25
ReRAM-Based Neuromorphic Computing

Giacomo Indiveri, Eike Linn, and Stefano Ambrogio

25.1
Neuromorphic Systems: Past and Present Approaches

Artificial neural networks are currently being used as a promising means to solve problems in machine learning and computer science [1–3]. These networks are loosely inspired by massively parallel biological neural systems, but they are often implemented on mostly sequential hardware computing platforms, ignoring key features of real neural processing systems, such as their ability to carry out robust and efficient computation using noisy components with limited resolution, which are highly variable and unreliable, and which have extremely low-power consumption characteristics. Neuromorphic systems represent a new and alternative class of neural network hardware architectures that attempt to implement these features by employing hybrid analog–digital electronic circuits that directly emulate the physics of neural processing elements by exploiting the physics of the devices and materials used [4, 5]. Recent developments in nanotechnologies are making available not only extremely compact and low-power but also highly variable and unreliable solid-state devices with memristive properties that can potentially extend the offerings of current neuromorphic systems [6–9]. In particular, memristors are considered as a promising solution for modeling key features of biological synapses due to their nanoscale dimensions, their capacity to store multiple bits of information per element and the low energy required to change their state.

25.2
Neuromorphic Engineering

Although the history of implementing electronic models of neural circuits extends back to the construction of perceptrons [10] and retinas [11], the modern wave of research utilizing very-large-scale integration (VLSI) technology and emphasizing the nonlinear current characteristics of the transistor began in

1) Text adapted from Ref. [5].
the mid-1980s with the collaboration that sprung up between prominent scientists Max Delbrück, John Hopfield, Carver Mead, and Richard Feynman [12]. Inspired by graded synaptic transmission in the retina, Mead sought to use the graded (analog) properties of transistors, rather than simply operating them as on–off (digital) switches. He showed that analog neuromorphic circuits share many common physical properties with proteic channels in neurons [4]. As a consequence, these types of circuits require far fewer transistors than digital approaches to emulating neural systems.

Through the *Physics of Computation* course at Caltech (led by Carver Mead, John Hopfield, and Richard Feynman), Mead’s textbook *Analog VLSI and Neural Systems* [4], and the creation of the *Telluride Neuromorphic Engineering Workshop*, the field of Neuromorphic Engineering was established. Prominent in the early expansion of the field were scientists and engineers such as Christof Koch, Terry Sejnowski, Rodney Douglas, Andreas Andreou, Paul Mueller, Jan van der Spiegel, and Eric Vittoz, training a generation of cross-disciplinary students.

It has been argued that neuromorphic circuits are ideal for developing a new generation of computing technologies that use the same organizing principles of the biological nervous system [13–15]. In addition to the computations of a single neuron, many neuromorphic circuits also utilize spiking representations for communication, learning and memory, and computation. The use of asynchronous spike-based or digital event-based representations in electronic systems can be energy efficient and fault tolerant, making them ideal for building modular systems and creating complex hierarchies of computation. The most successful neuromorphic systems to date have been single-chip devices that emulate peripheral sensory transduction such as silicon retinas, visual motion sensors, and silicon cochleas for a wide variety of applications.

In recent years, many larger multichip neuromorphic systems have begun to emerge that have raised new issues and challenges. Neuromorphic engineering now aims to use these technologies for developing larger-scale neural processing systems and move from the predominantly feed-forward, reactive neuromorphic systems of the past to adaptive behaving ones that can be considered cognitive [5, 16].

25.3
Neuromorphic Computing (The Present)

The idea of linking the type of information processing that takes place in the brain with theories of computation and computer science (i.e., *neurocomputing*) dates back to the origins of computer science itself [17, 18]. Neurocomputing has been very popular in the past [10, 19], eventually leading to the development of the artificial neural networks field, with both software simulations and dedicated hardware implementations, applied to a wide variety of practical problems [1, 20–23].
Both artificial neural networks and neuromorphic computing architectures are now receiving renewed attention thanks to progress in information and communication technologies (ICTs) and to the advent of new promising nanotechnologies. Some of present day neurocomputing approaches attempt to model the fine details of neural computation using standard technologies. For example, the Blue Brain project, launched in 2005, made use of a 126 kW Blue Gene/P IBM supercomputer to run software that simulated with great biological accuracy the operations of neurons and synapses of a rat neocortical column [24]. Similarly, the BrainScaleS EU-FET FP7 project had the goal of developing a custom neural supercomputer by integrating standard CMOS analog and digital VLSI circuits on full silicon wafers to implement about 262 000 I&F neurons and 67 million synapses [25]. Although configurable, the neuron and synapse models were hardwired in the silicon wafers, and the electronic models ran about 10 000 times faster than real biology, with each wafer consuming about 1 kW of power, excluding all external components.

Another large-scale neurocomputing project based on conventional technology is the SpiNNaker project [26]. The SpiNNaker is a distributed computer, which interconnects conventional multiple integer precision multi-ARM core chips via a custom communication framework. Each SpiNNaker package contains a chip with 18 ARM9 CPU on it, and a memory chip of 128 Mbyte Synchronous DRAM. Each CPU can simulate different neuron and synapse models. If endowed with simple synapse models, a single SpiNNaker device ARM core can simulate the activity of about 1000 neurons in real time. More complex synapse models (e.g., with learning mechanisms) require more resources and decrease the number of neurons that can be simulated in real time. A complete SpiNNaker board contains 47 of these packages, and the goal is to assemble 1200 of these boards. A full SpiNNaker system of this size would consume about 90 kW.

The implementation of custom large-scale spiking neural network hardware simulation engines is being investigated also by industrial research groups. For example, IBM recently unveiled a multicore spiking neural network chip called “TrueNorth” [27], comprising one million neurons distributed across multiple “neurosynaptic” cores. Each of these cores comprises 256 digital I&F neurons, with 1024 × 256 binary valued synapses, configured via an SRAM cross-bar array, and uses an asynchronous event-driven design to route spikes from neurons to synapses. The goal is to simulate networks of simplified spiking neurons with human-brain dimensions (i.e., approximately $10^{10}$ neurons and $10^{14}$ synapses) in real time. In the mean time, IBM simulated 2.084 billion neurosynaptic cores containing $53 \times 10^{10}$ neurons and $1.37 \times 10^{14}$ synapses in software on the Lawrence Livermore National Lab Sequoia supercomputer (96 Blue Gene/Q racks), running 1542× slower than real time [28], and dissipating 7.9 MW.

A diametrically opposite approach is represented by the Neurogrid system [29]. This system comprises an array of sixteen $12 \times 14$ mm$^2$ chips, each integrating mixed analog neuromorphic neuron and synapse circuits with digital asynchronous event routing logic. The chips are assembled on a 16.5 × 19 cm$^2$ PCB, and the whole system can model over one million neurons connected by billions of synapses in real-time, and using only about 3 W of power [30]. As opposed
to the neurocomputing approaches that are mainly concerned with fast and large simulations of spiking neural networks, the Neurogrid has been designed following the original neuromorphic approach, exploiting the characteristics of CMOS VLSI technology to directly emulate the biophysics and the connectivity of cortical circuits. In particular, the Neurogrid network topology is structured by the data and results obtained from neuroanatomical studies of the mammalian cortex. Although offering less flexibility in terms of connectivity patterns and types of synapse/neuron models that can be implemented, the Neurogrid is much more compact and dissipates orders of magnitude less power than the other neurocomputing approaches described earlier.

25.4
Neuromorphic ReRAM Approaches (The Future)

As discussed in the previous section, most neuromorphic approaches are based on pure CMOS technology. However, the soon projected availability of novel nanodevices [31] offers some unique properties that are highly advantageous for implementing hardware synapses [32]. Especially, redox-based resistive switching random access memory (RAM) devices (ReRAM) [33], also called memristive devices [34, 35], are considered to be very promising candidates for enabling high-density and ultimately scaled synaptic arrays in neuromorphic architectures. ReRAM cells are two terminal devices, which can be arranged in passive crossbar arrays of minimum feature size or potentially integrated in 3D synaptic arrays [6, 36]. The scaling properties of ReRAM cells have been shown to be excellent and could pave the path to very energy-efficient neuromorphic hardware systems (see Section 2.2 for details on the scaling aspects).

In Section 25.4.1, we briefly review the available ReRAM-based neuromorphic approaches. Then, in the following sections, we discuss the key properties of ReRAM useful for realization of hardware synaptic functions in more detail, showing how, depending on the approach followed, one or more of the following properties are exploited:

- nonvolatility/volatility of resistive states
- nonlinear switching kinetics
- multilevel resistance behavior
- capacitive properties
- switching statistics.

25.4.1
ReRAM-Based Neuromorphic Approaches

The basic idea in most ReRAM-based neuromorphic approaches is to consider ReRAM devices, or small ReRAM-based circuits, as artificial synapses. The idea of using ReRAM devices for neuromorphic applications goes back
to Likharev [37], who introduced the concept of “Crossnets,” where ReRAM devices serve as programmable interconnects, that is, binary synapses. On top of CMOS-based neurons, a crossbar array of nanoscaled ReRAM devices is used for reconfigurable wiring of the neurons. An example is given in Ref. [38], and shown in Figure 25.1a, indicating at the top a crossbar array of ReRAM devices providing reconfigurable, binary connection to the CMOS front-end circuit underneath. Binary ReRAM synapses could also be directly used to implement associative memories [39]. Note that the ReRAM crossbar array requires select elements [31] as shown in Figure 25.1b (see Chapters 22 and 24 for more details).

In a more general approach, the tunable resistive state of ReRAM device is used as a synaptic weight, which can be adapted by suitable methods for changing ReRAM resistance in analogy with synaptic plasticity rules. This is schematically shown in Figure 25.2a: by updating the weight of the synapses, the electrical connection between a presynaptic neuron and a postsynaptic neuron changes, thus enabling the possibility of implementing a variety of learning models for pattern recognition and memory storage.

The weight of the synapses is typically updated following the learning rule called STDP, where the conductance change of the synapse depends on the relative timing of the electrical pulses delivered from the presynaptic neuron and the postsynaptic neuron. If the presynaptic spike precedes the postsynaptic one, the synapse conductance is enhanced, whereas if the postsynaptic spike precedes the presynaptic one, the synapse is depressed [24, 41, 43]. Experimental results measured from a rat-hippocampal neuron are shown in Figure 25.2b [41].
The first experimental evidence of the feasibility of the STDP learning curve with ReRAM devices was reported by Jo et al. [42]. Figure 25.2c shows the experimentally measured STDP, namely the conductance update as a function of interspike delay. The STDP characteristic shows a conductance increase for a time delay $\Delta t = t_{\text{pre}} - t_{\text{post}} < 0$, that is, for the presynaptic spike preceding the postsynaptic spike. Conversely, when the presynaptic spike precedes the postsynaptic ($\Delta t < 0$), the ReRAM conductance is decreased. The resulting STDP characteristic shows an exponentially decaying behavior, which is consistent with biological data in Figure 25.2b.

The experimental demonstration of memristive STDP in Figure 25.2c was achieved by the approach presented by Snider [44]. Figure 25.3 schematically shows the ReRAM-based synaptic array (a) and the spiking scheme to achieve STDP behavior in the synapse (b). According to this idealized scheme, a single spike cannot alter the resistive state of the synapse, and a weight update requires the presence of a postsynaptic spike and a presynaptic spike overlapping in time.
Figure 25.3 (a) ReRAM-based synaptic array and (b) spiking scheme to either induce long-term potentiation (LTP) or long-term depression (LTD) in the synapse. (Reprinted with permission from Ref. [44].)

Note that the pulse width of the spikes gradually decreases with time according to an exponential law. Moreover, the synaptic weight can be either increased or decreased, inducing either long-term potentiation (LTP) or long-term depression (LTD) changes in the synapse.

25.4.2 Nonvolatility and Volatility of Resistive States

The nonvolatility of the resistive states, that is, the hysteretic memristive behavior, is one of the most appealing ReRAM properties to enable synaptic functionality. In general, ReRAM devices offer long-term state retention up to 10 years at typical temperatures below 85 °C [45]. However, this is not true for all ReRAM devices and even the same ReRAM device can display volatile/nonvolatile retention depending on the programming/erasing conditions. In Reference [46], the feasibility of short-term plasticity (STP) and LTP was demonstrated for Ag₂S-based devices. Figure 25.4a shows the STP behavior of a ReRAM device under the application of a sequence of pulses of 0.5 s pulse width, separated by a relatively long interpulse delay \( T = 20 \) s. Note the volatile change of weights, where the conductance is only transiently increased with no stable potentiation even after nine pulses. In contrast, Figure 25.4b shows that a permanent weight change to high conductance can be achieved when the interpulse delay is reduced to \( T = 2 \) s. Thus, the volatility of high ohmic ON states enables implementation of STP within a single ReRAM device depending on the spike rate, whereas the nonvolatility of the permanent high-conductance state enables implementation of LTP. Similarly, internal nonequilibrium states that cause an emf (electromotive force) voltage [47] may also influence the resistance states [48]. This means that ReRAM states can offer exponential forgetting, that is, the resistive state is varied exponentially with time, depending on internal nonequilibrium states. In general, internal nonequilibrium states allow for using biorealistic nonoverlapping input spikes that enable
the emulation of different synaptic plasticity effects, for example, the behavior of the cell’s Ca\textsuperscript{2+} concentration as a second-order memristive effect [49, 50].

25.4.3
Nonlinear Switching Kinetics

ReRAM devices in general offer a highly nonlinear switching kinetics [51–54], where the set time $t_{\text{set}}$ exponentially depends on the applied pulse height (see Chapter 11). This feature can be used to implement STDP in a simple manner [44, 55]. For a certain pulse length, one can define a threshold voltage below which no switching occurs for the specific pulse duration. If we select a voltage $V_{\text{pulse}}$ as the presynaptic spike voltage and $-V_{\text{pulse}}$ as the postsynaptic spike voltage, which are both well below the threshold voltage, the application of either the presynaptic or postsynaptic pulse alone would lead to a negligible change of the synaptic weight. However, if pre- or postsynaptic signals occur simultaneously, the total voltage at the junction will be equal to $2V_{\text{pulse}}$, which shall be well above the threshold voltage. In this case, therefore, the synaptic weight will significantly be affected, thus enabling STDP. Figure 25.5 shows a corresponding STDP implementation according to [55].

To improve the accuracy of the spike shape in terms of biological signals, a similar STDP approach was proposed in Ref. [56]. Various STDP schemes using memristive devices are reviewed in Ref. [57], where the impact of ReRAM device models (see Chapters 13, 14, and 17) is also discussed.

25.4.4
Multilevel Resistance Behavior

Another useful property of ReRAM devices is the capability of multilevel resistances [53, 58–60]. Multiple levels offer the possibility to have multiple discrete
weights of the synapses instead of binary weights [9, 32, 46, 60–65]. Gradual tunable ReRAM devices offer the potential of mimicking biological synaptic behavior more realistically than bistable devices (e.g., SRAM-based synapses). However, for typical ReRAM devices, the set process is an abrupt process. This is shown in the $I-V$ curve of an HfO$_x$ ReRAM device in Figure 25.6a, where the set transition at positive voltage shows a vertical change of the current, while the reset transition is more gradual [60]. To reliably tune the resistance to a given analog value, either a current compliance (e.g., $I_c = 500$ $\mu$A in Figure 25.6a) or a series resistor are required. Figure 25.6b shows the control of resistance state during set transition by using increasing compliance currents, from $I_{c1}$ to $I_{c3}$, resulting in a decreasing value of resistance. The compliance current $I_c$ can be controlled by the voltage applied to the gate of the transistor in a 1T1R structure [60]. Furthermore, the resistance during reset transition can be controlled by the value of the stop voltage as shown in Figure 25.6c.

To take advantage of the analog tuning of resistance in both the set and reset transitions, a 2-transistor/1-resistor (2T1R) synapse capable of both communication and plasticity was proposed [66]. In the 2T1R synapse, the resistor consists of a bipolar HfO$_x$ ReRAM connected to two transistors in parallel, as displayed in Figure 25.7a. The synapse circuit has three access terminals, namely (i) the top electrode (TE), (ii) the communication gate (CG), and (iii) the fire gate (FG). The bottom electrode (BE) serves as output of the current from the synapse to a postsynaptic neuron (POST). The 2T1R synapse is inserted between two neurons, namely a presynaptic neuron (PRE) driving TE and CG with voltage spikes, and POST that receives the current resulting from a PRE spike and driving the
Figure 25.6 (a) Exemplary I-V curve of an HfO\textsubscript{2} ReRAM cell. (b) I-V curves at positive voltage, where resistance after set transition decreases at increasing compliance current ($I_{c1} = 0.5$ mA, $I_{c2} = 0.75$ mA and $I_{c3} = 1$ mA). (c) I-V curves at negative voltage, where resistance after reset transition increases at increasing stop voltage ($V_{\text{stop},1} = 0.5$ V, $V_{\text{stop},2} = 0.7$ V and $V_{\text{stop},3} = 0.9$ V). Reprinted with permission from Ref. [60].

Figure 25.7 Schematic layout of the 2T1R synapse circuit (a) and experimental STDP characteristic (b), showing the probability for a conductance change $R_0/R$ as a function of the delay $\Delta t$ between PRE/POST synaptic pulses. (Reprinted with permission from Ref. [66].)

FG by a pulse (fire) generated in correspondence of the POST spike. Every neuron behaves as an integrate & fire circuit, capable of integrating all the current received at the input node and transforming the current in an internal (membrane) potential. Once the membrane potential reaches a characteristic threshold, a fire event is generated, where the POST delivers a spike pulse driving the FG terminal of all synapses. If overlapped with the TE pulse delivered by the PRE, this pulse can result in set or reset transitions in the ReRAM element, thus inducing synapse potentiation/depression and enabling learning in the neuromorphic network. The shape of the TE and spike pulses causes a specific STDP response; namely LTP/LTD is achieved when the pulse is following/preceding the PRE spike (delay $\Delta t > 0$ or $\Delta t < 0$, respectively).

Figure 25.7b shows the experimental STDP characteristic for the 2T1R synapse [66]. To characterize this behavior in the synapse, 1T1R samples with HfO\textsubscript{x} ReRAM elements were subjected to random spikes with variable delay $\Delta t$ and random initial state (LRS, HRS, and various intermediate states). The STDP was then reported as the probability to display a conductance change $R_0/R$ for
a given delay $\Delta t$ (see color scale of Figure 25.7). The statistical representation is needed due to the largely stochastic behavior of the ReRAM device, which can be explained both by the switching variability in the set/reset processes of the HfO$_x$ memristor [67], and the dependence of STDP response to the initial resistance $R_0$ (i.e., set state, reset state, and intermediate state). Learning of a visual pattern using this scheme has been demonstrated by simulations of a 2-layer neural network (64 neurons in the first layer, 1 neuron in the second layer) linked by 64 2T1R synapses [66]. The results support the ability to learn visual/auditory patterns by gradually changing the synapse weight in a neural network, as opposed to the case of binary abrupt switching [68].

25.4.5

**Capacitive Properties**

Although ReRAM devices are generally considered as resistive devices, the metal/insulator/metal (MIM) structure of ReRAM also offers a capacitance in parallel [69]. This capacitance can be exploited when considering complementary resistive switch (CRS) structures as binary synapses as described in Figure 25.8a. The resistance of elements A and B can be switched between the high resistance state (HRS) and the low resistance state (LRS), whereas the capacitance is controlled by the device geometry and the permittivity of the insulator material (see Figure 25.8b). For the logic state 0, element A is in the LRS, and thus the capacitance $C_A$ is short-circuited. In this case, the overall capacitance is determined by $C_B$. Similarly, for state 1, element B is short-circuited and capacitance $C_A$ dominates the overall device behavior. As a result, the stored information of the CRS cell influences the detectable capacitance, which thus can be read out nondestructively. In Figure 25.8b a capacitive voltage divider is formed with $C_{out}$, and $V_{out}$ is evaluated using a sense amplifier. The controllable capacitance feature of this device can be used to extend the functionality of the binary synapse and may enable pattern recognition tasks [70].

25.4.6

**Switching Statistics**

ReRAM devices are known to display a switching statistics where set/reset parameters (resistance, voltage, and current) randomly change from cycle to cycle [32, 67]. The stochastic switching can be used to emulate the nondeterministic update of synaptic weight, which has also been observed in biological synapses. From the biological point of view, synapses are characterized by a synaptic conductance or weight $w$ [71, 72], which depends on three different processes, namely

$$w = abc,$$  \hspace{1cm} (25.1)

where $a$ represents the number of quantal neurotransmitter release sites, or channels, $b$ is the probability of synaptic release for each site, and $c$ represents a measure of the postsynaptic effect of the synapse [72]. The majority of the STDP models
and demonstrations consider \( a \) and \( b \) to be constant, leaving all the weight tunability in the parameter \( c \), which gives the form of the STDP learning rule [73]. Analog STDP models thus neglect the natural variability of the number of release sites and probability of synaptic release. Another equivalent approach consists in emulating biological synapses by leaving \( c \) constant and varying the transmission probability \( b \). This method fully exploits the variability of nanoscale ReRAM, since deterministic multilevel and stochastic binary synapses are functionally equivalent [71]. This enables synapses, hence ReRAM devices, to assume only two different resistance levels irrespective of their precise values.

Figure 25.9 shows the switching probability for reset (a) and set (b) for a Ag-based ECM cell in 1T1R configuration. By applying relatively weak switching conditions, a certain switching probability, for example, 50%, can be achieved. The switching probability is a function of both the pulse length and the applied voltage, as shown in Figure 25.9. For a given gate voltage (\( V_g = 1.5 \) V in the figure) and pulse length (500 ns and 1 us), the set/reset probability increases with the applied voltage across the 1T1R. For increasing pulse width, the voltage to achieve a given switching probability decreases due to the time–voltage relationship for set/reset processes [54]. For ReRAM devices that exhibit abrupt set/reset transitions, thus most suitable for binary synapses, stochastic switching allows to gradually tune the synapse weights for learning and recognition, even with no multilevel capabilities [68]. This method reveals that variability arising from scaling could provide some advantage in the design of new neuromorphic architectures.

Another way to practically deal with variability is similar to the previously described method and consists in directly emulating the single biological ion channels, which are affected by stochasticity, in the synaptic cleft through a resistive element. In this way, one resistive device can represent one ion channel, thus employing the intrinsic variability of the nanoelectronic cell to emulate
25.4 Neuromorphic ReRAM Approaches (The Future)

Figure 25.9 Switching probability of a 1T1R Ag-based ECM cell. Two different pulse lengths of 500 ns and 1 μs are considered, and a gate voltage of $V_g = 1.5$ V is applied.

(a) RESET probability for bitline voltages in the range of 1.5–2.5 V. (b) SET probability for anode voltages of 1.5–2.5 V. (Reprinted with permission from Ref. [8].)

Figure 25.10 Probabilistic STDP learning rule used in Ref. [68]. $\Delta t$ is defined as the difference between the postsynaptic and the presynaptic spike. (Reprinted with permission from Ref. [68].)

the intrinsic variability of the single-ion channel [32]. Therefore, one synapse is implemented through a number of parallel ReRAMs.

However, scaled ReRAM can allow for a deterministic STDP learning rule by means of a gradual resistance modulation. Yu et al., [62] demonstrated the possibility of employing scaled devices with elevated variability by programming them with very short voltage pulses with a time width of 10 ns. The application of many short pulses allows to overcome the variability of the device since, averaging on many pulses, the device resistance follows a deterministic behavior. Figure 25.11 shows three different resistance states, which are gradually reset through the application of hundreds of pulses. It is thus possible to gradually vary and control the resistance change, also allowing for sub-pJ switching programming energy. With the described method, it has been simulated a robustness of an entire video pattern recognition system at typical experimental variability levels with resistance fluctuations up to 9%. Only a slight decrease of the selectivity properties has been observed with an additional increase of the variability of resistance values.
25.5 Scaling in Neuromorphic ReRAM Architectures

Nowadays, Moore’s law is reaching its ultimate scaling limits, due to intrinsic variability and excessive off-state leakage issues in CMOS technology. To extend Moore’s law in future generations, the semiconductor industry is considering new technologies, such as ReRAM devices, allowing for denser structures and more compact memory and logic architectures [74]. The same need for scaling is even more important for neuromorphic applications. Indeed, human brains have approximately $10^{11}$ neurons and $10^{14}$ synapses. Implementing brain-like processing architectures that comprise even a small fraction of these elements, without employing huge and expensive silicon areas, requires an aggressive scaling of devices and optimization of their organization in dense and low-power passive architectures. In this frame, ReRAM technology is a promising candidate for neuromorphic computing, thanks to its low power consumption (<10 fJ/bit) [75], fast switching (<10 ns) [76], multibit storage [59], extremely small size (<10 nm) [77], and capability of integration in a $4F^2$ area, where $F$ is the minimum feature size. However, despite these promising aspects, several serious issues arise when scaling down the device area, which can be partly overcome with intelligent and alternative neuromorphic approaches.

One of the key issues for scaling down ReRAM devices is the variability of the switching parameters [67, 78]. Recent works have pointed out that switching variations are due to discrete injection of ionized defects [79], which can be modeled by a Poisson distribution [67, 78]. The scaling down of the current and of the CF size also gives rise to undesired random telegraph noise (RTN), which affects the device resistance during read operation [80]. Reducing the filament dimensions also leads to structural instabilities that, through chemical dissolution and electromigration, cause unwanted resistance increase from the set state [81, 82].
As the interconnect lines within the crossbar array are downscaled, the series resistance due to the metal lines might raise additional issues. For instance, the resistivity of Cu, which is generally used for low-resistance metal interconnect in memory circuits, is going to increase over $10 \mu \Omega \text{cm}$ [31] as the cross section is reduced to roughly $10 \times 10 \text{nm}^2$. As a result, the effective voltage drop across the ReRAM device in the array might be significantly smaller than the applied voltage, due to the voltage drop across the interconnections. To solve this issue, the resistance of the device must be significantly higher than the metal lines, which can be achieved by small filaments obtained at small operation current during set [67]. This, in turn, leads to size-dependent reliability issues such as statistical fluctuations and RTN, as previously discussed.

The problems of device scaling can be mitigated when considering neuromorphic implementations: neuromorphic networks have the advantage of being able to deal with unreliable elements as real biological synapses, thanks to a highly parallel computation. For this reason, scaling drawbacks of ReRAM such as switching variability and noise do not represent critical issues for neuromorphic circuits.

25.6 Applications of Neuromorphic ReRAM Architectures

Potential applications that can best exploit the properties of neuromorphic ReRAM architectures are those that involve the processing of real-world signals and that require compact and low-power devices. Typical examples can be found in the robotic domain, in intelligent brain–machine interfaces, in sensor networks, embedded systems, and portable devices. As neuromorphic computing is mainly concerned with the implementation of the computational principles used by the nervous system, most existing neuromorphic systems adopt learning mechanisms in neural networks to solve computational problems [32, 56, 57, 71, 83–85]. In these systems, synapses represent the essential component for both learning and signal processing. Synaptic weight values can be stored in ReRAM devices and updated following the prescription of different types of learning algorithms that can depend on the pre- and postsynaptic neuron activity [61, 86, 87]. In general, within this context, much research has been dedicated to the use of resistive synaptic devices [32, 88] for implementing learning and neural computation, with classification and pattern recognition being the most explored applications [62, 63, 68, 89]. A typical example of a pattern recognition application is in the domain of auditory signals: for example, in Ref. [68] the authors first analyze auditory signals by an array of band-pass filters; then they connect each band-pass filter output to a corresponding input neuron, which is part of the first neural layer (see Figure 25.12). All neurons in this input layer are then connected to a single-output neuron, via CBRAM synapses (three per connection). The output neuron is then trained, following a classical perceptron learning scheme [10]. As training proceeds, the audio pattern is memorized, selectivity becomes very high, and the number of errors decreases. It is possible
to imagine multiple applications in the auditory scene analysis domain, using this scheme.

Another common example of pattern recognition comes from image processing of natural scenes. In the example shown in Figure 25.12b from Suri et al., [68], an address-event representation (AER) silicon retina [90] provides transient visual sensory signals to a neural network layer, via synapses implemented by resistive devices. A second neural layer, again connected through CBRAM synapses to the first layer, constitutes the output layer. A visualization of the visual information present at each level of the network is shown on the right, starting from the sensor’s output up to the top layer output, in which only the trained pattern is highlighted (which was set to be the smaller pattern on the right, representing a car on the right lane of a freeway). Similar visual pattern recognition applications have been proposed by Zamarreño-Ramos et al. [56], Thorpe [89], Yu et al. [62], Park et al. [63].

As these examples demonstrate, the neuromorphic approach makes use of massively parallel architectures, that typically operate using hybrid analog and digital components, often with inhomogeneous or even unreliable outputs. This practically means that developing computing architectures that can operate using unreliable devices is potentially not a problem, as some studies have already pointed out [55]. Neuromorphic architectures can also exploit the probabilistic nature of the components they have and the signals they process, without using hard logic [32, 91]. Therefore, they are an ideal medium for exploiting the properties of memristive devices and applying them to the solution of real-world problems.

In conclusion, the neuromorphic approach with memristive devices is promising for a new class of applications and revolutionary logic perspectives. For the vast majority, these types of applications have only been simulated up to now. Many efforts are still needed to improve the memristive switching performances and integrate them onto standard CMOS processes, but real memristive neuromorphic future applications are not inconceivable.
References

8. Suri, M., Bichler, O., Querlioz, D., and Palma, G. (2012) CBRAM devices as binary synapses for low-power stochastic neuromorphic systems: auditory (Cochlea) and visual (Retina) cognitive processing applications. *International Electron Devices Meeting (IEDM)*, pp. 10.3.1–10.3.4.


References


78. Fantini, A., Goux, L., Degraeve, R., Wouters, D.J., Raghavan, N., Kar, G.,


