analog-type behavior. The conductance of a memristor can be nearly continuously increased or decreased depending on the amplitude, duration, or number of input pulses.

1.7. Paper Outline

This review is constructed in the following order. First, the lengthy introduction described the motivation to write this review which is related with the unsustainable energy consumption along with the abrupt IT evolution. Second part will be started with the description on history of the memristor, and followed by a short review of the physical implications of the memristance. The main purpose of this section is describing the possible roles of the memristor within the von Neumann architecture. This section includes discussions on stateful logic application of the memristor within the scope of the von Neumann architecture. Readers are presented with the viewpoint that logic information flow along the time dimension, as opposed to the logic flow along the spatial dimension, which is the configuration of current logic circuits that use CMOSFETs. The material implication (IMP) or the fourth logic element, as suggested by Whitehead and Russell in 1910[33] and later by Claude Shannon in 1937[34] within the form of switch logic, and the three logic gates of AND, OR, and NOT at the present time are also discussed briefly in terms of their pros and cons. Third, the principle, role, and applications of memristors in the neuromorphic computing area (which adopts architectures different from that of the von Neumann) are addressed in detail. This part is composed of several sub-sections that describe generalities of neuromorphic computation based on online and offline learning algorithms as well as the neuromorphic system implementation. This section also discusses the possible usefulness of memristors as the artificial synapse, and limitations of them as the circuit component in such applications are also discussed. Possible approaches used to supplement this weakness are discussed. Fourth, recent notable improvements in material aspects are summarized, including electrodes and fabrication methods. More specifically, three-dimensional integration techniques are highlighted because application of these materials will require rather highly integrated devices for both von Neumann and newly developing computing architectures. Finally, the summary and outlook for this field are provided. **Figure 2** illustrates the hierarchy of the emerging computing methodologies that are addressed in this review in detail.

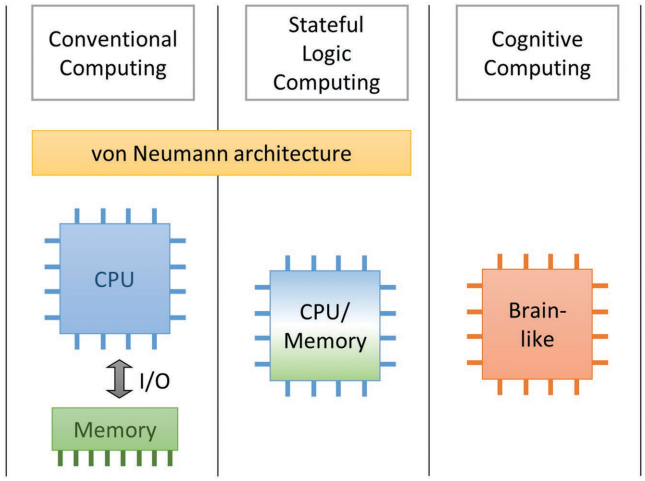


Figure 2. Hierarchy of computing methodologies. In conventional computing, the roles of the CPU and memory are distinguishable in that the CPU performs the logic operations and the memory store the data. Therefore, this process requires I/O elements for communication. In stateful logic computing, the device contains both logic operations and data storage capabilities. Both computing methodologies can be placed under the von Neumann architecture. In neuromorphic/cognitive computing, no distinguishable boundary exists between logic and memory, and this structure does not rely on von Neumann architecture.

2. Von Neumann Logic Application of the Memristor

This section is composed of three subsections: a short history of memristors, FPGA type implementation, and stateful logic using the memristor.

2.1. A Short History of the Memristor

Symmetry is one of the most dominant principles in science and was the fundamental motivation underlying the introduction of the memristor by L. Chua in 1971.[15] At that time, four fundamental variables in electrostatic circuits, i.e., current, voltage, flux, and charge, and three circuit components (defined by the relations of two of the four variables) were known, which was formidable in terms of the symmetry among the fundamental variables. Additional information was required to fill the missing piece of the symmetry that defines the relation between the charge and flux, and this element was referred to as the memristor by Chua. He also designed the equivalent circuit of the memristor composed of the existing circuit elements and described its theoretical aspects.[15,35] Ironically, this study was slightly too pioneering to be widely accepted at that time, and thus, it remained obsolete until 2008 (the original paper (Ref. [15]) was cited only 22 times over 37 years, but after 2008, it was cited over 1,400 times). In 2008, Strukov et al. reported that the unique current–voltage (I–V) relation, which is known as a resistance switching behavior in TiO_2, can be interpreted using Chua's memristor theory.[16] At that time, resistance switching related topics had been intensively studied such that the material and device technologies had been mostly established. Furthermore, the theoretical definition of the memristor was expanded as experimental data for the memristor accumulated; any type of material system that shows pinched I–V characteristics can be a memristor, meaning that most of the ReRAM system or even phase change memory material can be considered as a memristor.[36,37] Since the pioneering work by Strukov et al., the memristor has been recognized as offering high potential for emerging electronics, not only for conventional nonvolatile memory applications but also for new computing paradigms that are currently drawing great attention, as shown in this review. **Figure 3** displays the search results on the number of publications filtered using the specific keywords of ReRAM only, memristor including ReRAM, and memristor excluding ReRAM, which corresponds to the triggered area after the memristor was found.[38] This figure shows that the number of memristor-related publications continues to increase, whereas ReRAM related research is nearly saturated and has even begun a downward trend.

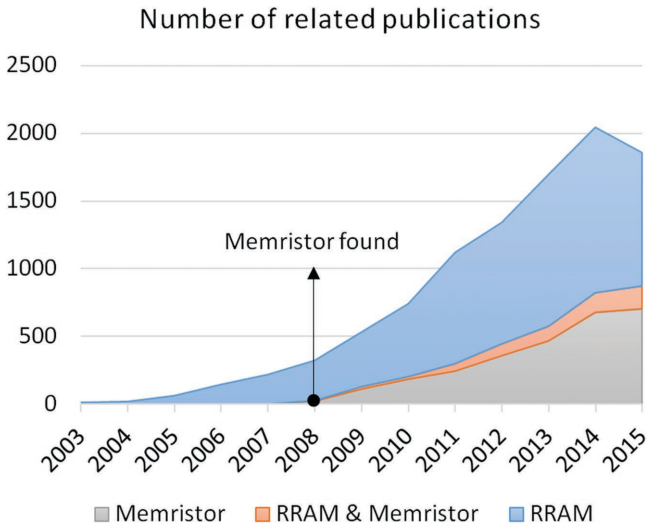


Figure 3. Number of publications filtered by specific keywords: ReRAM only (blue area), ReRAM and memristor (red area), and memristor excluding ReRAM (gray area). The gray area greatly increases after the memristor was claimed to be founded by Strukov et al. in 2008.[16] Data collected from Web Of Science, Thomson-Reuters in June 2016.

2.2. FPGA-Type Application

In 1998, Heath et al. proposed the defect-tolerant computer architecture.[39] The basic concept of this architecture is configuration of the wiring network to allow detour paths that enable the system to avoid the defective computing units. This wiring network was built on the crossbar structure, and a pair of memory and switch units were located at the cross point of the crossbar, where the programmable memory unit, composed of SRAM, stored the connectivity information and the switch unit (which was basically a transistor) received the information from the memory and turned the connectivity on or off accordingly. This architecture worked well and later resulted in the "Teramac" computer designed by HP labs.[39] This computer is an example of a new application FPGA device that has been used as a compromise between the CPU and ASIC devices. The general FPGA contains blocks and arrays of CMOSFETs for which connections can be configured to arbitrary logic circuits using the abovementioned pairs of memory and switching units. Due to this unique feature, the FPGA chips generally contain a much higher proportion of routing devices (a pair of memory and switch units) than that of other CMOS logic circuits, partly due to the high area consumption of SRAM. Another problem is that SRAM is a volatile memory device such that once power is turned off, the implemented circuit configuration disappears. Therefore, it is highly desirable to replace the SRAM with a certain type of nonvolatile memory with a much smaller cell size. It is also preferable if the memory cells can be placed on top of the logic circuit block to further conserve the chip surface area. Therefore, the memristor, or more specifically ReRAM, has been seriously considered for such an application. Replacement of the SRAM on a general FPGA chip with a memristor (which could be more appropriately described as a ReRAM in this case) does not necessarily correspond to the defect-tolerable architecture mentioned above. However, if the integration density of the memory and logic cells in a chip increases to a much higher value than that at the current time, it is reasonable to explore any type of defect-tolerable architecture because checking the functionality of all cells and replacing the defective ones with redundant cells is not an economically feasible option.

At nearly the same time, molecular electronics have been extensively studied.[40] The molecules sandwiched between two electrodes exhibited various types of electronic behaviors, i.e., conductor, diode, and even memory (which can also be assigned to memristors), such that they were promising alternatives for replacement of inorganic materials such as metal, oxide, and semiconductor, although various issues related to thermal stability and process integration still remain. Because the electronic functions of molecules can be achieved with a two-terminal structure, the crossbar structure that allows higher density and lower cost was the best platform to maximize the advantages of molecular electronics. The first step toward realizing the molecular electronics for this device was embedding them into the conventional CMOS technology. To this end, one of the most notable studies was that reported by Ziegler and Stan in 2003,[41] who proposed various applications of the crossbar-based molecular electronics in conjunction with CMOS technology as a new device paradigm. Nevertheless, the thermally fragile properties of the molecular layer, the precise location of the molecules at the specified positions on a large-scale wafer without adverse effects from other regions (contact), and accurate thickness control of the layers still impose significant challenges in this field. Therefore, inorganic-based memristor materials appear to be more promising.

In 2005, Strukov and Likharev merged the existing crossbar-based molecular electronics technology and the reconfigurable wiring network concept of defect-tolerant computing, both commonly based on the crossbar structure. Additionally, they proposed a new semiconductor-molecule hybrid architecture[42] in which they used one of the interesting behaviors shown by specific molecules sandwiched by two electrodes known as the "latching switch". This latching switch was actually a memristor, although it was not clearly realized at that time. Because the latching switch included the functions of memory and switching and could be implemented in the crossbar platform, a single cell of this device could directly replace the functions of memory and switching of the reconfigurable wiring network, which might result in a large improvement in the density (one cross-point can replace seven transistors and the extra connections between them). This concept was realized by Xia et al. in 2009 and was known as the memristor–CMOS hybrid integrated circuits.[43] In this work, Xia and coworkers fabricated the CMOS arrays and then integrated the inorganic TiO_2 memristor crossbar array on top of it in an arrangement in which the CMOS arrays and the memristor arrays are connected through vertical vias. The connectivity between the individual CMOS circuits for specific logic gates was configured via the crossbar memristor by specifically programming the states of the memristors located at each cross-point of the crossbar. This approach provided a convenient way to personalize the logic gates depending on the user purpose, but a portion of the CMOSs had to be idle such that the efficiency could not be optimized. Therefore, such an approach is suitable for FPGA-type applications rather than CPU-type applications.

2.3. Stateful Logic Application

2.3.1. A Short History

Modern computational processes are based on the Boolean algebra suggested by G. Boole in 1854.[44] In Boolean algebra, only two values exist for the variables, 1 or 0 (i.e., true or false). Using the relationship between the variables, logic operations such as 'AND', 'OR', and 'NOT' are addressed, and have become the fundamental base of digital electronics. The machine that was realized to perform Boolean algebra is the digital computer; the basic principle of digital computer was initiated by Turing and Church, and it was further refined by von Neumann with random access memory. Binary digital logic was also formulated by Shannon through switching devices. Thereafter, many digital logic circuits have been established and now dominate the IT era via CMOS logic devices and circuits.

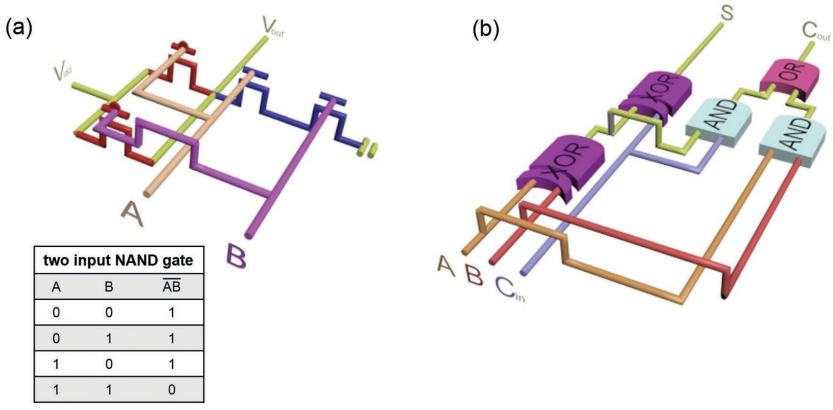
Switching components that can input signals during the operation period or otherwise be turned off are necessary for practical realization of electronic logic devices, in which the digital signals of 1 or 0 can be represented by the presence or absence of voltage and corresponding current flow. For early computing machines, due to diode's simple switching characteristic, diodes were used extensively for logic operations, and this is referred to as diode logic (DL). However, the disadvantage of this method is quite obvious. Since diode is a unipolar-type device, combinations of these devices cannot provide all logic operations, allowing only AND and OR operations. In addition, these components are passive-type devices wherein the output signal intensity becomes increasingly weaker as the logic sequences progress. This weakness can be complemented by adopting a signal amplifying system in the high-density array. The circuit configuration adopted to amplify the decaying signals with transistors is known as diode transistor logic (DTL). By replacing the diode with the bipolar junction transistor (BJT), the logic is further enhanced and is called transistor–transistor logic (TTL); BJT consists of back-to-back connection of two p–n junctions (diodes) that can easily mimic a diode. CMOS logic was finally developed to reduce operation power, becoming the core technology of modern computers. For all cases mentioned, the logic operations are based on switch technology.

2.3.2. Material Implication Logic using Memristors

Considering the memristor is a two-terminal switch, it is obvious that it can also be utilized as the switching element for logic operations. Moreover, owing to the memristor's memory functionality, unlike diode or transistor, the memory function can be used during the logic operation. This functionality has been realized and implemented within the "stateful" logic operation by Borghetti et al. in 2010,[6] as mentioned in Section 1. The suggestion was based on the material implication (IMP) gate, which has been obsolete for a long time as mentioned previously. It was also proven that almost all Boolean algebras could be achieved by combining several IMP operations. Various types of materials and operational schemes of the memristor can be adopted for IMP logic. As a result, stateful logic could have critical implications for digital computers. According to the original concept of a universal Turing machine, the conventional computing system requires memory as well as a processing unit, and these functions are not necessarily separated. In fact, the modern CPU contains a relatively high density (several hundreds of Mb) of embedded memory (SRAM), however, it is not sufficient for most of the demanding computational tasks. In order to alleviate this issue, DRAM plays the role of the main memory and continues to receive/feed data from and to the CPU during the operation. Still, this process invokes performance mismatch between the CPU and DRAM, and congestion of data through the input/output (I/O) system, requiring additional energy and cost. In memristor-based stateful logic operation, in principle, the I/O could be significantly decreased because data can be stored in the stateful logic circuit itself and used directly for the next operation. This logic operation can provide a fundamentally new paradigm for computer architecture within the von Neumann framework. Nonetheless, "stateful" itself invokes several critical problems, as discussed later, in addition to the limited functionality of IMP for certain tasks. The details of IMP operation and its circuit implementation are explained in the original paper by Borghetti et al. and a recent monograph by Vourkas and Sirakoulis (Chapter 4, Memristor-based logic circuits).[45] The monograph contains an extensive review on memristor-based logic circuits, including IMP and CMOS-like circuit implementation using complementary resistive switching (CRS) devices, and their simulations using the SPICE tool.

2.3.3. Comparison with CMOS Circuits

To explain the advantages and disadvantages of stateful logic using memristors compared to the conventional CMOS logic devices in terms of calculation efficiency, the circuit implementation of the NAND gate and full adder (FA) via CMOS devices (n- and p-type MOSFETs) are briefly presented. **Figure 4**a shows the schematic circuit diagram of NAND and its truth table whereas Figure 4b displays the schematic block diagram of FA. The CMOS NAND gate is a typical example used to explain conventional logic functionality. When the two inputs (A and B) are applied to the gates of the coupled inverters composed of four n- and p-type MOSFETs, the output is immediately determined according to the truth table. Therefore, this process requires only one unit of processing time. A small time delay occurs between the inputs and outputs mostly due to the RC delay of the circuit. At the same time, this operation requires at least four MOSFETs of which the layout can be integrated in a $\approx 40F^2$ area. In an ideal CMOS NAND circuit, no current flow should occur at a given state once the switching operations are completed, which is actually not the case owing to the gate leakage and source-drain leakage. The output of a NAND gate is directly connected to the input of the next gate. A MOSFET is an active device, meaning that the input voltage is not directly transferred to an output wherein output is always connected to either the drive voltage (V_{dd}) or ground through the channel of p-MOSFET or n-MOSFET. Thus, even noisy input is filtered at the output, and signal amplification can be achieved. When the power is turned off, the logic state is removed immediately. These are the critical characteristics of a NAND gate composed



two input NAND gate		
A	B	$\overline{AB}$
0	0	1
0	1	1
1	0	1
1	1	0

Figure 4. a) Schematic circuit diagrams of NAND and its truth table and b) schematic block diagram of full adder (FA).

of CMOSFETs. Next, how a similar NAND function can be constructed using memristor-based stateful logic IMP devices is discussed, and the schematic diagram and truth table, which are reproduced from Borghetti et al.,[6] are shown in **Figure 5**a and b, respectively. The structure consists of three memristors, which are referred to as P, Q and S, and their respective on and off states correspond to logic states 1 and 0. According to the truth table, p NAND q is equivalent to q IMP (p IMP 0), where p, q and 0 can be defined as the logic states of P, Q, and S, respectively. In this case, the inputs are applied to P and Q, and the output is achieved as the stateful logic state of S after the NAND operation composed of three steps is completed. Hence, the states of P and Q are defined first by applying the appropriate voltages sequentially to P and Q wherein the other memristors remain floated. The first step of the NAND operation is performed by applying a clocking voltage of $V_S = V_{clear}$ to S while the P and Q memristors remain floated, turning the S off. At the second step, $V_P = V_{cond}$ and $V_S = V_{set}$ are applied to P and S while Q is floated; the state of S is changed accordingly to the input data of P and it is described by the second truth table in Figure 5b. For the third step, $V_Q = V_{cond}$ and $V_S = V_{set}$ are applied to Q and S while P is floated, which changes the logic state of S again with respect to the input data of Q, and it is described by the third truth table in Figure 5b. Finally, the logic state of S, corresponding to "s", is read out using a read voltage. Therefore, it can be understood that a total of six sequential clocking voltage steps (two input steps + three steps of logic operation + one read step) are required to execute one NAND operation, and this is a critical drawback of such method compared to the one-step NAND operation in a CMOS-based NAND gate. This situation might be understood as follows: in the CMOS NAND gate, the logic operation is accomplished by changing the state of the n- and p-MOSFETs located along the chip surface, meaning that the logic data are transferred along the spatial dimension. For more complicated circuits such as the FA shown in Figure 4b, the same reasoning can be applied, i.e., the input data are applied at one end of the circuit, and the output is achieved at the opposite end along the spatial dimension despite taking a longer time than the simple NAND case due to the longer RC delay. In contrast, in the case of a NAND via the IMP memristor, the state of S changes with time during the operation in a three-step manner while the input is applied at the adjacent memristors. For this reason, the logic data flow along the spatial dimension as well as the time dimension, and the time domain has higher significance. To realize FA, memristor circuits can be used in several different ways. As shown in Figure 4b, the CMOS FA combines circuits of AND and XOR gates, each of which can also be constructed using IMP operations in the form of p AND q = (p IMP (q IMP 0)) IMP 0 = (p NAND q) IMP 0, and p XOR q = (p IMP q) IMP ((q IMP p) IMP 0), respectively. Although these circuit implementations have certain complications, as will be discussed shortly, they can be achieved by memristors nonetheless. Examining the equivalence of the AND circuit, one extra IMP operation step is included between S (the output of p NAND q) and 0. Hence, two approaches exist for how the AND is implemented by the memristor circuits: one is to use a fourth memristor cell T to input 0 ($t = 0$), corresponding to the spatial arrangement of the circuit elements, and the other is to re-use the same memristor cell after the first p NAND q operation is completed. In this case, the output of the first NAND operation S is stored in a separate memory, and S is reconfigured to 0. The data from the previous NAND are input to P (or Q) and the subsequent IMP operation between P (or Q) and S (of which logic state is now 0) results in the final output of AND, which is now stored as the logic state of S that must be subsequently read out. As a result, this process can be understood as the allocation of logic functionality along the time dimension. For FA implementation,

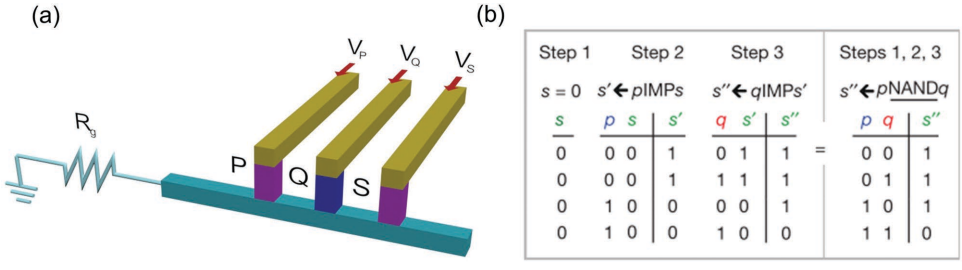


Figure 5. Memristor-based stateful logic IMP devices. a) Schematic diagram of the device and b) truth table showing the sequential data processing results. Reproduced with permission.[6] Copyright 2010, Nature Publishing Group.

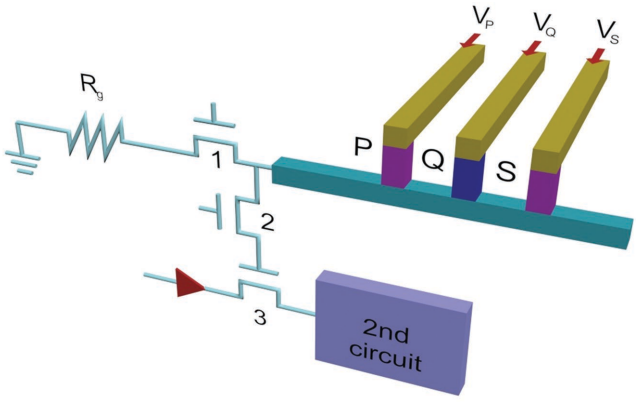


Figure 6. An exemplary configuration of the CMOS-stateful logic array for delivery of the output signal to the next operation without signal loss. This configuration corresponds to spatio-temporal-type data processing. Adapted with permission.[6] Copyright 2010, Nature Publishing Group.

another set of memristors must be available to calculate Carry. If the previous set of memristors were used to calculate the Sum, it must be re-configured using the same set of memristors. Compared to the purely temporal circuits, perhaps, the spatio-temporal combination of the memristors results in a more practical implementation of the IMP circuits.

Two main problems exist for spatial IMP circuits. As previously mentioned, the memristor is a passive two-terminal device, and thus the signal decay along the long logic chain must be considered, which is not necessarily the case for the temporal circuit. Therefore, signal amplification is necessary. Another problem is that the logic state of the output (S in the above examples) must be transferred to the next logic circuit by different voltage step. One solution to these problems is to adopt MOSFETs, as shown in **Figure 6**, where three n-MOSFETs are used between the parallel memristors and load resistance (R_G). For logic operation of the memristor circuit, MOSFET 1 is turned on, and MOSFET 2 is turned off such that the next stage of the logic circuit is effectively isolated from the circuit of interest. Once the logic operation is completed and the output is assumed to be transferred to the next stage of the circuit, MOSFET 1 is turned off, and MOSFET 2 is turned on. With an appropriate read voltage, sufficiently low to not disturb the logic state of S at the read step but high enough to turn on MOSFET 3, the data of S can be transferred to the next stage with appropriate amplification. This situation is related to the spatio-temporal arrangement of the memristor circuits, and it requires the combination of a CMOS circuit with memristors. As discussed in the previous section, the CMOS-memristor combined circuit is a feasible contender with FPGA, which is also applicable in this case.

Utilizing CMOS circuit would also aid in the cases of the XOR mentioned above combined with the EQUAL operations, such that CMOS memory can provide the necessary backup data when p and q inputs are needed twice in one operation, which would be invalid without additional data back-up sequence.

Another critical difference from the CMOS circuit is that the input data can be modified by the IMP operations for the NAND operation whereas in the CMOS logic circuits, the output has no influence on the inputs. It is unclear whether this situation poses a critical drawback at the moment. Nevertheless, the critical merits of such memristor-based logic gates are the use of two-terminal devices with limited numbers as it can effectively decrease the chip area, and data non-volatility lowering power consumption. The memristors can be integrated in a crossbar structure in which one memristor can occupy $4F^2$; accordingly, three memristors can be integrated in $12F^2$ which is ≈1/3 size of the CMOS NAND gate. Even taking additional transistors into account for signal amplification and delivery, the compactness of the area corresponds to the primary advantage that partially compensates for the disadvantage of complexity in the operational sequence. The effective area per device can be further reduced if the crossbar arrays are stacked, and this is quite difficult to realize with CMOS gate arrays. Likewise, memristor-based stateful logic can be reconfigured both spatially and temporally, depending on the specific requirements, which will increase efficiency in general applications compared to the CMOS logic gate. In standard CMOS logic, the number of specific logic gates are predetermined with respect to the optimized design, and thus the logic gates assigned for a specific logic operation cannot be used for different processes. Also, it will be idle during those periods, decreasing the computational and energy efficiency. In that case, it is important to determine the appropriate number of each logic gate depending on the purpose of the CMOS processor. This problem in the CPU can be partly overcome by implementing a GPU. Previously, the CPU performed all computational tasks, including graphic data processing mostly composed of specific mathematical formulae and transformations. Those graphics-related computations could be more efficiently handled in a GPU since it is dedicated to graphics-related processing. However, the GPU cannot replace the CPU for other purposes. On the contrary, in stateful logic, logic operations are performed by a temporal combination of signals, and it is possible to perform various operations at any logic gates, in principle. This aspect of IMP circuit might be seen as a type of an FPGA-type processor, but the new programming of the logic gates occurs in the time dimension as well as the spatial dimension. **Figure 7**a and b displays the conceptual diagrams for spatio-temporal assignments of logic gates and their operations for stateful logic and CMOS logic gates, respectively. Overall, the higher density associated with the crossbar platform in regard to the potentially higher working efficiency of stateful logic constitute a good rationale for further intensive research in this field, despite many remaining obstacles in the design architecture and material processing areas. It is expected that to perform the aforementioned operations in the crossbar memristor array, CMOS-based control parts that can determine the operation sequences, optimize the operation procedure, and address/assign the inputs and outputs appropriately are required. Therefore, not only reliable integration of the memristor crossbar array but also design of the control CMOS circuit are expected to be important components of this research. To this end, the recent suggestion of a CMOL configuration in which the memristor array is located on top of the back-end portion of CMOS layer is a great advancement for this direction.[42]

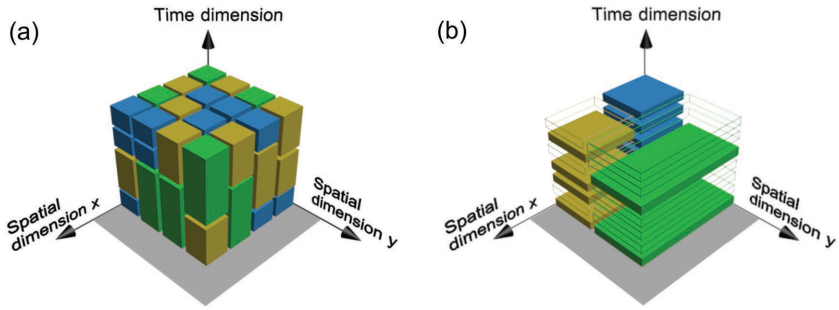


Figure 7. Conceptual diagram for spatio-temporal assignments of logic gates and their operations for a) stateful logic and b) CMOS logic gates. Different colored regions represent different operations.

2.3.4. Alternatives

Another type of stateful logic using CRS devices has been suggested by Linn et al., who originally reported CRS ReRAM for a sneak-leakage free crossbar array (CBA).[46,47] CBA memory based on CRS is gaining less attention at the moment because it might be less feasible for achieving sufficiently high integration density due to the relatively low on-to-off current ratio of the reported CRS devices.[48] The CRS-based logic circuit is currently gaining interest, however, and the configuration is quite CMOS-like, i.e., the parallel (anti-parallel) and serial (anti-serial) combinations of the directionality of the memristors composing the CRS are quite reminiscent of the various combined configurations of CMOSFETs.[46] The threshold-like *I*–*V* characteristics of the CRS cell are especially preferable for achieving CMOS-like logic functionality. Details for this CMOS-like implementation of memristor logic circuit can be found in the monograph by Vourkas and Sirakoulis.[45] However, these CMOS-like circuits (NAND, NOR, and NOT) require separate steps for input, logic operation, and reading of the output and also need rather complicated drive circuits and switches to discriminate among the different stages of the logic operations, which are the critical drawbacks of such an implementation. This situation occurs due to the CRS (or memristor) being a two-terminal device, but CMOSFETs are three-terminal devices, and mimicking the three-terminal device using two-terminal devices requires supplemental devices and operational steps in the circuitry, which was also the case for the aforementioned memristor's IMP-based logic circuit. Furthermore, the situation creates problems related to the spatio-temporal allocation of the logic functionalities mentioned above. Nevertheless, interest in this field is increasing owing to the critical merits of low power consumption and small cell size. One example is the stateful logic operation using SiO_2-based unipolar memristors that Zhou et al. demonstrated,[49] which can be operated by either of the bias polarities. It must be mentioned that material reliability issues (which are a large concern for any memristor materials reported to date) must be addressed, although it is not yet clearly understood how high the reliability of the memristor needs to be for reasonably operating nonvolatile logic circuits. A fundamentally different approach using the exciting functionality of the memristor lies in the exploration of a computing paradigm that is completely disparate from the von Neumann architecture and is the topic of the subsequent sections.

3. Neuromorphic Computing

This section is composed of five subsections: definition of neuromorphic computing, machine learning based on software implementation, neuromorphic circuit architecture, explanation on the advantage of neuromorphic system over the conventional von Neumann system, and finally outlook for memristive neuromorphic systems. This topic is viewed from cross-disciplinary standpoints, e.g. neuroscience, computer science, electrical engineering, and materials science, such that different approaches can be adopted. Generally, a neuromorphic system has a hierarchical structure ranging from building block (neuron and synapse) that form a small network to large network of such small networks. In this case, a building-up principle is of great importance, which is strongly related to learning and recognition algorithms. Given this hierarchical structure, emphasis perhaps differs for different disciplinary standpoints. Progress in neuromorphic computing algorithm may be justified by means of readily available CMOS components (conventional building blocks), which may correspond to a top-down approach. This approach offers great efficiency if the building blocks are suitable for the algorithms such that they do not impose constraints on algorithm realization. Otherwise, the algorithm should compromise with the building blocks. Alternatively, building blocks can be emphasized which renders diverse building blocks available. This may consequently enrich neuromorphic computing algorithms, particularly, algorithms customized for such new building blocks. In this regard, this section mainly overviews recent attempts to seek new building blocks for neuromorphic systems and several customized learning protocols, which are believed to enrich the neuromorphic research field alongside the conventional CMOS-based building blocks.

3.1. Definition of Neuromorphic Computing

The first proposal for neuromorphic engineering dates back to the late 1980s and originally covered the very-large-scale integration (VLSI) implementation of silicon-based analog circuits that function as the building blocks of a spiking neural network (SNN) and conduct recognition tasks on the network scale.[50,51] Nowadays, the scope of neuromorphic engineering has been widened to encompass hybrid analog/digital circuits and even fully digital circuits such as FPGAs,[52–54] which enable recognition tasks. Neuromorphic computing implies computation for recognition tasks achieved by neuromorphic systems that are distinctive from the conventional von Neumann computing architecture. Recognition in this context usually means pattern recognition. The data subject to recognition can be classified as several groups (patterns) according to correlations among each datum in one set of data for a specific recognition. Furthermore, the classified data can have semantic labels as a consequence of supervised learning. Handwritten digit recognition is a typical example of pattern recognition; humans are able to recognize

digits in spite of large variability in their shape because spatial correlations among the digits in the same group that have unconsciously been learned. The details of unprocessed raw images have relatively low significance in recognition. Neuromorphic computing thus aims at realizing such a recognition capability using neuromorphic circuits that are highly inspired by human brain's activities. Visual object recognition is one main subject that a number of neuromorphic engineers have attempted to implement using neuromorphic circuits, and thus, has been relatively well developed, particularly its front-end transducer, i.e., silicon retinas.[50,55,56]

Neuromorphic systems should differ in learning algorithm and circuit for different recognition tasks: handwritten digit (and still cut), moving visual object, and natural language recognitions. The first task, such as handwritten digit recognition, deals with time-invariant (static) input data, so that working memory is unnecessarily involved in the system. By contrast, the other two tasks need to encode time-variant objects in time domains, which may need working memory. Additionally, essential to neuromorphic system design is the learning scheme: either online or offline learning. The former generally means real time learning, whereas the latter distinguishes learning and recognition phases. Neuromorphic engineering attempts to offer solutions to diverse recognition tasks and eventually a universal solution as a platform engine applicable to various tasks.

3.2. Software-Based Machine Learning

For the moment, it is worth introducing software-based approaches to recognition tasks within the conventional von Neumann architecture in attempt to address the current status (including pros and cons) of (hardware-based) neuromorphic computing in comparison with its software counterpart. Deep learning is a type of a machine learning based on a virtual deep neural network (DNN) encompassing hidden layers between the input and output layers.[57,58] **Figure 8**a illustrates a schematic diagram of a DNN with a few layers. In this diagram, neurons are wired to other neurons. The input information is highly abstracted through several layers until a decision is eventually made in the output layer. Figure 8b illustrates the unidirectional signal transmission from a presynaptic to a postsynaptic neuron.

Deep learning is a fully software-based technology within the von Neumann framework and is thus distinguishable from the hardware-based neuromorphic systems. The DNN works as a universal mathematical hypothesis that is able to classify multidimensional input data as proper groups given their correlations.[8] The hidden layers offer nonlinear classification boundaries, rendering the DNN universal.[8] The DNN generally consists of binary neuron and analog synapse; the former makes a decision on its output, either '1' or '0', and the latter defines weight for the connection between neurons. The most commonly used type of neuron in DNN is a sigmoid neuron that is described by a sigmoid function. This construction outputs '1' if the sum of inputs exceeds zero to a certain extent and '0' if the sum falls below zero (also to a certain extent). The DNN includes several subsets that are distinguished according to architecture or learning algorithm, such as the multilayer perceptron network,[59] convolutional neural network,[60,61] and deep belief network.[58,62]

Learning (training) in deep learning adjusts the weight (model parameters) for classification. Deep learning depends on a fully mathematical learning algorithm, e.g., the backpropagation algorithm[57,60,63] and energy-based model.[64] The backpropagation is an error correction algorithm that is most widely deployed in various recognition problems. This algorithm minimizes the difference between the expected (desired) and actual output values. The weight update sequence (from the output to the input layer) is opposite to the information transmission through feedforward connection such that this algorithm is referred to as backpropagation.[65]

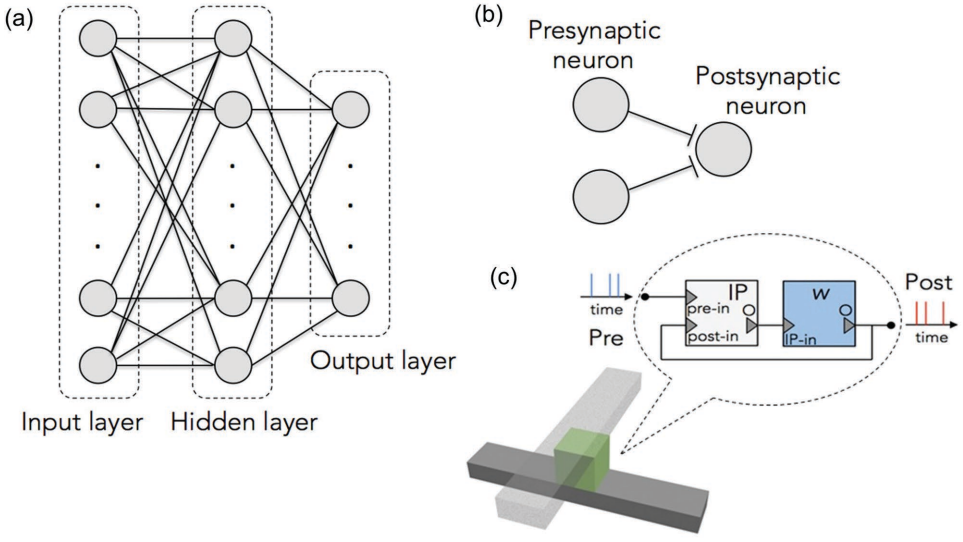


Figure 8. a) Schematic diagram of a neuronal network with a few layers. b) Unidirectional synaptic transmission from a presynaptic to a postsynaptic neuron. c) Artificial synaptic device including IP block and synaptic weight memory block used in the crossbar structure.

Deep learning requires a substantial amount of computer resources. The requirement becomes more severe as the DNN includes more neurons and synapses, which substantially increase the runtime. A solution to this computational inefficiency is to reinforce parallel computation using GPUs. A GPU contains thousands of cores designed originally to process graphics data in parallel, and thus it is compatible with, particularly, convolutional neural network.[66] GPUs customized for deep learning have already been commercialized.[8] Nonetheless, it should be noted that the improvement in computational efficiency minimally reduces the power consumption, and the energy efficiency of the GPU tends to decline with performance.

3.3. Neuromorphic Circuit Architecture

Neuromorphic computation produces neuronal processing that endows the electronic replica (SNN) with recognition capability as a consequence of training with a given experience. Analogous to the software-based DNN, the SNN provides a universal mathematical hypothesis that classifies multidimensional input data as groups. However, a significant difference lies in the fact that the SNN conducts analog computing (rather than digital computing as for deep learning) that is fully supported by physical phenomena. Therefore, the runtime barely scales with the network size, so that the SNN most likely outperforms the DNN when dealing with large networks.

As such, the SNN consists of spiking neurons and synapses. The spiking neuron produces a spike train or burst that is equivalent to a bit stream (e.g., 000010010101...) given the all-or-nothing property of a spike, i.e., '1' when spiking and '0' otherwise. Such spike firing occurs only if the summed input exceeds the threshold for firing, whereas the neuron is silent otherwise. The spiking neuron is analogous to the binary neuron in the DNN with regard to the threshold output behavior.

The DNN allows unidirectional signal flow through a single synapse, and the same holds for the SNN. This unidirectional signal transmission reflects the nature of a chemical synapse and is distinguishable from an electrical synapse that allows bidirectional signal flow.[67] The unidirectional information transmission in biological system is attributed to the asymmetry of the chemical synapse in which neurotransmitters are released from the presynaptic neuron (exocytosis) and are received by receptors only on the postsynaptic neuron. Similar to the synapse in the DNN, that in the SNN has a weight value (synaptic weight) to remember. The synaptic weight determines the excitatory postsynaptic current (EPSC); the higher the weight, the more probability that the spiking of the presynaptic neuron evokes postsynaptic spiking.

To implement a learning protocol in a neuromorphic system, the artificial synaptic device or circuit should include at least two components, as illustrated by the block diagram in Figure 8c. The induction protocol (IP) block outputs a signal to the synaptic weight memory block (indicated using w in Figure 8c), depending on the presynaptic and postsynaptic spiking conditions. A silicon VLSI synaptic circuit frequently uses an analog circuit in the IP block to evaluate an appropriate change in synaptic weight under the given spiking conditions, and the result is stored in a capacitor (thus, w is an array of capacitors in this case). The IP circuit design varies depending on the implemented learning protocol. A significant challenge from a materials standpoint is achievement of these two separate functional blocks in a single two-terminal synaptic device that is capable of synaptic weight evaluation and long-term memorization. To this end, the IP block realized by an analog circuit in a silicon synapse must be hosted in a two-terminal synaptic device at the atomic scale (the crossbar structure in Figure 8c).

The SNN needs hidden layers to represent nonlinear classification boundaries as for the DNN. The basic structure of the SNN is analogous to the DNN (Figure 8a and b). In the SNN, each wire in Figure 8b denotes a lumped axon and synapse. Note that the intra-layer wiring between neurons must also be implemented to form a recurrent network.[68] A winner-take-all network with full inhibitory neurons wired to each other is an example of a recurrent network.[68]

The following sub-sections (3.3.1–3.3.4) separately deal with the artificial spiking neuron (3.3.1), SNN-compatible learning algorithm in offline format (3.3.2), SNN-compatible learning algorithm in online format with an emphasis on artificial synapse (3.3.3), and network design (3.3.4) in a neuromorphic system.

3.3.1. Artificial Spiking Neuron

Within the framework of SNN, the external input is encoded into a train or burst of spikes, and the spikes propagate to the topmost layer (i.e., the output layer) through hidden layers in between. Neuronal ensembles (population) dynamically represent their states via spiking dynamics, which are referred to as attractor networks.[69] Thus, implementing the "appropriate" spiking neurons is of significant importance in the SNN-based neuromorphic system. The spiking neuron must meet the following requirements: i) integrate-and-fire (I&F) behavior, ii) low power consumption, iii) low spiking rate, and iv) active operation.

The I&F behavior is a prototypical framework of artificial spiking neuron design and denotes input signal integration until the integrated level (membrane potential) reaches a threshold for spiking and consequent spiking when the threshold is reached.[69] Both the input and output of the I&F neuron consist of a train/burst of spikes that is equivalent to a bit stream (e.g., 000010010101...). Because spiking occurs in the time domain, each bit in the bit stream represents spiking-or-nothing within a given time bin. The length of the bit stream, therefore, corresponds to the input or output time, and the number of '1's divided by the bit stream length is equal to the firing rate or neuronal activity that is primarily taken as the input or output quantity in the framework of leaky I&F (LIF).

The integrator in the I&F neuron counts the number of '1's and compares it with a threshold value in an attempt to spike when the threshold is exceeded. The interval between neighboring '1's is referred to as the inter-spike interval (ISI). The LIF is a neuron design framework with higher fidelity to the biological neuron.[68–70] As the name implies, the integrator is leaky such that the integrated level at a given moment decays with time with no incoming spike in close succession. This LIF behavior realizes a dynamic (time-dependent) integration procedure. Therefore, the absolute number of '1's rather than the firing rate (i.e., how many '1's per unit time) is a meaningful variable in this case.

Power consumption is significantly concerned to neuromorphic engineers, so that the artificial spiking neurons must consume as little power as possible. To accomplish this goal, one of the popular strategies for CMOS-based spiking neurons is to maintain operation of MOSFETs in the subthreshold regime, which markedly reduces power consumption compared with above-threshold operation.[71–73] However, this strategy reduces operational fidelity due to the involvement of inherent random noise, e.g., white noise such as thermal and shot noises and non-white noise such as flicker and burst noises.[71,72] Nevertheless, such noise effects perhaps offer the possibility of stochastic computing if the correlation between the noises is not established.[71,74]

Additionally, because each spike generation requires a certain power, the lower neuronal activity consumes less power. Thus, the optimal design of a spiking neuron might include limiting the maximum neuronal activity below approximately 100 Hz (comparable to the biological neuron), particularly if the neuromorphic system is aimed at real-time interaction with physical environments. However, this design strategy faces a severe obstacle in scale-down of the size of a neuronal circuit. The leaky integrator of an LIF neuron is commonly based on a capacitor-based low-pass filter whose cutoff frequency is determined by the RC time constant.[75,127] To integrate incident spikes at a given neuronal activity (<100 Hz), the cutoff frequency should be less than the neuronal activity, and thus the time constant must be remarkably large. This situation accordingly requires large capacitor area, which seriously hinders scale-down of the neuronal circuit. A workaround for this issue may be to replace the capacitance-based integrator by a floating gate integrator as proposed by Kornijcuk et al.[72]

Moreover, a spiking neuron should essentially be *active*. The biological neuron is capable of active operations due to pumps for sodium and potassium ions[76] that are embedded in the bilipid membrane. The ion pumps consume chemical energy to store electrical energy, i.e., voltage, across the membrane in the resting state, meaning that the neuron has a power reservoir.[67] The same situation should apply for its electronic replica, otherwise spikes generated at a neuron cannot propagate through successive postsynaptic neurons due to signal dissipation. Therefore, the use of active devices such as MOSFETs is perhaps unavoidable, which is a quite similar circumstance to the circuit design for the stateful logic discussed in Figure 6.

Although a simple two-terminal passive device such as the memristor cannot serve as an entire replacement for a CMOS neuronal circuit, such a simple device can be used in a neuronal circuit in an attempt to alleviate the large area overhead of the fully CMOS-based neuronal circuit. To date, a few attempts have been made to achieve this objective.[30,77,80] Essentially, the two-terminal device based on functional materials should not represent the memory effect because a neuronal operation is minimally dependent on history. Additionally, the two-terminal device should possess two distinctive (resistance) states, and the state should oscillate between them when excited. This oscillation behavior is converted to oscillation of the output voltage, which resembles spiking behavior. The oscillation frequency must vary upon an input signal to make use of neuronal activity as a state variable.

A potential approach is based on the relaxation oscillation achievable via the Pearson-Anson effect.[30,74,77,80] A key component of this effect is a threshold switch that represents monostable resistive switching accompanied by the S-shape NDR effect.[30,74,78,79] Threshold switching behavior has been observed in a wide range of materials systems, such as higher chalcogenides,[80–82] Mott insulators,[30,79] and Shockley diodes.[83] Recently, Pickett et al. proposed an LIF neuron model using a pair of Pearson-Anson oscillators referred to as a "neuristor."[30] Each oscillator includes an NbO_2-based Mott insulator sandwiched between inert Pt electrodes that exhibits threshold switching with respect to an applied voltage, which in turn leads to a change in lattice temperature crossing the critical temperature. The dynamics of the neuristor-based LIF (NLIF) neuron model are detailed in a phase-plane analysis by Lim et al.[78] The two constant voltage (pull-up voltage) sources in the NLIF neuron circuit are capable of active operation given the output voltage gain arising from the voltage sources. However, this scheme uses a threshold switch directly as a pull-up resistor, and the pull-up voltage must be close to the threshold for the S-shaped NDR to enlarge the voltage gain. Consequently, the pull-up voltage is most likely to cause significant reliability issues.[78]

Another approach was proposed by Krzysteczko et al. in which the instability of the magnetic configuration in a magnetic tunnel junction (MTJ) was exploited.[84] The oscillation of resistance through the MTJ was viewed as similar to neuronal spiking behavior. The oscillation in the MTJ is estimated to originate from the thermodynamic instability of the intermediate magnetic configuration mediated by the magnetic domain configuration in the free electrode.[84] However, the results remained at rather primitive level yet, and no proof-of-concept circuit was proposed to meet the aforementioned requirements for artificial spiking neurons.

3.3.2. Learning Algorithm with Memristor-Based Synapse (Offline Learning)

An emerging approach to neuromorphic engineering is to make use of a passive synapse array[85,86] as a replacement for silicon synapses each of which generally requires dozens of MOSFETs.[73,87] Each synaptic device in the synapse array is a passive two-terminal device that meets the design rule of $4F^2$ such that the passive synapse array has a remarkable advantage over mainstream silicon synapses in terms of synapse density (areal compactness). As such, each synaptic device has a synaptic weight to remember (memory) to control the excitatory current into the postsynaptic neuron circuit (i.e., EPSC). In this case, conductance or resistance is perhaps the suitable physical quantity for the synaptic weight, and therefore, resistance-based nonvolatile memory bits (e.g., memristor, phase-change memory, magnetic memory) are potential candidates. Likewise, these passive synaptic devices are highly scalable down to a few tens of nanometers in diameter, thus maximizing the areal density of synaptic devices (see Section 4). However, it should be conceded that this emerging approach is not as mature as the former mainstream in silico technologies at present time.

The simplest way to strengthen the above-mentioned advantage of memristors as synaptic devices is to replace the IP block in Figure 8c by external digital computing, which is often

termed as offline learning. Offline learning refers to training a machine in a learning phase that is separate from a real-time decision-making phase, so that the machine is unable to learn in real time. An advantage of such offline configuration is that the proper synaptic weight values can be predetermined through the learning phase by deep learning algorithms such as the backpropagation algorithm,[88,89] and an energy-based model for the Restricted Boltzmann Machine.[90] Therefore, the IP block is simply replaced by digital calculation, which remarkably mitigates the circuit area overhead. In this regard, the memristor-array merely works as a look-up table that remembers a synaptic weight value for each synaptic connection. The conductance of each memristor in the array is adjusted to the preset value. Essential to this end is the analog-like representation of memristor conductance, good memory retention, and minimal crosstalk between passive cells.

3.3.3. Learning Algorithm with Memristor-Based Synapse (Online Learning)

A learning (training) process corresponds to adjusting the synaptic weight values in a neural network to classify similar input patterns that share essential features with the training datasets. In contrast to offline learning, online learning does not endow the learning machine with a separate learning phase; instead, learning occurs in real-time, interacting with the physical environment. A neuromorphic system as a standalone learning machine essentially needs the IP block (Figure 8c) and the learning principle achieved by the IP block. Frequently, neuromorphic engineers benchmark the Hebbian learning rule proposed by D. Hebb in 1949.[91] The Hebbian learning rule generally covers two protocols: activity-dependent plasticity (ADP) and spike timing-dependent plasticity (STDP).[92–98]

ADP means the change in synaptic weight with presynaptic and postsynaptic activities, i.e., firing rate. The induced change is maintained for a long duration, representing a long-term memory (LTM) effect. LTM represents long-term potentiation (LTP) and long-term depression (LTD), denoting an increase and decrease in synaptic weight under a particular circumstance, respectively. ADP is empirically described by the BCM rule (proposed by Bienenstock, Cooper, and Munro) that adopts a moving threshold for LTP to prevent unlimited growth of the synaptic weight.[92] ADP captures the Hebb's seminal hypothesis of "neurons that fire together wire together," implying that simultaneous spiking of neurons connected by a synapse causes an increase in the synaptic weight.[91]

STDP is another seminal learning protocol that relates temporal coding with respect to the relative timing between the presynaptic spike's arrival at the chemical synapse and postsynaptic spiking.[97–101] The presynaptic spike's arrival that precedes postsynaptic spiking leads to LTP, and the reverse time order leads to LTD. In neurophysiology, the implication of STDP is not clear at the neural network scale other than its role in reducing spike-firing latency[97] and short synaptic chain formation.[102,103] Phenomenologically, STDP bifurcates synaptic weight depending on causality; the synapse directly causing postsynaptic spiking is reinforced (potentiated), whereas others are ruled out (depressed). In fact, ADP and STDP appear to share the same universal feature of learning but are viewed from different viewpoints in different input domains, i.e., in the activity and spike timing domains, respectively.[95,96,104]

Employing the Hebbian learning rule, the neuromorphic system may be capable of real-time learning (online learning) without external digital programming. The Hebbian learning rules are customized to these two-terminal synaptic devices with abstraction and/or reinterpretation.[52,105] Particularly, STDP in abstracted and/or modified form is a widely adopted learning protocol. This section covers three approaches to memristor-based synapse realization within the framework of the Hebbian learning: i) memristor plus a CMOS IP block, ii) memristor subject to temporal overlap between presynaptic and postsynaptic spikes, and iii) memristor engineered on the atomic scale to learn without the temporal overlap.

A general strategy for artificial synapse adopts a mathematical formula for a learning rule using dozens of MOSFETs.[73,87] The evaluated synaptic weight under a given circumstance is stored in storage devices such as a capacitor[73,106,107] or floating-gate transistor.[87] For instance, a stored voltage across the capacitor, i.e., synaptic weight, is designed to gate a MOSFET that drives an EPSC through the channel towards a postsynaptic neuron circuit.[107,108] Implementing the synaptic weight in terms of a voltage across a capacitor simplifies the silicon synapse circuit design. However, the charge on the capacitor spontaneously decays in due course given the presence of finite gate leakage and off-state source-drain leakage of MOSFETs, and thus, long-term memory is not achieved solely by the capacitor.[73] Memristor is an alternative to these mainstream memory elements, which merely works as components of the w block in Figure 8c, as shown by Kornijcuk et al.[109] A main advantage of this strategy is such that well established learning protocols within the framework of CMOS-based neuromorphic engineering can readily be used given the separated IP block. Nonetheless, this case imposes substantial area overhead on the IP block shown in Figure 8c.

A similar approach was applied to phase-change memory.[110] STDP was successfully achieved without direct overlap between presynaptic and postsynaptic spikes with the aid of the axon driver based on CMOS. Instead, the preset voltage pulse produced by the axon driver overlaps with the postsynaptic spike in time such that the desired conductance state is programmed in the phase-change memory.[110]

The memristor encodes the height and duration of an applied voltage pulse into the consequent resistance state. The memristive synaptic device is popularly subject to presynaptic and postsynaptic spikes applied to the two different electrodes. Given the aforementioned unique characteristics of the memristor, a commonly used method for STDP implementation is to vary the voltage across the memristor depending on the spike timing. In this framework, the overlap in time between the presynaptic and postsynaptic spikes is essential to drive the resistance change. To this end, the spike shape must be engineered carefully such that the overlap of the spikes leads to the desired effective programming pulse duration and/or amplitude with spike timing. This approach requires the CMOS neurons to provide a spike of the desired shape. The shape of a spike must be customized to different memristive systems of different switching polarities and switching voltages.

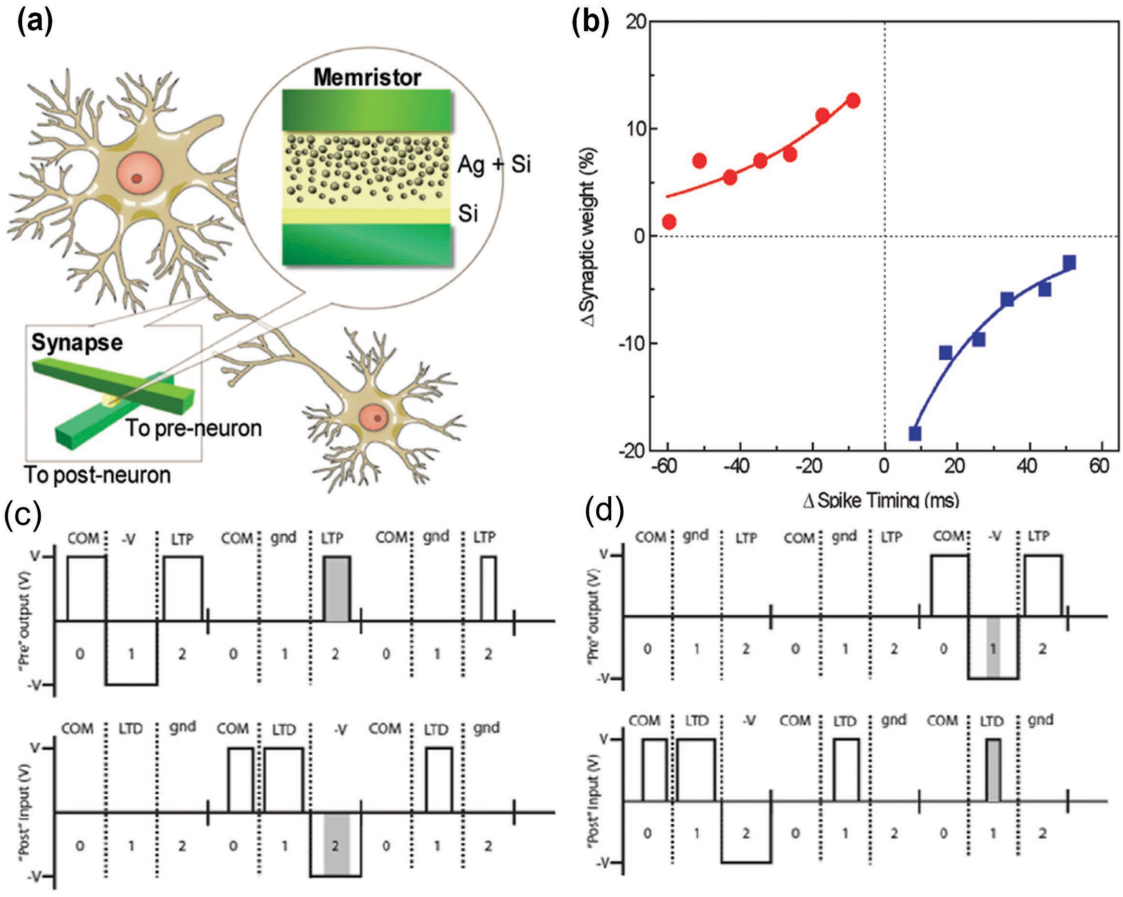


Figure 9. a) Schematic illustration of the concept of using memristors as synapses between neurons. b) Demonstration of STDP in the memristive synaptic device. The measured change of the synaptic weight (device conductance) vs. the relative timing of pre- and postsynaptic spiking. Examples of pre- and postsynaptic spiking patterns (interpreted) for c) LTP and d) LTD are shown. The overlap between potential at the pre- and post-terminals (indicated with a gray zone) exceeds the threshold for a resistance change and consequently induces resistance change. The spiking timing was interpreted as LTP- and LTD-evoking voltage pulse widths in this experimental design. Reproduced with permission.[137] Copyright 2010, American Chemical Society.

Figure 9 shows the STDP behavior in a cation-based memristive synaptic device, which was measured by applying the spike-overlap scheme.[85] This Ag-based memristive synaptic device represents a resistance state that varies for different programming voltage pulse widths for both set (LTP) and reset (LTD) processes.[15,85] In this experimental design, each spike consisted of a square voltage pulse with a height between $V_{th}/2$ and V_{th}, where V_{th} denotes the threshold voltage for resistance change. Therefore, a single spike was unable to induce a resistance change. The presynaptic and postsynaptic spiking patterns used in these experiments are displayed in Figure 9c and d. The overlap between the presynaptic and postsynaptic spikes endows the device with a voltage above the threshold, leading to a resistance change. A mixed analog-digital circuit working as the IP block generated a potentiating/depressing pulse across the memristor synapse when the presynaptic neuron spiked before/after the postsynaptic neuron. The feasibility of ADP behavior has been identified also in memristive synaptic devices in an attempt to describe learning process with respect to neuronal activity. The basic scheme for ADP implementation is akin to that for STDP, i.e., overlap between presynaptic and postsynaptic spikes that are particularly designed. Li et al.[94,111] and He et al.[95,112] recently demonstrated ADP in Ag/AgInSbTe/Ag and Pt/FeO_x/Pt, respectively, using this scheme. Such spike-overlap strategies have been employed in various memristive systems which include cation-based systems,[85,113] anion-based systems,[114–117] phase-change materials,[118,119] MTJs,[84,120] and ferroelectrics.[121] Various candidates for synaptic materials have been reviewed in recent review papers.[67,122,123]

However, such STDP induction protocols based on spike overlap are not completely aligned with the principles of neuromorphic engineering, particularly with respect to the requirement for sparsity of spikes, i.e., presynaptic spikes seldom overlap with postsynaptic spikes in time. At a low neuronal activity comparable to that of the biological neuron, the probability of spike overlap is notably low, and thus, the learning appears quite inefficient. This limitation on learning protocol is inherently attributed to the fact that the memristor's resistance is primarily determined by the internal state variable (conduction channel size). In some case, the size of channel is solely controlled by an applied voltage. This type of memristor is referred to as a first-order memristor.[59] Assuming good retention of the programmed resistance states, a state change occurs only if a voltage ($>V_{set}$ or V_{reset}) is applied to the memristor. This

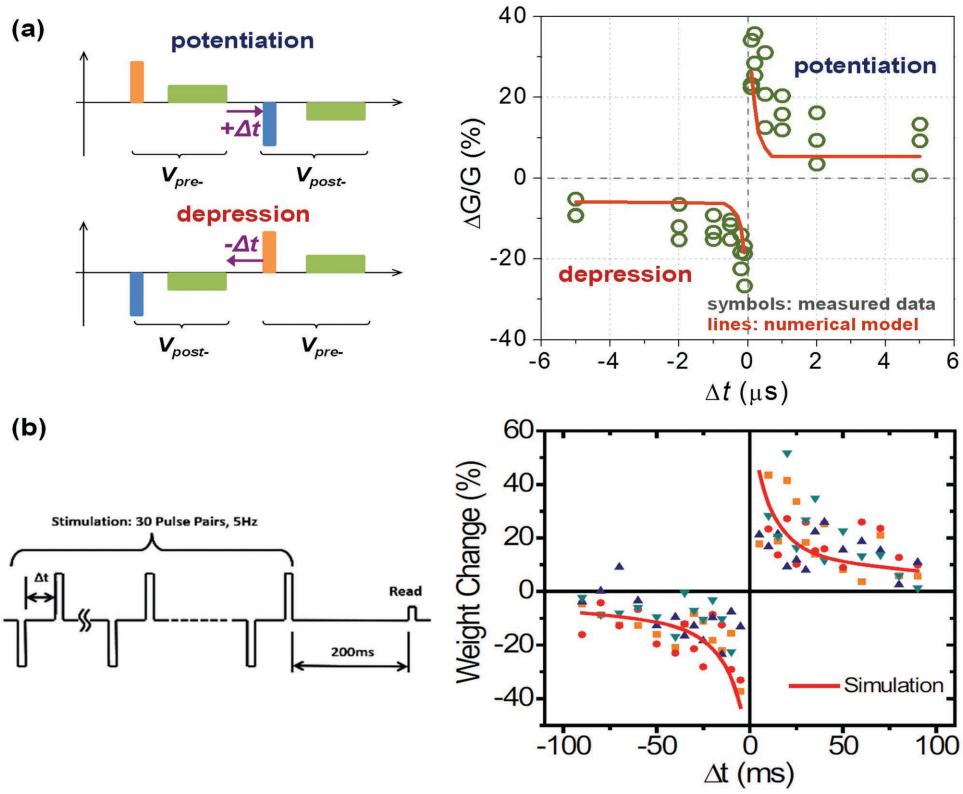


Figure 10. a) Implementation of spike-timing dependent plasticity in the second-order memristor. Non-overlapping pre- and postsynaptic spikes successfully lead to STDP results. b) Memristor weight changes as a function of the relative timing between the pre- and postsynaptic non-overlapping pulses. Reproduced with permission.[115] Copyright 2015, American Chemical Society.

observation underlies the spike-overlap scheme for the previously mentioned learning protocols. An alternative to this fairly limited scheme is perhaps to involve an auxiliary variable, i.e., the indirect cause of a state change but a long-lasting physical parameter between spikes in close succession. Consequently, even non-overlapping presynaptic and postsynaptic spikes might result in a state change via the interaction between the long-lasting auxiliary variable and spike. This type of memristor that involves an auxiliary state variable is referred to as a second-order memristor.[115,116]

Recently, the experimental demonstration of second-order memristor was reported in TaO_x-[115] and WO_x-[116] based memristors, which was captured by the STDP behavior induced by a non-overlapping pair of presynaptic and postsynaptic spikes, as shown in **Figure 10**. The second-order memristive effect in the TaO_x-based memristor is believed to be caused by the auxiliary variable (lattice temperature), the dynamics of which are responsible for the relatively long-term effect that fills the gap between non-overlapping successive incident spikes.[115] Joule heating is the direct cause of an increase in lattice temperature during the auxiliary voltage application, and the temperature decays when the voltage terminates. Fortunately, the decay rate is sufficiently low to induce a remanent effect on the conducting channel size within a spike timing range of a few microseconds.[115] A similar second-order memristive effect is also observed in a WO_x-based memristor.[116] The oxygen vacancy mobility is believed to be an auxiliary variable responsible for the remanent effect.[116] Given this second-order memristive effect, a wide range of synaptic plasticity behaviors, including pair-pulse facilitation, STDP, and ADP, were successfully demonstrated using simple non-overlapping spikes. These results offer probable compatibility of memristive synaptic devices with sparsely spiking artificial neurons, although the available activity range is still fairly high. In addition, the addition of voltage pulses with a lower amplitude (auxiliary variable) between the voltage spikes is not very compatible with the biological system, but it can be realized in artificial neuronal system with the help of periphery circuits.

3.3.4. Network Architecture for an Efficient Artificial System

Each neuron in a network is wired to a number of other neurons through axons such that the neuron device requires the same number of wires. An efficient method for direct wiring of neurons was proposed by Likharev et al. and is referred to as CMOL (See Section 2.1, discussions related to molecular electronics).[124] However, the use of simple two-terminal synaptic devices is required for the practical application of this network design technology. A sensible alternative for rather complex neuronal and synaptic circuits is to route a signal (spike) from a presynaptic neuron circuit (with its own address) to a target postsynaptic neuron circuit (also with its own address) using a digital communication protocol. This protocol is known as address-event representation (AER).[125,126] In this protocol, the

neuron circuits do not necessarily require physical contact via a synaptic circuit, and therefore, a neuron circuit block can be separate from a synaptic circuit block, rendering it efficient for design of a neuromorphic chip. The information in spikes from a vast number of neuronal circuits (e.g., neuron addresses and spike time labels) is sent through the same data bus by time-multiplexing the connections. The data are de-multiplexed, and each spike is transmitted to a target postsynaptic neuron. The synaptic weight for each transmission is listed in a look-up table that is temporally stored in a memory block, such as SRAM. Thus, each connection is endowed with an appropriate weight value for a component in the look-up table (e.g., presynaptic neuron address, postsynaptic neuron address).[105,106,125,126] The synaptic weight evaluation can be performed with a separate synaptic circuit block.

3.4. Advantages of Neuromorphic Computation over von Neumann Computation

In an "ideal" neuromorphic system, unlike in a digital computer, no CPU clock exists, which implies asynchronous operation of each neuronal circuit only when it is triggered by external input, i.e., it is event-driven. Event-driven operation offers a significant reduction in power consumption compared with the (almost) continuous power consumption of the CPU, which is controlled by a clock. Additionally, the neuromorphic system can operate at the real-time scale, i.e., neuronal activity (spiking frequency) similar to that of a biological neuron (<100 Hz), to reduce power consumption. Given that each spike costs power, the real-time scale outperforms the "accelerated-time" scale that implements higher neuronal activities in terms of energy consumption, particularly when the system interacts with its real physical environment.[127] Additionally, the real-time scale offers flexibility of system architecture. For instance, the aforementioned AER (time-multiplexing) protocol can be used effectively because the lower the neuronal activity the more neurons that can share a data bus without overloading it with the time-multiplexed addresses of the sending neurons. By contrast, modern digital computers based on the von Neumann architecture obviously differ from the brain with respect to the number of sequential data processing steps (>10^6) and the limited capability of parallel data processing (number of cores < 20 for high-end CPUs). Thus, such computers are not fully compatible with the brain in terms of architecture.

Additionally, neuromorphic architecture is known to offer relatively large tolerance for false operation of units, compared with the von Neumann architecture. This advantage is mainly attributed to the redundancy of neurons and their connections; units in false operation are backed up by some of redundant units such that no macroscopic false operation is caused. Interestingly, in deep learning, some neuron units are purposefully halted or removed during a learning process to avoid parameter overfitting, which is referred to as dropout.[66] However, the degree of false operation tolerance depends on the network structure including the number of neurons and synapses, task assigned to the network, and algorithm in use. Moreover, neuromorphic architecture is much more immune to noise (thermal, shot, flicker, and burst noise) unlike the von Neumann architecture. Rather, such uncorrelated noise endows the neuromorphic architecture with operational uniformity as theoretically predicted by Burkitt and Clark[128] and even creates a new functionality that cannot be achieved in the von Neumann architecture.[73] Therefore, the neuromorphic system generally can offer more energy-efficient and fault-tolerant architecture compared with conventional von Neumann computer for several recognition-oriented tasks.

3.5. Outlook for Memristive Synaptic Devices

Despite recent progresses made in memristive synaptic devices, practical implementation of high-density synapse arrays remains highly challenging due to many technical problems and the low maturity level of this research field. As addressed in Section 3.3.3, a standard learning algorithm for memristor-based neuromorphic systems is not available; instead, there exist a large number of algorithm proposals that were demonstrated in small memristor-based networks. Therefore, it is believed that a breakthrough in this research field can be made with an optimal algorithm that is suitable for memristor-based neuromorphic systems. Otherwise, the desired behavior of a single memristor largely differs for different approaches so that attention is divided.

An exemplary question is whether the memristive synaptic device should represent analog-type conductance or digitized conductance for neuromorphic applications. Generally, it is believed that the former is desirable because of the two main reasons: first, a vast number of neurophysiological studies on chemical synapses report an analog-type synaptic weight, albeit it is rather noisy; the second reason might stem from DNN. A DNN is commonly built using binary neurons and synapses that have analog-type weight values. However, regarding the first reason, it is widely accepted that the synaptic weight eventually bifurcates between the maximum and minimum values.[129–131] In other words, the intermediate synaptic weight values might be merely temporary, resulting from an initial few spike-pairs. Therefore, the analog-type conductance representation may not be a necessary condition for the learning process. An objection to the second reason is the learning algorithm with a binary synapse, where binary values rather than analog-type weight values can be adopted in the learning process.[132,133] These factors mean that a wide range of opportunities exists for memristive synaptic devices for various suitable learning protocols that unnecessarily need analog-type conductance.

Irrespective of learning algorithm, an obvious requirement that should be carefully considered for the realization of high-density synapse arrays is low power consumption as repeatedly stressed. The rule-of-thumb calculation reveals the following approximate power consumption of a single synaptic event in the human brain. Among the 10^{14} synapses in the brain, approximately 1% are simultaneously active at a given time.[134] Additionally, each neuron produces spikes at a frequency of ≈10–100 Hz on average, and the total power consumption of the brain is approximately 10–20 W. Therefore, the power consumption per synaptic event is estimated to be ≈10^{-11} W. Given that each synaptic event has a duration of ≈100 ms, each synapse consumes an energy of ≈1 pJ per synaptic event. Meeting this energy consumption level using the current

memristor technology requires the following operational conditions based on the rule-of-thumb calculation. The energy consumption of a memristor per synaptic event (E) can be calculated by multiplying the programming pulse amplitude (V) with the current through the device (I) and the programming pulse width (t) ($E = V \times I \times t$). With a programming voltage of ≈2 V and a programming time in the range 10–100 ns, the current should be in the range 5–50 μA to meet the biologically comparable energy consumption level. This requirement is intended merely for a single synaptic event, and the number of events per second is proportional to the spiking rate. Therefore, at the accelerated time scale, achievement of biologically comparable power consumption becomes quite daunting. Moreover, the power consumption of spiking neurons should also be considered, which makes the requirement even more severe.

Additionally, a significant issue that should be urgently addressed for high-density passive synaptic arrays is the sneak current arising from its parallel connection structure.[26,135,136] Such sneak current issue has been significantly dealt with in crossbar ReRAMs,[26,135,136] and the same problem applies to synaptic arrays. A promising solution to this problem is the use of a passive selector in series with the memristor, which endows the chosen bit with high selectivity. **Figure 11** shows the representative neuromorphic system, which was developed recently to emulate the cognition of characters z, v, and n using the combined CMOS neurons and memristive synaptic crossbar array, where the Al_2O_3/TiO_{2-x} bilayer plays the critical role (see Section 4.1).[137] These researchers stated that the uniformity and reliability of the synaptic crossbar array enabled such system to work fluently.

4. Improvements in the Materials Aspects

For stateful logic and neuromorphic/analog-type computing applications, diverse memristive materials can be used.[85,86,138,139] It is understood that memristive switching in oxides is related to the redox reaction process that competes between oxidation and reduction of the conducting channel, which might easily and successfully emulate several critical aspects of the biological systems. More specifically, it is believed that the conducting filament (CF) plays an important role in most functionalities of the memristor. This section deals with material and processing aspects of memristors. In fact, these aspects have been extensively studied for ReRAM applications, and thus, this section selectively refers to the studies which aimed at the neuromorphic applications and the memory applications that could be used for the new computing applications. The three-dimensional integration of the memristors, which is of utmost importance for the ReRAM application, is also reviewed as it will be a critical ingredient for synaptic applications. Although the discussion could slightly overlap with the previous sections for some aspects, the main focus in this section is on the materials and their processing. Extensive reports for the neuromorphic application of the memristors were already made, but those for the stateful logic were quite rare as it has a much shorter history than others, which makes the following discussion meaningful as the stateful logic application is similar to that of ReRAM.

The unproven reliability of memristor materials is one of the significant concerns about the usefulness of the memristive materials for the new computational applications. Nonetheless, there are several notable reports about the quite extensive reliability improvement, such as Ta_2O_5 memristor.[140,141] It is also noted that if the non-volatile logic or stateful logic is realized, the number of necessary switching cycles per a device will be significantly decreased, which is what the stateful logic is about as the energy-saving device. Nevertheless, it cannot be claimed that the memristor can show immediate possibility to compete with the existing CMOS-based solutions in terms of the material reliability at this moment.

4.1. Memristive Materials for New Computing Paradigms

Memristive materials are generally categorized into anion-migration-based and cation-migration-based resistance

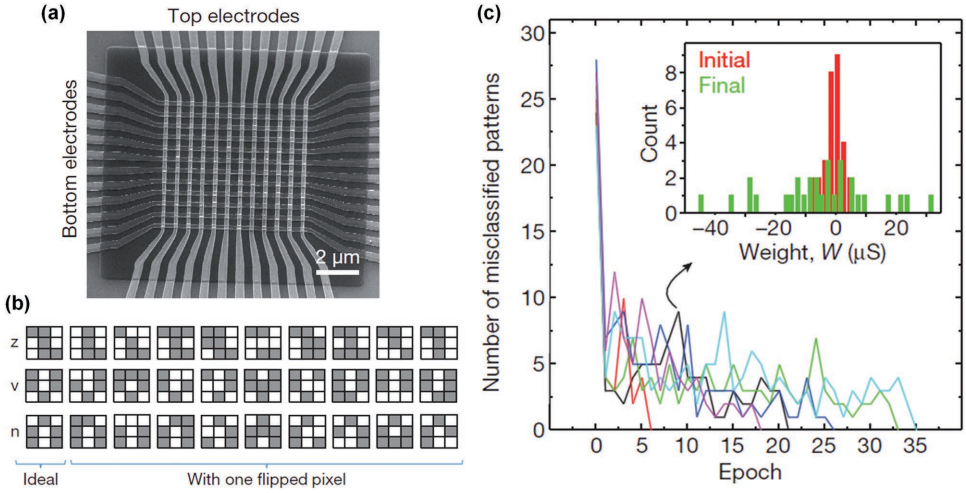


Figure 11. a) Integrated 12 × 12 crossbar with an Al_2O_3/TiO_{2-x} memristor at each cross-point. b) Input pattern set of 3 × 3 binary images used for the pattern classification experiment. c) Pattern classification experiment results: Convergence of network outputs during the training process to the perfect value (zero). Reproduced with permission.[137] Copyright 2015, Nature Publishing Group.

switching materials.[22,27] Numerous memristive materials, such as binary transition metal oxides,[115,137,142–148] and multinary complex oxides,[149–152] have been extensively studied and can be placed into the anion-migration-based memristive materials category. Cation-migration-based memristors include an active electrode-containing diffusive metal (Ag, Cu, or their alloys) with a solid electrolyte, such as chalcogenides,[138,142,145,153–156] amorphous silicon[85,142] or other ionic conductors.[143,157–159] Despite the prevalence of the ionic resistive switching system described above, the electronic resistive switching system can also be included in memristive materials.[160,161] In the electronic switching system, memristive switching phenomena rely on carrier (mainly electron) trapping/de-trapping at the defective sites. Defective TiO_{2-x}[160,161] or a bilayer of Ta_2O_5/HfO_2[162] could exhibit such bipolar-type electronic switching phenomena. Chalcogenide-based phase change materials were also investigated in this regard.[89]

In the case of the anion-migration-based memristor, the formation of a reduced phase or metallic channel (or metallic precipitation) was revealed in several transition metal oxides, including TiO_2, VO_2, WO_3, NiO, ZnO, and TaO_x.[163–171] The Ag or Cu dendrites or precipitations have been observed in a solid electrolyte in the case of a cation-migration-based memristor.[142,172–174] Such conducting phase is the key to maintaining the resistance state, i.e., the logic states in stateful logic circuit and neuronal function of LTP. It is believed that the conducting phase significantly reduces the restoring force, which implies the increase of resistance with elapsing time due to ionic relaxation or oxidation, back to the high-resistance state. It is interesting to note that the formation of a Ca^{2+} ionic channel is crucially involved in LTP in a biological system via release of neurotransmitters from a presynaptic neuron.[67]

Molas et al. reported that an Hf dopant in the metal oxide electrolyte layer for Cu ion migration could reduce the formation voltage at the cost of a reduced ON/OFF ratio, whereas a small concentration of Al dopant could improve the memory window and thermal stability.[175] Similarly, an oxygen vacancy could be created in the doped electrolyte layer, leading to easier Cu migration while the remaining defects prevent recovery of the pristine state during the reset operation. In addition, chalcogenide (Ag_2S),[138,155] Ag-doped amorphous Si (a-Si:Ag),[85,142] and other electrolyte layers with Ag being diffused by thermal annealing or ultraviolet-radiation[176,177] have been used as an electrolyte layer to aid the formation of a CF for an atomic switch or electrochemical resistive memory. Furthermore, Yang et al. recently suggested that the evolution of Ag nanoclusters in an electrolyte is determined by the local ion supply, the compensating electronic charge distribution, and the electric field such that complex and adaptive evolution of the nanocluster configurations can be expected by applying the field and ion supply from multiple terminals, as shown in **Figure 12**.[142] Such a multi-terminal system can be used for heterosynaptic plasticity to enable the important biological functions that correlate more than three neurons engaged for the purpose of, e.g., sensory perception, associative learning, and prevention of synaptic runaway dynamics as well as LTP.[142,154]

For anion-migration-based memristive materials, a hypostoichiometric switching layer has been adapted using reactive sputtering with an oxygen-deficient atmosphere, applying an oxygen-reactive electrode without or with post-annealing, and alloying with metallic inclusions. A binary oxide Al_2O_3/TiO_{2-x} bilayer has been used to demonstrate both neuromorphic networks and resistive switching memory.[137,178] Reactive sputtering at low-temperature (<300 °C),[137] atomic layer deposition (ALD) and subsequent post-deposition-annealing at 60 0°C[178] were used to reduce the TiO_2 layer and induce an oxygen vacancy profile. An oxygen vacancy-abundant Al_2O_3/TiO_{2-x} memristive switching layer could exhibit a low formation voltage and nonlinear *I–V* curves, which would allow selector-free neural networks or vertically integrated self-rectifying ReRAM.[137,178] The film deposition method and subsequent annealing can modify the film structure and oxygen vacancy concentration profile. Song et al. reported that ALD-grown TiO_2 films are weakly crystallized to the anatase phase with randomly oriented tiny crystallites, whereas reactive sputter-grown TiO_2 film has a columnar grain in a rutile phase, which is structurally similar to a CF composed of Magnéli phases.[179] Such concurrent multi-phases might lead to locally disparate memristive switching behaviors, which are related to the growth kinetics of the CF. Ta/Ta_2O_5, Hf/HfO_2, and Ti/HfO_2 stacks have been widely used as a memristive switching layer and electrode because the reactive metal creates oxygen vacancy in the interface.[22,170,180–184]

Recently, metal oxides with dispersed nanoparticles or metallic/semiconducting dopants, such as Pt-dispersed SiO_2, Mn-doped HfO_2, Na-doped WO_{3-x}, Si-doped HfO_2, and Si-doped Ta_2O_5, have been reported as resistive switching materials.[182,185–190] The switching mechanism has not been fully established, but regardless of whether purely electronic switching or partly ionic switching is involved,[185,190] such materials exhibit promising switching characteristics such as low variability and low-power operation. Choi et al. reported

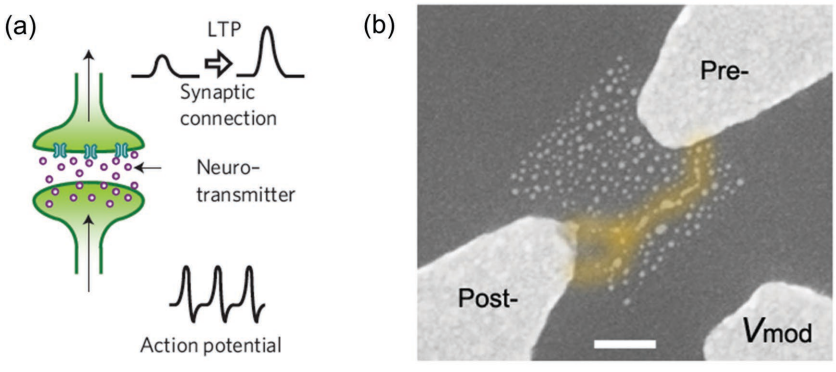


Figure 12. a) In the case of a biological synapse, the release of neurotransmitters is caused by the arrival of action potentials generated by firing, and a signal is subsequently transmitted as a synaptic potential. Reproduced with permission.[138] Copyright 2011, Nature Publishing Group. b) SEM image of the device after heterosynaptic facilitation showing a filament connecting the presynaptic and postsynaptic terminals, which operates as an LTP. Reproduced with permission.[142] Scale bar: 100 nm.

that highly uniform and rapid (≈100 ps) resistive switching characteristics were demonstrated in Pt/Pt-dispersed SiO_2/Ta material for memory applications.[185] Mandal et al. reported a novel synaptic memory device made of Mn-doped HfO_2 material at the 20 nm × 20 nm scale.[187] Compared with the in silico VLSI synapse, a nano-device with a Mn-doped HfO_2 switching layer showed a ≈10 times reduction in area and $>10^6$ times reduction in power consumption per learning cycle.

For the bipolar-type electronic switching system, the electron-trapping site plays a crucial role in controlling the stability of the resistance state. Kim et al. and Yoon et al. revealed that electron trapping sites with ≈0.8 eV of trap depth could be created in the ruptured region of a CF in the Pt/TiO_2/Pt system.[160,161] It was believed that such a moderate trap depth could offer 10-year stability. Fluent charge injection is possible via a ruptured CF region due to a lowered Schottky barrier height, a decrease in Schottky barrier thickness, or a trap-assisted conduction mechanism. However, Lim et al. reported that long-term stability of synaptic weights was not maintained without CF, but initial resistance state was recovered (i.e., short-term depression occurred) when different 'reactive' metals such as Cr, Ni, and Ti rather than inert metals such as Pt and Au were used as the top electrodes in the top metal/TiO_2/Pt system.[146]

Reports of a new type of memristor with self-rectifying or self-limiting memristive switching properties have increased. These devices commonly show low current operation with high *I*–*V* nonlinearity, which are highly desirable for the vertically integrated crossbar array architecture designed to increase the device density per area. Such behavior is also essential which provides each cell with the necessary selectivity for synaptic devices. Self-rectifying behavior has been reported from bi-layered metal oxides, e.g., Pt/TiO_2/HfO_2/TiN, Pt/TiO_2/HfO_2/Ti, Pt/Al_2O_3/NiO/W, Ti/HfO_2/TiO_x/Pt, TiN/Al_2O_3/TiO_2/TiN, Ta/TaO_x/TiO_2/Ti, etc.[162,184,191–193,195,197,198,200,201] One of these layers plays the role of the switching layer, and the other layer rectifies the current injection by forming a Schottky emission contact or tunnel barrier. In addition to bi-layered oxides, hybrid memory and selectors such as HfO_2/CuGeS, TiO_2/VO_2, and Nb_2O_5/NbO_2 have been reported.[194,196,199] CuGeS is known as a mixed ion-electron conductor (MIEC) selector, and VO_2 and NbO_2 are known as insulator-metal-transition (IMT) selector materials that deliver superior *I*–*V* nonlinearity and high current density. Due to the simple stack and absence of an individual selector device with a metal electrode, it is believed that the self-rectifying memristor is considered the only choice for vertical ReRAM. Details on the three-dimensional crossbar array of the memristor are provided in the next section.

4.2. Electrodes and Fabrication Issues for the Three-Dimensional Crossbar Array

Low energy consumption per operation (spike) is required for energy-efficient parallel computing. The lower reliability (endurance and uniformity) of memristors compared with CMOS circuit can be supplemented by the low power operation. In addition, the leakage current in the off state should be low to minimize standby power. For these reasons, low reset current and nonlinear *I*–*V* characteristics matter for effectively decreasing the power consumption in both active and standby modes. Moreover, multiple stacking of CMOS neurons and memristor synapses is of great interest in mimicking the multi-layered human brain. Therefore, implementation of a vertically integrated architecture of memristor synapses should be seriously considered.

At the moment, few studies exist on highly integrated memristor synapses. Therefore, recent attempts related to the three-dimensional stacking of memristors, interconnects, and a fabrication method for high-density nonvolatile memory are reviewed in this section. A more detailed review of the three-dimensional resistive switching crossbar array memory from integration to materials can be found in another review paper.[5,26]

It is well known that memristors can be integrated two-dimensionally into crossbar arrays using a simple metal-insulator-metal (MIM) stack. Such crossbar arrays can be further stacked in three dimensions to increase density. **Figure 13**a shows the conventional structure of horizontally stacked 3D crossbar arrays (H-CBA), which could be the simplest way to stack 2D crossbar arrays on top of each other. However, the lithographic process and its cost are expected to surge with increases in the number of stacks.[26,202]

Alternatively, the MIM junction in a crossbar array can be formed in the vertical direction, i.e., the so-called vertically stacked 3D crossbar arrays (V-CBA), which can be considered as a 90° rotation of H-CBA, as shown in Figure 13b.[26] This type of architecture can reduce the large number of lithographic processes as well as the fabrication cost. Moreover, the aforementioned self-rectifying memristors can be implemented in the V-CBA architecture, thus avoiding the metallization process in the middle for individual selector devices. Otherwise, a selective etching process for the middle electrode layer is necessary, which is generally rather challenging.

If word-lines are replaced by word-planes, as shown in Figure 13c, further reduction in the number of lithographic processes is expected and of even greater importance is that such a "word-plane-type 3D crossbar array" can simplify the interconnection between the bit-lines and word-planes connected to the periphery circuits.[5,26,203,204] Although the realization of such architecture is highly challenging, with difficulties in the design complexity of interconnections and its fabrication process, several preliminary works have been recently reported. Lee et al. reported the adoption of two graphene word-planes and Pt word-planes intervened by a SiO_2 layer to demonstrate word-plane-type V-CBA devices, as shown in **Figure 14**a and b.[204] These researchers stated that 60 stacked layers in a 3D V-CBA ReRAM were shown to be possible for a lithographic half-pitch (*F*) = 22 nm using Pt as the word-planes, whereas 200 stacked layers for V-CBA can be achieved in principle by repeating a word-plane consisting of 0.3-nm-thick graphene and a 6-nm-thick isolation SiO_2 layer. The graphene word-plane could add additional functionalities to the device operation, e.g., a built-in selector that delivers moderate nonlinearity, a thermal barrier that confines the heat generated inside, and an oxygen diffusion barrier with low activation energy for oxygen migration.[204–209] Yang et al. reported that graphene/TaO_y/Ta_2O_{5-x}/graphene exhibited nonlinear *I*–*V* curves (nonlinearity ≈ 280), and it was believed that such nonlinearity is likely to originate

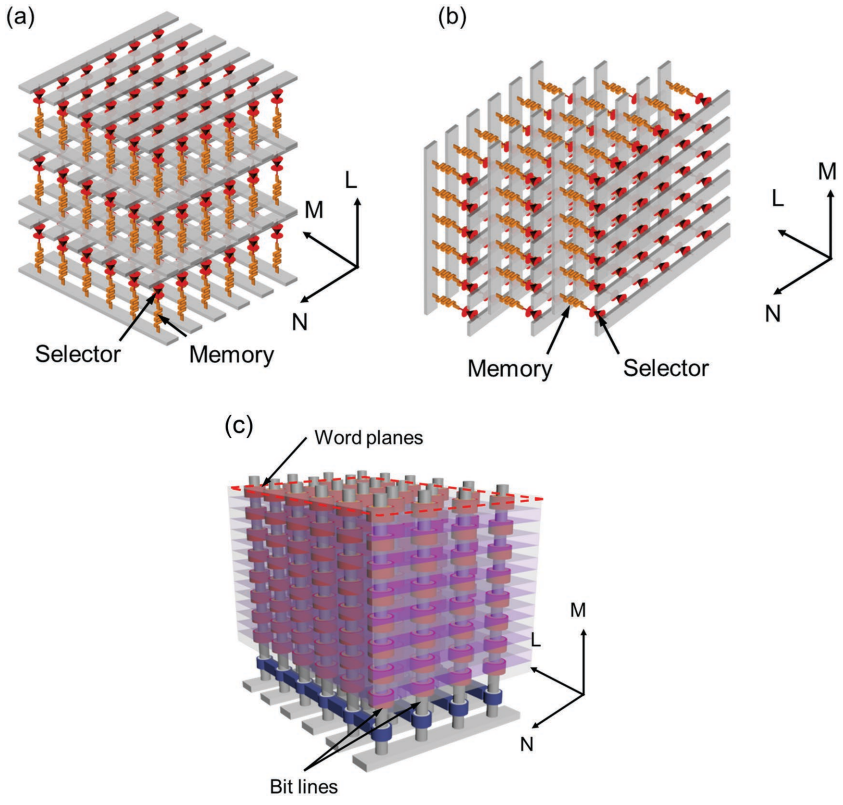


Figure 13. Standard structures of the cross-line type 3D CBA: a) vertical stacking and b) horizontal stacking of 2D CBA and c) word-plane type CBA.[5,26]

from the functionalized (oxidized) graphene at the bottom interface during the plasma-assisted deposition process for the Ta_2O_{5-x} switching layer.[205]

Because the device feature size is decreasing and a longer bit-line or word-line is needed to meet the requirements of memory density, the high resistance of the narrow metal lines could be a problem for power consumption and latency due to the high operation voltage and RC delay time, respectively.[210–213] Different voltages along a given word- or bit-line at different locations in the array structure are another critical problem. If this line resistance exceeds a certain critical value, feasible switching of the memory cell is hindered.[214] To address this issue, one possible approach is 3D integration of a memristor connected with single-walled carbon nanotubes (SWNTs), graphene/reduced graphene oxide (rGO), graphene nano-ribbons (GNRs), and topological insulators (TIs).[207,212,213,215] Low-dimensional carbon allotropes are free from the negative effects suffered by metal interconnects, e.g., electro-migration, grain boundaries, and edge scattering at highly scaled technology nodes.[213] Nevertheless, these low-dimensional materials suffer from low mobility at elevated temperatures and

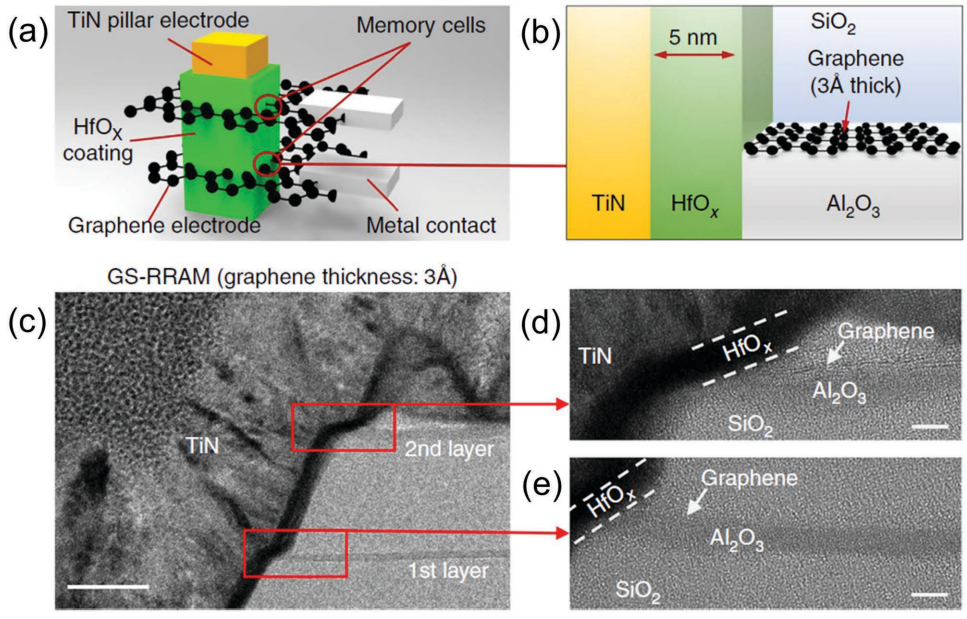


Figure 14. Structure of graphene-based and Pt-based ReRAM in a vertical 3D cross-point architecture. a) Illustration of graphene-based ReRAM in a vertical cross-point architecture. The ReRAM cells are formed at the intersections of the TiN pillar electrode and the graphene plane electrode. The resistive switching HfO_x layer surrounds the TiN pillar electrode and is also in contact with the graphene plane electrode. b) A schematic cross-section of the graphene-based ReRAM. c) High-resolution TEM image of the two-stack V-CBA structure. The ReRAM memory elements are highlighted in red. Scale bar = 40 nm. d,e) First and second layer of graphene-based ReRAM with graphene on top of the Al_2O_3 layer. Scale bars = 5 nm. Reproduced with permission.[204] Copyright 2015, Nature Publishing Group.

heat dissipation through thermal interfaces at the contacts in addition to their low material quality (defects, rough edges, etc.).[213,216] In fact, achieving a resistance (and not resistivity) identical to that of metal lines for the given line width from the two-dimensional materials is highly challenging due to their extremely thin thickness.

For fabrication of the 3D CBA architecture, numerous challenges appear in manufacturing, process integration, and characterization. It is expected that the higher demands placed on the dry etch, film deposition, and planarization processes are related to lithography for manufacturing of 3D CBA.[212] Considering the 3D V-CBA shown in Figure 13, alternating metal and insulating layers can be grown to form word-planes via conventional chemical vapor deposition (CVD) of the insulating (SiO_2) layer and sputtering of the metal layer.[26] Additionally, the subsequent etching process used to form deep holes with a high aspect ratio for introducing the bit-lines is highly challenging. Multiple etching of the insulating and metal layers should be developed with minimal damage to the etched surface because it will become the contact interface for the functional memristor and selector layers. Multiple layers of memristor materials (usually insulators) and bit-lines (metals) should be grown in the deep holes. Considering the dimensions of the deep hole (a few tens of nm in diameter and hundreds to thousands of nm in depth) and necessity of a highly uniform and conformal deposition method, only ALD can satisfy such requirements for both the metal and insulating layers. The ALD processes for high-*k* dielectric materials, the contact metal layer, and the diffusion barrier for Cu metallization have been used in the current semiconductor industry.[5]

It was envisioned that 3D CBA fabrication would be underpinned mostly by ALD, which is known as the only feasible method for meeting the various stringent requirements for 3D CBA fabrication. It was indeed the case that many reports related to the memristor, selector layer, and 3D CBA included processing with ALD. It is especially noted that self-rectifying memristors (e.g., $Pt/Ta_2O_5/HfO_{2-x}/Ti$),[198] memristive materials engineering (e.g., $TiN/Hf_{1-x}Al_xO_y/TiN$)[217] and multilayered tunnel selectors (e.g., $Pt/TaN_{1+x}/Ta_2O_5/TaN_{1+x}/Pt$)[218] could be prepared using consecutive or multi-component ALD processes. The scaling limits of the area (≈10 nm × 10 nm) and thickness (≈2 nm) of the memristive layer were also verified by the ALD-grown HfO_2 switching layer.[159,219] Park et al. worked on the electrical characterizations of electro-forming and resistive switching in 0.8-, 1.3-, 1.8-, and 2.3-nm-thick ALD-grown Ta_2O_5 switching layers formed on a 28-nm-diameter contact plug.[220,221] It was notable that a Ta_2O_5 switching layer as thin as 0.8 nm showed fluent resistance switching performance, thus demonstrating the extreme scalability of the material.

5. Concluding Remarks

The tremendous growth in information technology that humans have enjoyed over the past several decades is expected to face severe challenges within the next few decades simply because of unsustainable energy consumption, which will be over ≈100 times greater than all of the usable energy in 2040 combined with the current trend. Decreasing the energy per unit operation is certainly an option, but this approach might not be feasible due to the involvement of thermal noise at room temperature. Therefore, the most feasible approach is to decrease the number of (binary) operations itself. Although the digital computation methodology based on von Neumann architecture has been extensively developed and optimized and will still remain as the appropriate method for deterministic computational tasks, many other computationally more demanding tasks require more optimum architecture than that of the von Neumann. As the human-machine-interface becomes more intimate, such tasks that often require certain decision steps under the given circumstances will become even more important in the future. The recent successes of the "deep learning" algorithm in finding the "optimal" solution for a given information set, although it might not be the "correct" solution, is representative of such trends, i.e., mimicking of the human brain. Nevertheless, the deep learning algorithm is still a software-based solution that uses Si-based devices. Therefore, this approach might not be the correct answer to the aforementioned energy crisis in computation. A seminal match occurred in March 2016 between Google's alpha-Go machine and Sedol Lee, who has been the Go game world champion for 10 years, with victory at the hand of the machine and the score of 4:1. However, human is definitely the winner in terms of the energy consumption (≈200 KW for alpha-Go vs. ≈20 W for Sedol Lee).

In this long review, the authors have attempted to provide deeper views on newly emerging computing paradigms, including stateful logic and neuromorphic computing. Because the data volatility in current computers is a large source of energy consumption, stateful logic is an attractive option for a more energy-efficient device that still lies within the von Neumann architecture. However, "stateful" logic poses significant challenges because the output of certain logic operation (gate) is retained within the devices, and thus a separate readout step is required, which not only complicates the operation but also poses a risk of higher overall energy consumption. Therefore, the most feasible configuration of such stateful logic is a combination with current CMOS logic circuits, more preferably in three-dimensional form. This arrangement must be suitable for deterministic computation. Material implication logic realized by the memristor is a strong contender for such a device configuration.

For other computational areas, i.e., human-like tasks, a cognitive/neuromorphic computing architecture will be necessary, which is the most probable case if no well-defined "correct" answer exists under a given circumstance. An automatic car-driving algorithm could be a good example for such scenario. In this case, the hardware must contain neurons and synapses rather than CPUs and memories. The in silico approach was a first step, as demonstrated by IBM's True North chip, but eventually, a genuine hardware approach must be pursued, i.e., materials and devices that mimic biological neurons and synapses. To this end, the memristor appears to offer great possibility albeit there are still significant concerns about its reliability. At the same time, a defect-tolerant architecture similar to the brain must be developed. Great improvements have been reported in this direction during past decade in terms of material processing, understanding of properties, and device fabrication. A subset of the most significant improvements is included in this review but other points could be missing.

Impressive and significant improvements have occurred in the understanding of the human brain during the past decades, which means that designers, architects, and process engineers working on computers and semiconductor chips have dared to mimic the brain with various hardware and software approaches. Memristors are expected to play a critical role in this challenging area, both with conventional and new computing architectures. This trend will be strengthened when the roles of the chaotic processes in the brain are better understood and the extreme dynamics of memristors near the chaotic edge can be correlated with such neuronal behaviors. Nevertheless, the fundamental question of the ultimate performance of a computing machine that mimics the brain functionality will remain unanswered for a long time because the question of the origin of "intelligence" and "consciousness" is not yet clearly answered; is it a consequence of types of "physical" operations that occur in so many computing elements (neurons and synapses) and thus is a computable entity, or a consequence of something unknown that is not yet captured? With respect to this question, the 480-page monograph written by R. Penrose, titled "Shadows of the Mind", offers the community insightful viewpoints.[222]

Acknowledgements

D.S.J. and K.M.K. contributed equally to this work. CSH acknowledges support from the Global Research Laboratory Program (No. NRF-2012K1A1A2040157) of the Ministry of Science, ICT, and Future Planning, and the National Research Foundation of Korea (NRF) grant (No. NRF-2014R1A2A1A10052979) from the Republic of Korea.

Received: March 1, 2016
Revised: July 18, 2016
Published online:

[1] M. Hilbert, P. López, *Science* **2011**, *332*, 60.
[2] J. Gantz, D. Reinsel, IDC iView: IDC Analyze the future 2012, **2007**, 1.
[3] R. Landauer, *IBM J. Res. Dev.* **1961**, *5*, 183.
[4] C. Sun, M. T. Wade, Y. Lee, J. S. Orcutt, L. Alloatti, M. S. Georgas, A. S. Waterman, J. M. Shainline, R. R. Avizienis, S. Lin, *Nature* **2015**, *528*, 534.
[5] C. S. Hwang, *Adv. Electron. Mater.* **2015**, *1*, 6.
[6] J. Borghetti, G. S. Snider, P. J. Kuekes, J. J. Yang, D. R. Stewart, R. S. Williams, *Nature* **2010**, *464*, 873.
[7] a) https://www.research.ibm.com/software/IBMResearch/multimedia/Computing_Cognition_WhitePaper.pdf (accessed: August, 2016); b) H. Markram, *Sci. Am.* **2012**, *306*, 50.
[8] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[9] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, D. S. Modha, *Science* **2014**, *345*, 668.
[10] https://www.qualcomm.com/news/onq/2013/1010/introducing-qualcomm-zeroth-processors-brain-inspired-computing (accessed: August, 2016).
[11] G. Roth, U. Dicke, *Trends Cognit. Sci.* **2005**, *9*, 250.
[12] http://cpudb.stanford.edu/ (accessed: November, 2015).
[13] B. Jacob, S. W. Ng, D. T. Wang, *Memory Systems: Cache, DRAM, Disk*, Elsevier, San Francisco, **2007**.
[14] http://arstechnica.com/gadgets/2013/08/samsungs-3d-vertical-nand-crams-a-terabit-on-a-single-chip/ (accessed: August, 2016).
[15] L. O. Chua, *IEEE Trans. Circuit Theory* **1971**, *18*, 507.
[16] D. B. Strukov, G. S. Snider, D. R. Stewart, R. S. Williams, *Nature* **2008**, *453*, 80.
[17] A. Beck, J. Bednorz, C. Gerber, C. Rossel, D. Widmer, *Appl. Phys. Lett.* **2000**, *77*, 139.
[18] T. Hickmott, *J. Vac. Sci. Technol* **1969**, *6*, 828.
[19] T. W. Hickmott, *J. Appl. Phys.* **1962**, *33*, 2669.
[20] T. W. Hickmott, *J. Appl. Phys.* **1964**, *35*, 2679.
[21] B. Cho, S. Song, Y. Ji, T. W. Kim, T. Lee, *Adv. Funct. Mater.* **2011**, *21*, 2806.
[22] D. S. Jeong, R. Thomas, R. Katiyar, J. Scott, H. Kohlstedt, A. Petraru, C. S. Hwang, *Rep. Prog. Phys.* **2012**, *75*, 076502.
[23] K. M. Kim, D. S. Jeong, C. S. Hwang, *Nanotechnol.* **2011**, *22*, 254002.
[24] W. P. Lin, S. J. Liu, T. Gong, Q. Zhao, W. Huang, *Adv. Mater.* **2014**, *26*, 570.
[25] A. Sawa, *Mater. Today* **2008**, *11*, 28.
[26] J. Y. Seok, S. J. Song, J. H. Yoon, K. J. Yoon, T. H. Park, D. E. Kwon, H. Lim, G. H. Kim, D. S. Jeong, C. S. Hwang, *Adv. Funct. Mater.* **2014**, *24*, 5316.
[27] J. J. Yang, D. B. Strukov, D. R. Stewart, *Nat. Nanotechnol.* **2013**, *8*, 13.
[28] Y. Yang, W. Lu, *Nanoscale* **2013**, *5*, 10076.
[29] K. Mainzer, L. O. Chua, *Local Activity Principle: The Cause of Complexity and Symmetry Breaking*, Imperial College Press, London **2013**.
[30] M. D. Pickett, G. Medeiros-Ribeiro, R. S. Williams, *Nat. Mater.* **2013**, *12*, 114.
[31] J. King, S. Yarkoni, M. M. Nevisi, J. P. Hilton, C. C. McGeoch, arXiv:1508.05087 **2015**.
[32] K. Chen, F. Lombardi, J. Han, *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition* **2015**, *293*.
[33] A. N. Whitehead, B. Russell, *Principia mathematica*, Vol. *2*, Cambridge University Press, Cambridge **1912**.
[34] C. E. Shannon, *Trans. Am. Inst. Electr. Eng.* **1938**, *57*, 713.
[35] L. O. Chua, S. M. Kang, *Proc. IEEE* **1976**, *64*, 209.
[36] L. O. Chua, *Appl. Phys. A Mater. Sci. Process.* **2011**, *102*, 765.
[37] L. O. Chua, *Semicond. Sci. Technol.* **2014**, *29*, 104001.
[38] Data collected from ISI Web of Knowledge (as of June **2016**)
[39] J. R. Heath, P. J. Kuekes, G. S. Snider, R. S. Williams, *Science* **1998**, *280*, 1716.
[40] R. F. Service, *Science* **2001**, *294*, 2442.
[41] M. M. Ziegler, M. R. Stan, *IEEE Trans. Nanotechnol.* **2003**, *2*, 217.
[42] D. B. Strukov, K. K. Likharev, *Nanotechnol.* **2005**, *16*, 888.
[43] Q. Xia, W. Robinett, M. Cumbie, N. Banerjee, T. J. Cardinali, J. J. Yang, W. Wu, X. Li, W. M. Tong, D. B. Strukov, G. S. Snider, G. Medeiros-Ribeiro, R. S. Williams, *Nano Lett.* **2009**, *9*, 3640.
[44] G. Boole, *An Investigation of the Laws of Thought, On Which Are Founded the Mathematical Theories of Logic and Probabilities*, Walton and Maberly, London **1854**.
[45] I. Vourkas, G. C. Sirakoulis, *Memristor-Based Nanoelectronic Computing Circuits and Architectures*, Springer, New York, **2016**.
[46] E. Linn, R. Rosezin, S. Tappertzhofen, U. Böttger, R. Waser, *Nanotechnol.* **2012**, *23*, 305205.
[47] E. Linn, R. Rosezin, C. Kügeler, R. Waser, *Nat. Mater.* **2010**, *9*, 403.
[48] K. J. Yoon, S. J. Song, J. Y. Seok, J. H. Yoon, T. H. Park, D. E. Kwon, C. S. Hwang, *Nanoscale* **2014**, *6*, 2161.
[49] F. Zhou, L. Guckert, Y. F. Chang, E. E. Swartzlander, J. Lee, *Appl. Phys. Lett.* **2015**, *107*, 183501.
[50] C. Mead, *Analog VLSI and Neural Systems*, Adison-Wesley, Reading, MA **1989**.
[51] C. Mead, *Proc. IEEE* **1990**, *78*, 1629.
[52] A. Cassidy, S. Denham, P. Kanold, A. Andreou, *IEEE Biomedical Circuits and Systems Conference (BIOCAS)* **2007**, 27.

[53] A. Cassidy, A. G. Andreou, J. Georgiou, *Annual Conference on Information Sciences and Systems (CISS)* **2011**, 23.
[54] J. Li, Y. Katori, T. Kohno, *Front. Neurosci.* **2012**, *6*, 183.
[55] T. Delbruck, *IEEE Trans. Neural Netw.* **1993**, *4*, 529.
[56] T. Delbrück, S.-C. Liu, *Vision Res.* **2004**, *44*, 2083.
[57] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[58] G. E. Hinton, *Trends Cognit. Sci.* **2007**, *11*, 428.
[59] E. B. Baum, *J. Complexity* **1988**, *4*, 193.
[60] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, *Proc. IEEE* **1998**, *86*, 2278.
[61] Y. LeCun, K. Kavukcuoglu, C. Farabet, *IEEE International Symposium on Circuits and Systems (ISCAS)* **2010**, 253.
[62] G. E. Hinton, S. Osindero, Y.-W. Teh, *Neural Comput.* **2006**, *18*, 1527.
[63] D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Nature* **1986**, *323*, 533.
[64] Y. LeCun, S. Schopra, R. Radsell, M. A. Ranzato, F. J. Huang, *Predicting Structured Data*, MIT Press, Cambridge, **2007**.
[65] http://neuralnetworksanddeeplearning.com (accessed: August, 2016).
[66] A. Krizhevsky, I. Sutskever, G. E. Hinton, *Conference on Neural Information Processing Systems (NIPS)* **2012**, 25.
[67] D. S. Jeong, I. Kim, M. Ziegler, H. Kohlstedt, *RSC Adv.* **2013**, *3*, 3169.
[68] P. Dayan, L. F. Abbott, *Theoretical Neuroscience*, MIT Press, London **2001**.
[69] C. Eliasmith, C. H. Anderson, *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*, MIT Press, Cambridge **2003**.
[70] W. Gerstner, W. M. Kistler, *Spiking Neuron Models: Single Neurons, Populations, Plasticity*, Cambridge University Press, New York **2002**.
[71] T. J. Hamilton, S. Afshar, A. van Schaik, J. Tapson, *Proc. IEEE* **2014**, *102*, 843.
[72] V. Kornijcuk, H. Lim, J. Y. Seok, G. Kim, S. K. Kim, I. Kim, B. J. Choi, D. S. Jeong, *Front. Neurosci.* **2016**, *10*, 212.
[73] M. R. Azghadi, N. Iannella, S. F. Al-Sarawi, G. Indiveri, D. Abbott, *Proc. IEEE* **2014**, *102*, 717.
[74] H. Lim, H.-W. Ahn, V. Kornijcuk, G. Kim, J. Y. Seok, I. Kim, C. S. Hwang, D. S. Jeong, *Nanoscale* **2016**, *8*, 9629.
[75] S. A. Bamford, A. F. Murray, D. J. Willshaw, *IEEE Trans. Bio-Med. Circuits Syst.* **2012**, *6*, 385.
[76] D. C. Gadsby, *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 344.
[77] S. O. Pearson, H. S. G. Anson, *Proc. Phys. Soc.*, London **1921**, *34*, 204.
[78] H. Lim, V. Kornijcuk, J. Y. Seok, S. K. Kim, I. Kim, C. S. Hwang, D. S. Jeong, *Sci. Rep.* **2015**, *5*, 9776.
[79] M. Ignatov, M. Ziegler, M. Hansen, A. Petraru, H. Kohlstedt, *Front. Neurosci.* **2015**, *9*, 376.
[80] D. S. Jeong, H. Lim, G.-H. Park, C. S. Hwang, S. Lee, B.-k. Cheong, *J. Appl. Phys.* **2012**, *111*, 102807.
[81] H.-W. Ahn, D. S. Jeong, B.-k. Cheong, S.-d. Kim, S.-Y. Shin, H. Lim, D. Kim, S. Lee, *ECS Solid State Lett.* **2013**, *2*, N31.
[82] M.-J. Lee, D. Lee, S.-H. Cho, J.-H. Hur, S.-M. Lee, D. H. Seo, D.-S. Kim, M.-S. Yang, S. Lee, E. Hwang, M. R. Uddin, H. Kim, U. I. Chung, Y. Park, I.-K. Yoo, *Nat. Commun.* **2013**, *4*, 2629.
[83] X. Tong, H. Wu, L. Zhao, H. Zhong, *ECS Trans.* **2013**, *52*, 105.
[84] P. Krzysteczko, J. Münchenberger, M. Schäfers, G. Reiss, A. Thomas, *Adv. Mater.* **2012**, *24*, 762.
[85] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, W. Lu, *Nano Lett.* **2010**, *10*, 1297.
[86] G. S. Snider, *Nanotechnol.* **2007**, *18*, 365202.
[87] J. Hasler, H. B. Marr, *Front. Neurosci.* **2013**, *7*, 118.
[88] G. W. Burr, R. M. Shelby, C. di Nolfo, J. W. Jang, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. Kurdi, H. Hwang, *IEEE International Electron Devices Meeting (IEDM)* **2014**, 29.5.1.
[89] G. W. Burr, P. Narayanan, R. M. Shelby, S. Sidler, I. Boybat, C. di Nolfo, Y. Leblebici, *IEEE International Electron Devices Meeting (IEDM)* **2015**, 4.4.1.
[90] P. Merolla, J. Arthur, F. Akopyan, N. Imam, R. Manohar, D. S. Modha, *IEEE Custom Integrated Circuits Conference (CICC)* **2011**, 1.
[91] D. O. Hebb, *The organization of behavior*, Wiley & Sons, New York **1949**.
[92] E. Bienenstock, L. Cooper, P. Munro, *J. Neurosci.* **1982**, *2*, 32.
[93] T. V. P. Bliss, G. L. Collingridge, *Nature* **1993**, *361*, 31.
[94] S. A. Siegelbaum, E. R. Kandel, *Curr. Opin. Neurobiol.* **1991**, *1*, 113.
[95] T. Toyoizumi, J.-P. Pfister, K. Aihara, W. Gerstner, *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 5239.
[96] J. Gjorgjieva, C. Clopath, J. Audet, J.-P. Pfister, *Proc. Natl. Acad. Sci.* **2011**, *108*, 19383.
[97] S. Song, K. D. Miller, L. F. Abbott, *Nat. Neurosci.* **2000**, *3*, 919.
[98] N. Caporale, Y. Dan, *Annu. Rev. Neurosci.* **2008**, *31*, 25.
[99] G.-Q. Bi, M.-M. Poo, *J. Neurosci.* **1998**, *18*, 10464.
[100] M. C. W. van Rossum, G. Q. Bi, G. G. Turrigiano, *J. Neurosci.* **2000**, *20*, 8812.
[101] Y. Dan, M.-M. Poo, *Physiol. Rev.* **2006**, *86*, 1033.
[102] I. R. Fiete, H. S. Seung, *The New Encyclopedia of Neuroscience*, Elsevier, New York **2008**.
[103] I. R. Fiete, W. Senn, C. Z. H. Wang, R. H. R. Hahnloser, *Neuron* **2010**, *65*, 563.
[104] J.-P. Pfister, W. Gerstner, *J. Neurosci.* **2006**, *26*, 9673.
[105] R. J. Vogelstein, F. Tenore, R. Philipp, M. S. Adlerstein, D. H. Goldberg, G. Cauwenberghs, *Advances in Neural Information Processing Systems*, MIT Press, Cambridge **2002**, *15*, 1147.
[106] A. van Schaik, *Neural Netw.* **2001**, *14*, 617.
[107] G. Indiveri, E. Chicca, R. Douglas, *IEEE Trans. Neural Netw.* **2006**, *17*, 211.
[108] P. Häfliger, M. Mahowald, L. Watts, *Advances in Neural Information Processing Systems* **1996**, *9*, 692.
[109] V. Kornijcuk, O. Kavehei, H. Lim, J. Y. Seok, S. K. Kim, I. Kim, W.-S. Lee, B. J. Choi, D. S. Jeong, *Nanoscale* **2014**, *6*, 15151.
[110] S. Kim, M. Ishii, S. Lewis, T. Perri, M. BrightSky, W. Kim, R. Jordan, G. Burr, N. Sosa, A. Ray, *IEEE International Electron Devices Meeting (IEDM)* **2015**, 17.1.1.
[111] Y. Li, Y. Zhong, J. Zhang, L. Xu, Q. Wang, H. Sun, H. Tong, X. Cheng, X. Miao, *Sci. Rep.* **2014**, *4*, 4906.
[112] W. He, K. Huang, N. Ning, K. Ramanathan, G. Li, Y. Jiang, J. Sze, L. Shi, R. Zhao, J. Pei, *Sci. Rep.* **2014**, *4*, 4755.
[113] M. Suri, D. Querlioz, O. Bichler, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, B. DeSalvo, *IEEE Trans. Elec. Dev.* **2013**, *60*, 2402.
[114] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, H. P. Wong, *IEEE Trans. Elec. Dev.* **2011**, *58*, 2729.
[115] S. Kim, C. Du, P. Sheridan, W. Ma, S. Choi, W. D. Lu, *Nano Lett.* **2015**, *15*, 2203.
[116] C. Du, W. Ma, T. Chang, P. Sheridan, W. D. Lu, *Adv. Funct. Mater.* **2015**, *25*, 4290.
[117] Z. Wang, S. Ambrogio, S. Balatti, D. Ielmini, *Front. Neurosci.* **2015**, *8*, 438.
[118] D. Kuzum, R. G. D. Jeyasingh, B. Lee, H. S. P. Wong, *Nano Lett.* **2011**, *12*, 2179.
[119] M. Suri, O. Bichler, D. Querlioz, B. Traore, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, B. DeSalvo, *J. Appl. Phys.* **2012**, *112*, 054904.
[120] A. F. Vincent, J. Larroque, W. S. Zhao, N. Ben Romdhane, O. Bichler, C. Gamrat, J. O. Klein, S. Galdin-Retailleau, D. Querlioz, *IEEE International Symposium on Circuits and Systems (ISCAS)* **2014**, 1074.
[121] Y. Nishitani, Y. Kaneko, M. Ueda, T. Morie, E. Fujii, *J. Appl. Phys.* **2012**, *111*, 124108.
[122] D. Kuzum, Y. Shimeng, H. S. P. Wong, *Nanotechnol.* **2013**, *24*, 382001.
[123] S. Saïghi, C. G. Mayr, T. Serrano-Gotarredona, H. Schmidt, G. Lecerf, J. Tomas, J. Grollier, S. Boyn, A. Vincent, D. Querlioz,

S. La Barbera, F. Alibart, D. Vuillaume, O. Bichler, C. Gamrat, B. Linares-Barranco, *Front. Neurosci.* **2015**, *9*, 51.
[124] K. Likharev, A. Mayr, I. Muckra, Ö. TÜRel, *Ann. N.Y. Acad. Sci.* **2003**, *1006*, 146.
[125] T. S. Lande, *Neuromorphic Systems Engineering: Neural Networks in Silicon*, Kluwer Academy Publishers, Boston **1998**.
[126] K. A. Boahen, *IEEE Trans. Circuits Syst. II: Analog and Digital Signal Processing* **2000**, *47*, 416.
[127] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. van Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, K. Boahen, *Front. Neurosci.* **2011**, *5*, 73.
[128] A. N. Burkitt, G. M. Clark, *Neural Comput.* **2000**, *12*, 1789.
[129] C. C. H. Petersen, R. C. Malenka, R. A. Nicoll, J. J. Hopfield, *Proc. Natl. Acad. Sci.* **1998**, *95*, 4732.
[130] M. Graupner, N. Brunel, *Proc. Natl. Acad. Sci.* **2012**, *109*, 3991.
[131] P. J. Sjöström, G. G. Turrigiano, S. B. Nelson, *Neuron* **2001**, *32*, 1149.
[132] C. Baldassi, A. Braunstein, N. Brunel, R. Zecchina, *Proc. Natl. Acad. Sci.* **2007**, *104*, 11079.
[133] A. Braunstein, R. Zecchina, *Phys. Rev. Lett.* **2006**, *96*, 030201.
[134] P. Lennie, *Curr. Biol.* **2003**, *13*, 493.
[135] G. H. Kim, K. M. Kim, J. Y. Seok, H. J. Lee, D.-Y. Cho, J. H. Han, C. S. Hwang, *Nanotechnol.* **2010**, *21*, 385202.
[136] W. Y. Park, G. H. Kim, J. Y. Seok, K. M. Kim, S. J. Song, M. H. Lee, C. S. Hwang, *Nanotechnol.* **2010**, *21*, 195201.
[137] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, D. B. Strukov, *Nature* **2015**, *521*, 61.
[138] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. K. Gimzewski, M. Aono, *Nat. Mater.* **2011**, *10*, 591.
[139] Q. Lai, L. Zhang, Z. Li, W. F. Stickle, R. S. Williams, Y. Chen, *Adv. Mater.* **2010**, *22*, 2448.
[140] M.-J. Lee, C. B. Lee, D. Lee, S. R. Lee, M. Chang, J. H. Hur, Y.-B. Kim, C.-J. Kim, D. H. Seo, S. Seo, *Nat. Mater.* **2011**, *10*, 625.
[141] F. Miao, J. P. Strachan, J. J. Yang, M. X. Zhang, I. Goldfarb, A. C. Torrezan, P. Eschbach, R. D. Kelley, G. Medeiros-Ribeiro, R. S. Williams, *Adv. Mater.* **2011**, *23*, 5633.
[142] Y. Yang, B. Chen, W. D. Lu, *Adv. Mater.* **2015**, *16*, 7720.
[143] B. Gao, Y. Bi, H. Y. Chen, R. Liu, P. Huang, B. Chen, L. Liu, X. Liu, S. Yu, H. S. P. Wong, J. Kang, *ACS Nano* **2014**, *8*, 6998.
[144] D. Garbin, O. Bichler, E. Vianello, Q. Rafhay, C. Gamrat, L. Perniola, G. Ghibaudo, B. Desalvo, *IEEE International Electron Devices Meeting (IEDM)* **2014**, 28.4.1.
[145] B. DeSalvo, E. Vianello, O. Thomas, F. Clermidy, O. Bichler, C. Gamrat, L. Perniola, *IEEE International Symposium on Circuits and Systems (ISCAS)* **2015**, 3088.
[146] H. Lim, I. Kim, J.-S. Kim, C. Seong Hwang, D. S. Jeong, *Nanotechnol.* **2013**, *24*, 4005.
[147] H. Lim, H. W. Jang, D.-K. Lee, I. Kim, C. S. Hwang, D. S. Jeong, *Nanoscale* **2013**, *5*, 6363.
[148] F. Alibart, E. Zamanidoost, D. B. Strukov, *Nat. Commun.* **2013**, *4*, 2072.
[149] A. Chanthbouala, V. Garcia, R. O. Cherifi, K. Bouzehouane, S. Fusil, X. Moya, S. Xavier, H. Yamada, C. Deranlot, N. D. Mathur, M. Bibes, A. Barthélémy, J. Grollier, *Nat. Mater.* **2012**, *11*, 860.
[150] Z. Q. Wang, H. Y. Xu, X. H. Li, H. Yu, Y. C. Liu, X. J. Zhu, *Adv. Funct. Mater.* **2012**, *22*, 2759.
[151] J. Jang, S. Park, G. W. Burr, H. Hwang, Y. Jeong, *IEEE Elec. Dev. Lett.* **2015**, *36*, 457.
[152] S. Park, a. Sheri, J. Kim, J. Noh, J. Jang, M. Jeon, B. Lee, B. R. Lee, B. H. Lee, H. Hwang, *IEEE International Electron Devices Meeting (IEDM)* **2013**, 625.
[153] M. Suri, O. Bichler, D. Querlioz, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, B. DeSalvo, *IEEE International Electron Devices Meeting (IEDM)* **2012**, 10.3.1.
[154] M. Ziegler, R. Soni, T. Patelczyk, M. Ignatov, T. Bartsch, P. Meuffels, H. Kohlstedt, *Adv. Funct. Mater.* **2012**, *22*, 2744.
[155] T. Hasegawa, K. Terabe, T. Tsuruoka, M. Aono, *Adv. Mater.* **2012**, *24*, 252.
[156] W. Zhang, Y. Hu, T.-C. Chang, K.-C. Chang, T.-M. Tsai, H.-L. Chen, Y.-T. Su, T.-J. Chu, M.-C. Chen, H.-C. Huang, W.-C. Su, J.-C. Zheng, Y.-C. Hung, S. Sze, *IEEE Elec. Dev. Lett.* **2015**, *36*, 772.
[157] F. Alibart, S. Pieutin, D. Guérin, C. Novembre, S. Lenfant, K. Lmimouni, C. Gamrat, D. Vuillaume, *Adv. Funct. Mater.* **2010**, *20*, 330.
[158] F. Alibart, S. Pleutin, O. Bichler, C. Gamrat, T. Serrano-Gotarredona, B. Linares-Barranco, D. Vuillaume, *Adv. Funct. Mater.* **2012**, *22*, 609.
[159] S. Gaba, F. Cai, J. Zhou, W. D. Lu, *IEEE Elec. Dev. Lett.* **2014**, *35*, 1239.
[160] K. M. Kim, B. J. Choi, M. H. Lee, G. H. Kim, S. J. Song, J. Y. Seok, J. H. Yoon, S. Han, C. S. Hwang, *Nanotechnol.* **2011**, *22*, 254010.
[161] K. J. Yoon, M. H. Lee, G. H. Kim, S. J. Song, J. Y. Seok, S. Han, J. H. Yoon, K. M. Kim, C. S. Hwang, *Nanotechnol.* **2012**, *23*, 185202.
[162] J. H. Yoon, S. J. Song, I.-H. Yoo, J. Y. Seok, K. J. Yoon, D. E. Kwon, T. H. Park, C. S. Hwang, *Adv. Funct. Mater.* **2014**, *24*, 5086.
[163] J. P. Strachan, M. D. Pickett, J. J. Yang, S. Aloni, D. A. L. Kilcoyne, G. Medeiros-Ribeiro, R. S. Williams, *Adv. Mater.* **2010**, *22*, 3573.
[164] D.-H. Kwon, K. M. Kim, J. H. Jang, J. M. Jeon, M. H. Lee, G. H. Kim, X.-S. Li, G.-S. Park, B. Lee, S. Han, M. Kim, C. S. Hwang, *Nat. Nanotechnol.* **2010**, *5*, 148.
[165] F. Miao, J. P. Strachan, J. J. Yang, M.-X. Zhang, I. Goldfarb, A. C. Torrezan, P. Eschbach, R. D. Kelley, G. Medeiros-Ribeiro, R. S. Williams, *Adv. Mater.* **2011**, *23*, 5633.
[166] S. Kumar, M. D. Pickett, J. P. Strachan, G. Gibson, Y. Nishi, R. S. Williams, *Adv. Mater.* **2013**, *25*, 6128.
[167] T. Fujii, M. Arita, K. Hamada, Y. Takahashi, N. Sakaguchi, *J. Appl. Phys.* **2013**, *113*, 083701.
[168] G.-S. Park, Y.-B. Kim, S. Y. Park, X. S. Li, S. Heo, M.-J. Lee, M. Chang, J. H. Kwon, M. Kim, U.-I. Chung, R. Dittmann, R. Waser, K. Kim, *Nat. Commun.* **2013**, *4*, 2382.
[169] D. S. Hong, Y. S. Chen, Y. Li, H. W. Yang, L. L. Wei, B. G. Shen, J. R. Sun, *Sci. Rep.* **2014**, *4*, 4058.
[170] U. Celano, L. Goux, R. Degraeve, A. Fantini, O. Richard, H. Bender, M. Jurczak, W. Vandervorst, *Nano Lett.* **2015**, *15*, 7970.
[171] J. Y. Chen, C. L. Hsin, C. W. Huang, C. H. Chiu, Y. T. Huang, S. J. Lin, W. W. Wu, L. J. Chen, *Nano Lett.* **2013**, *13*, 3671.
[172] Y. Yang, P. Gao, S. Gaba, T. Chang, X. Pan, W. Lu, *Nat. Commun.* **2012**, *3*, 732.
[173] Q. Liu, J. Sun, H. Lv, S. Long, K. Yin, N. Wan, Y. Li, L. Sun, M. Liu, *Adv. Mater.* **2012**, *24*, 1844.
[174] X. Tian, L. Wang, J. Wei, S. Yang, W. Wang, Z. Xu, X. Bai, *Nano Res.* **2014**, *7*, 1065.
[175] G. Molas, E. Vianello, F. Dahmani, M. Barci, P. Blaise, J. Guy, A. Toffoli, M. Bernard, A. Roule, F. Pierre, C. Licitra, B. De Salvo, L. Perniola, *IEEE International Electron Devices Meeting (IEDM)* **2014**, 6.1.1.
[176] M. N. Kozicki, M. Mitkova, M. Park, M. Balakrishnan, C. Gopalan, *Superlattices Microstruct.* **2003**, *34*, 459.
[177] S. Tappertzhofen, R. Waser, I. Valov, *ChemElectroChem* **2014**, *1*, 1287.
[178] B. Govoreanu, a. Redolfi, L. Zhang, C. Adelmann, M. Popovici, S. Clima, H. Hody, V. Paraschiv, I. P. Radu, a. Franquet, J. C. Liu, J. Swerts, O. Richard, H. Bender, L. Altimime, M. Jurczak, *IEEE International Electron Devices Meeting (IEDM)* **2013**, 10.2.1.
[179] S. J. Song, J. Y. Seok, J. H. Yoon, K. M. Kim, G. H. Kim, M. H. Lee, C. S. Hwang, *Sci. Rep.* **2013**, *3*, 3443.
[180] Y. Y. Chen, M. Komura, R. Degraeve, B. Govoreanu, L. Goux, A. Fantini, N. Raghavan, S. Clima, L. Zhang, A. Belmonte,

A. Redolfi, G. S. Kar, G. Groeseneken, D. J. Wouters, M. Jurczak, *IEEE International Electron Devices Meeting (IEDM)* **2013**, 10.1.1.
[181] N. Ge, M.-X. Zhang, L. Zhang, J. J. Yang, Z. Li, R. S. Williams, *Semicond. Sci. Technol.* **2014**, *29*, 104003.
[182] Z. Wang, S. Ambrogio, S. Balatti, S. Sills, A. Calderoni, N. Ramaswamy, D. Ielmini, *IEEE International Electron Devices Meeting (IEDM)* **2015**, 7.6.1.
[183] C. Y. Chen, A. Fantini, L. Goux, R. Degraeve, S. Clima, A. Redolfi, G. Groeseneken, M. Jurczak, *IEEE International Electron Devices Meeting (IEDM)* **2015**, 10.6.1.
[184] Y.-F. Wang, Y.-C. Lin, I.-T. Wang, T.-P. Lin, T.-H. Hou, *Sci. Rep.* **2015**, *5*, 10150.
[185] B. J. Choi, A. C. Torrezan, K. J. Norris, F. Miao, J. P. Strachan, M.-X. Zhang, D. A. A. Ohlberg, N. P. Kobayashi, J. J. Yang, R. S. Williams, *Nano Lett.* **2013**, *13*, 3213.
[186] S. Kim, S. Choi, J. Lee, W. D. Lu, *ACS Nano* **2014**, *8*, 10262.
[187] S. Mandal, A. El-Amin, K. Alexander, B. Rajendran, R. Jha, *Sci. Rep.* **2014**, *4*, 5333.
[188] D. Shang, P. Li, T. Wang, E. Carria, J. Sun, B. Shen, T. Taubner, I. Valov, R. Waser, M. Wuttig, *Nanoscale* **2015**, *7*, 6023.
[189] S. Choi, P. Sheridan, W. D. Lu, *Sci. Rep.* **2015**, *5*, 10492.
[190] B. J. Choi, A. B. K. Chen, X. Yang, I.-W. Chen, *Adv. Mater.* **2011**, *23*, 3847.
[191] H. D. Lee, S. G. Kim, K. Cho, H. Hwang, H. Choi, J. Lee, S. H. Lee, H. J. Lee, J. Suh, S. Chung, Y. S. Kim, K. S. Kim, W. S. Nam, J. T. Cheong, J. T. Kim, S. Chae, E. Hwang, S. N. Park, Y. S. Sohn, C. G. Lee, H. S. Shin, K. J. Lee, K. Hong, H. G. Jeong, K. M. Rho, Y. K. Kim, J. Nickel, J. J. Yang, H. S. Cho, F. Perner, R. S. Williams, J. H. Lee, S. K. Park, *Symp. Very-Large-Scale Integration (VLSI) Technol.* **2012**, 151.
[192] C. Hsu, C. Wan, I. Wang, M. Chen, C. Lo, Y. Lee, W. Jang, C. Lin, T. Hou, *IEEE International Electron Devices Meeting (IEDM)* **2013**, 10.4.1.
[193] C. Hsu, I. Wang, C. Lo, M. Chiang, W. Jang, *Symp. Very-Large-Scale Integration (VLSI) Technol.* **2013**, T166.
[194] E. Cha, J. Woo, D. Lee, S. Lee, J. Song, Y. Koo, J. Lee, C. G. Park, M. Y. Yang, K. Kamiya, K. Shiraishi, B. Magyari-Köpe, Y. Nishi, H. Hwang, *IEEE International Electron Devices Meeting (IEDM)* **2013**, 10.5.1.
[195] S. Lee, D. Lee, J. Woo, E. Cha, J. Song, J. Park, H. Hwang, *IEEE International Electron Devices Meeting (IEDM)* **2013**, 10.6.1.
[196] Q. Luo, X. Xu, H. Liu, H. Lv, T. Gong, S. Long, Q. Liu, W. Banerjee, L. Li, J. Gao, N. Lu, S. S. Chung, *IEEE International Electron Devices Meeting (IEDM)* **2015**, 10.2.1.
[197] K. M. Kim, S. R. Lee, S. Kim, M. Chang, C. S. Hwang, *Adv. Funct. Mater.* **2015**, *25*, 1527.
[198] J. H. Yoon, K. M. Kim, S. J. Song, J. Y. Seok, K. J. Yoon, D. E. Kwon, T. H. Park, Y. J. Kwon, X. Shao, C. S. Hwang, *Adv. Mater.* **2015**, *27*, 3811.
[199] S. H. Chang, S. B. Lee, D. Y. Jeon, S. J. Park, G. T. Kim, S. M. Yang, S. C. Chae, H. K. Yoo, B. S. Kang, M.-J. Lee, T. W. Noh, *Adv. Mater.* **2011**, *23*, 4063.
[200] J. J. Yang, M.-X. Zhang, M. D. Pickett, F. Miao, J. P. Strachan, W.-D. Li, W. Yi, D. A. A. Ohlberg, B. J. Choi, W. Wu, J. H. Nickel, G. Medeiros-Ribeiro, R. S. Williams, *Appl. Phys. Lett.* **2012**, *100*, 113501.
[201] I. Wang, Y. Lin, Y. Wang, C. Hsu, T. Hou, I. Wang, C. Lo, M. Chiang, W. Jang, *IEEE International Electron Devices Meeting (IEDM)* **2014**, 28.5.1.
[202] I. G. Baek, C. J. Park, H. Ju, D. J. Seong, H. S. Ahn, J. H. Kim, M. K. Yang, S. H. Song, E. M. Kim, S. O. Park, C. H. Park, C. W. Song, G. T. Jeong, S. Choi, H. K. Kang, C. Chung, *IEEE International Electron Devices Meeting (IEDM)* **2011**, 31.8.1.
[203] H. Y. Chen, S. Yu, B. Gao, P. Huang, J. Kang, H. S. P. Wong, *IEEE International Electron Devices Meeting (IEDM)* **2012**, 20.7.1.
[204] S. Lee, J. Sohn, Z. Jiang, H. Chen, H. P. Wong, S. Lee, *Nat. Commun.* **2015**, *6*, 8407.
[205] Y. Yang, J. Lee, S. Lee, C. H. Liu, Z. Zhong, W. Lu, *Adv. Mater.* **2014**, *26*, 3693.
[206] H. Tian, H. Zhao, X.-F. Wang, Q.-Y. Xie, H.-Y. Chen, M. A. Mohammad, C. Li, W.-T. Mi, Z. Bie, C.-H. Yeh, Y. Yang, H.-S. P. Wong, P.-W. Chiu, T.-L. Ren, *Adv. Mater.* **2015**, *27*, 7767.
[207] A. Behnam, F. Xiong, A. Cappelli, N. C. Wang, E. A. Carrion, S. Hong, Y. Dai, A. S. Lyons, E. K. Chow, E. Piccinini, C. Jacoboni, E. Pop, *Appl. Phys. Lett.* **2015**, *107*, 123508.
[208] C. Ahn, S. W. Fong, Y. Kim, S. Lee, A. Sood, C. M. Neumann, M. Asheghi, K. E. Goodson, E. Pop, H.-S. P. Wong, *Nano Lett.* **2015**, *15*, 6809.
[209] J. Sohn, S. Lee, Z. Jiang, H. Chen, H. P. Wong, *IEEE International Electron Devices Meeting (IEDM)* **2014**, 5.3.1.
[210] Y. Deng, H. Y. Chen, B. Gao, S. Yu, S. C. Wu, L. Zhao, B. Chen, Z. Jiang, X. Liu, T. H. Hou, Y. Nishi, J. Kang, H. S. P. Wong, *IEEE International Electron Devices Meeting (IEDM)* **2013**, 25.7.1.
[211] P. Cappelletti, *IEEE International Electron Devices Meeting (IEDM)* **2015**, 10.1.1.
[212] N. Chandrasekaran, *IEEE International Electron Devices Meeting (IEDM)* **2013**, 13.1.1.
[213] A. D. Liao, A. Behnam, V. E. Dorgan, Z. Li, E. Pop, *IEEE International Electron Devices Meeting (IEDM)* **2013**, 15.1.1.
[214] G. H. Kim, K. M. Kim, J. Y. Seok, M. H. Lee, S. J. Song, C. S. Hwang, *J. Electrochem. Soc.* **2010**, *157*, G211.
[215] M. M. Shulaker, T. F. Wu, A. Pal, L. Zhao, Y. Nishi, K. Saraswat, H. P. Wong, S. Mitra, *IEEE International Electron Devices Meeting (IEDM)* **2014**, 27.4.1.
[216] G. Gupta, M. B. A. Jalil, G. Liang, *IEEE International Electron Devices Meeting (IEDM)* **2013**, 32.5.1.
[217] A. Fantini, L. Goux, S. Clima, R. Degraeve, A. Redolfi, C. Adelmann, G. Polimeni, Y. Y. Chen, M. Komura, A. Belmonte, D. J. Wouters, M. Jurczak, *IEEE International Memory Workshop (IMW)* **2014**, 1.
[218] B. J. Choi, J. Zhang, K. Norris, G. Gibson, K. M. Kim, W. Jackson, M.-X. M. Zhang, Z. Li, J. J. Yang, R. S. Williams, *Adv. Mater.* **2016**, *28*, 356.
[219] B. Govoreanu, G. S. Kar, Y. Chen, V. Paraschiv, S. Kubicek, 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. Altimime, D. J. Wouters, J. A. Kittl, M. Jurczak, B.- Leuven, K. U. Leuven, *IEEE International Electron Devices Meeting (IEDM)* **2011**, 31.6.1.
[220] T. H. Park, S. J. Song, H. J. Kim, S. G. Kim, S. Chung, B. Y. Kim, K. J. Lee, K. M. Kim, B. J. Choi, C. S. Hwang, *Sci. Rep.* **2015**, *5*, 15965.
[221] T. H. Park, S. J. Song, H. J. Kim, S. G. Kim, S. Chung, B. Y. Kim, K. J. Lee, K. M. Kim, B. J. Choi, C. S. Hwang, *Phys. status solidi – Rapid Res. Lett.* **2015**, *9*, 362.
[222] R. Penrose, *Shadows of the Mind: A Search for the Missing Science of Consciousness*, Oxford University Press, Oxford **1996**.

