Spike-driven threshold-based learning with memristive synapses and neuromorphic silicon neurons

E Covi[*1], R George[*2‡], J Frascaroli[1], S Brivio[1], C Mayr[2‡], H Mostafa[2§], G Indiveri[2] and S Spiga[1]

[*] These Authors contributed equally to this work.

[1] CNR-IMM, Unit of Agrate Brianza, Italy.

[2] Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland.

E-mail: erika.covi@mdm.imm.cnr.it, sabina.spiga@mdm.imm.cnr.it

Abstract.

Biologically plausible neuromorphic computing systems are attracting considerable attention due to their low latency, massively parallel information processing abilities, and their high energy efficiency. To achieve these features neuromorphic silicon neuron circuits need to be integrated with plastic synapse circuits capable of on-line learning and storage of synaptic weights. Within this context, memristive devices play a key role thanks to their non-volatility, scalability, and compatibility with Complementary Metal-Oxide-Semiconductor (CMOS) fabrication process. However, practical implementations of neuro-memristive systems are still facing difficult challenges for implementing efficient learning protocols. Here we propose and demonstrate in hardware a spike-driven threshold-based learning rule which goes beyond conventional Spike-Timing Dependent Plasticity (STDP) mechanisms, by taking into account also the neuron membrane potential and its firing rate. The mixed memristive-neuromorphic system we demonstrate comprises an oxide-based memristive synapse device placed between two silicon neurons implemented on a neuromorphic chip that comprises the proper interfacing and spike-based learning circuits designed to drive the memristive elements. We show how the system is able to emulate in real-time weight dependent post-synaptic activity and drive synaptic weight updates at the memristive synapse level following the spike-driven learning rule presented. We validate this spike-based learning mechanism with experimental results and quantify the system performance with basic learning experiments.

‡ Present address: Institute of Circuits and Systems, Technical University of Dresden, Germany.

§ Present address: Institute of neural computation, University of California San Diego, USA.

1. Introduction

Neural networks and deep learning algorithms are currently achieving impressive state-of-the-art results in a wide range of application areas [1–3]. At the same time, computational neuroscience has made tremendous progress in uncovering key principles of computation used by the brain [3–5]. However, despite the remarkable progress, today's artificial electronic systems are still not able to compete with biological ones in real-time sensory and cognitive tasks carried out in complex and uncertain settings. Understanding how to use the brain's computational principles to build compact ultra-low power cognitive computing systems based on unreliable and inhomogeneous nanoscale components has become a goal of utmost importance for the information and communication technology (ICT) industry. This is especially true now that Moore's scaling law, the main driving force behind the tremendous evolution of all electronic devices for the last 50 years, is predicted to come to an end at the horizon of 2020-2025 [6].

Mixed memristive-neuromorphic electronic systems offer a technology that can pave the road towards this objective [3, 7–10]. These systems combine in an optimal way the properties of standard CMOS technologies and those of emerging resistive memory technologies to emulate the biological neuron-synapse systems, where synapses can facilitate (potentiation) or inhibit (depression) the connection between neurons. On the neuronal side, asynchronous analogue sub-threshold CMOS circuits are used to emulate the ion diffusion observed in biological neurons whilst keeping the currents in the pA to nA range [1, 11]. On the synaptic side, oxide-based resistive switching (RS) memristive devices are promising candidates as artificial synapses because they allow on-line update and non-volatile storage of synaptic weights, which map to the device conductance [8, 12–16]. Moreover, RS device programming voltages are comparable with the power supply of typical neuromorphic chips, which makes the on-chip integration of RS memristive devices less problematic with respect to other non-volatile devices such as floating gate transistors.

Efficient bio-inspired learning protocols are being developed [17–19], but their implementation still poses a significant challenge [20, 21]. One of the most widely used learning rules in current neuromorphic architectures is Spike-Timing Dependent Plasticity (STDP). STDP was first observed in biological neurons [22] and establishes a relationship between the timing of the relative firing activity between two neurons (within a ms-time window), namely the pre-synaptic and post-synaptic neuron, and the weight modification of the synapse that connects them. Accordingly, STDP is able to catch temporal correlations among events that occur within a time window of few ms. Usually proposed implementations of STDP rely on the overlap of pre- and post-spikes that are as long as the correlation time the learning rule is expected to be sensitive to. [23, 24] This expedient, though, poses serious issues related to synaptic matrix overload and constraints in the throughput of a network. Moreover, biological evidences and computational neuroscience studies have shown that the basic pair-wise STDP rule has critical limitations in both biological plausibility and computational power [25–27]. Indeed, STDP in computational models [17, 18] and in hardware implementations [23, 28–30] is carried out considering combinations of few

pre- and post-spikes, whereas biological experiments demonstrate that the overall spiking activity within a certain time frame (e.g., including spike rates) plays also an important role in determining the occurrence and nature (potentiation or depression) of the plasticity event [31]. New learning schemes involving spike rate [32, 33] and new plasticity models taking the rate of the synaptic activity into account [34–36] have therefore been proposed.

In this work, we pursue the same approach of going beyond basic STDP mechanisms, and demonstrate in memristor-CMOS hardware a spike-driven plasticity rule belonging to the class of models where plasticity occurs depending on the post-synaptic neuron state at the time of the pre-synaptic spike [34]. Indeed, the information about the temporal correlation of incoming events is coded in the post-synaptic neuron state variables that are evaluated in real-time by a analog and sub-threshold CMOS circuitry. According to the state variable values at the time of a pre-synaptic spike, the conductance of a HfO_2-based memristor [37] synapse is changed either between two separated states (binary operation) or throughout many adjacent values (analogue or multilevel operation) by digital plasticity circuits. The proposed implementation features a number of positive advantages. (i) The learning rule is implemented with pulses that are much shorter than the time correlation among incoming events that the network is expected to be sensitive to because the temporal information is stored in the neurons. This fact ensures an easier management of large memristive arrays with respect to using long overlapping pulses responsible for the coding of the temporal correlations. (ii) The plasticity circuits interfacing neuron and synapses allow the management of write and read phases in an asynchronous manner and, in particular, avoids the integration of the synaptic current during the programming phase, which would result in an improper evaluation of the synaptic efficacy because of current overshoot during the memristor switching. (iii) The temporal information over ms-time windows is provided by relatively small capacitor thanks to the sub-threshold CMOS design which reduces the charging-discharging currents down to the femto- to pico-Ampére range.

In summary, the present work opens the way towards the implementation of several stimulation protocols for life-long on-line learning.

2. Methods

The memristive synapse is fabricated at laboratory level and at micron scale area and it is a (bottom to top) 40 nm TiN / 5.5 nm HfO_2 / 50 nm Pt 40×40 μm^2 structure. The details of the fabrication process are reported in [38]. In an initial phase, the device is in its pristine high resistance state ($\sim$1 MΩ) and needs to be formed to initiate the switching. Indeed, the forming process creates for the first time a conductive filament shorting the two electrodes which can be partially dissolved and re-instated with depression and potentiation operations [39–41], respectively. The forming operation is carried out with a current sweep up to 1 mA. The device characterisation shown in Section 3.1, are carried out using a standard automatic test equipment (ATE), more specifically a B1500A semiconductor parameter analyser by Keysight, and a custom printed circuit board (PCB) interfacing the sample with the Source Pulse Generator Unit (SPGU) and the Source Measuring Units (SMUs) of the Keysight

B1500A, as described in [42]. Pulses applied to the device in Section 3.1 are trains of square pulses with rise and fall time of 100 ns, time width of either 50 μs or 100 μs and different voltage amplitudes.

In all the experiments of Section 3, time widths in the order of tens to hundreds of μs are used to program the device because they are compatible with real-time applications.

The CMOS chip is a full-custom analogue/digital neuromorphic VLSI chip fabricated using a standard 0.35 μm CMOS process and it includes special pads in the back-end to allow external wire connection of memristive devices. The parameters which determine the neuron behaviour (refractory period, firing threshold, time constant of the calcium variable...) as well as the thresholds used in the plasticity rule (see Section 3.3) and the parameters which shape the pre- and post-spike (voltage amplitude and time width) can be set externally by the user. The chip and the interfacing circuits have been presented and fully characterized in [4,11,43].

To configure the chip parameters, and acquire data for analysis, we interface the chip to a host PC using a field programmable gate array (FPGA), and a custom digital to analogue (DAC) board. Details on the setup built around the chip can be found in the Supplementary Information, Section S1.

All the experiments from Section 3.2 on are carried out by stimulating one silicon neuron through the host PC with constant injection current, thus making the neuron fire spikes at a constant rate (e.g., of 100 Hz in our experiments).

The memristive device is wired-connected to the neuromorphic chip, but for testing purposes we need the possibility of monitoring the device resistive state and of setting the device resistance at a determined state. As these operations cannot be carried out by the chip, we need to connect the ATE to the device as well.However, the chip keeps the terminals of the synaptic device at 1.65 V, whereas the ATE channels are initially at 0 V. Therefore, manually switching the device between chip and ATE would result in damaging the device. Hence, the need for designing a switch PCB, which is used to allow an automatic and protected transition of the memristive devices terminals between the ATE and the neuromorphic chip, we designed a custom switch PCB using reed relays. Further details on the switch PCB can be found in the Supplementary Information, Section S2.

3. Results

To demonstrate the spike-driven plasticity rule, we configured the hardware system to use two silicon neurons connected by a memristive synapse. In the following, we first analyze the properties of each stand-alone element, then we describe the overall system behaviour.

3.1. Synaptic plasticity in memristive devices

The role of the synapse in the system is played by the memristive device, which exhibits plastic behaviour when properly stimulated. Figures 1a and 1b show the device behaviour when stimulated with trains of depressing and potentiating pulses, respectively. The trains of pulses have different voltage amplitudes and the same time width of 100 μs. In these

experiments, the device was first brought to its Low Resistance State, LRS (High Resistance State, HRS), before being depressed (potentiated) by DC sweeps. It can be noted that the pulse voltage amplitude determines different device behaviours. As an example, in Figure 1a, when the voltage amplitude is low, the stimulation does not affect the device state (stimulation at -0.50 V). The increase of the voltage amplitude determines a change in the device resistance, which can be gradual, featuring several resistance levels (stimulation at -0.95 V), or abrupt, as in the case of -1.40 V stimulation, where only two states are distinguishable. Moreover, it should be noted that higher voltages correspond also to wider resistance windows. The same behaviour is observable during potentiation (Figure 1b): The pulse voltage determines whether the resistance state remains unchanged (stimulation at 0.50 V) or whether the device shows analogue (stimulation at 0.80 V) or digital (stimulation at 1.15 V) behaviour. Plasticity is therefore closely related to the choice of the spike amplitude and it is a trade-off between the number of different synaptic levels and the obtainable programming window. Indeed, with high depressing and potentiating voltages, the device shows a digital behaviour, as in the case of Figure 1c, where potentiation / depression cycles carried out with alternating pulses are shown. The LRS is stable, whereas the HRS shows higher variability. Despite HRS variability, the resistance window of about one order of magnitude is constant throughout the cycles. The reduction of the stimulation voltage with respect to the one used in digital operation results in a gradual modulation of the synaptic weight.At the beginning, a preliminary characterisation of the synaptic device is needed to set the parameters during the learning experiments. Figure 1d shows analogue potentiation / depression cycles carried out with 300 pulses of 100 μs and with amplitude of 0.8 V and -1 V, respectively. The HRS to LRS ratio is lower than in Figure 1c, but the cycles are repeatable and the synaptic weight is gradually modulated in both operations.

3.2. Silicon implementation of neurons

The silicon neuron (neuron element, NE) is the current CMOS implementation that faithfully reproduces the properties of an adaptive exponential integrate and fire model [44].The model reproduces the biological channels and ions which take part in the normal operation of biological neurons, as detailed in the Supplementary Information, Section S3. Figure 2 shows the analogies between a biological neuron and its silicon counterpart. In biology, when an action potential reaches the end terminals of a pre-synaptic neuron, it causes the release of neurotransmitters, which cross the synapse and bind to the neurosynaptic receptors of the post-synaptic neuron dendrites (biochemical signal transmission in the Figure). The biochemical signal is translated into an electric one, the excitatory post-synaptic current (EPSC, I_{EPSC} in the Figure), which then propagates to the neuron's soma. All the EPSC signals coming from each dendrite sum in the neuron soma, thus altering the neuron's membrane potential (V_{mem} in the Figure) in the axon hillock. When the membrane potential overcomes a threshold, the neuron produces an action potential (i.e., a rapid increase and decrease of membrane voltage V_{mem}) and propagates this "spike" along the axon. The same mechanism is reproduced in the silicon neuron. Here, when a spike is fired by the pre-synaptic neuron, a current proportional to the synaptic weight flows through the synapse and is sensed by the post-

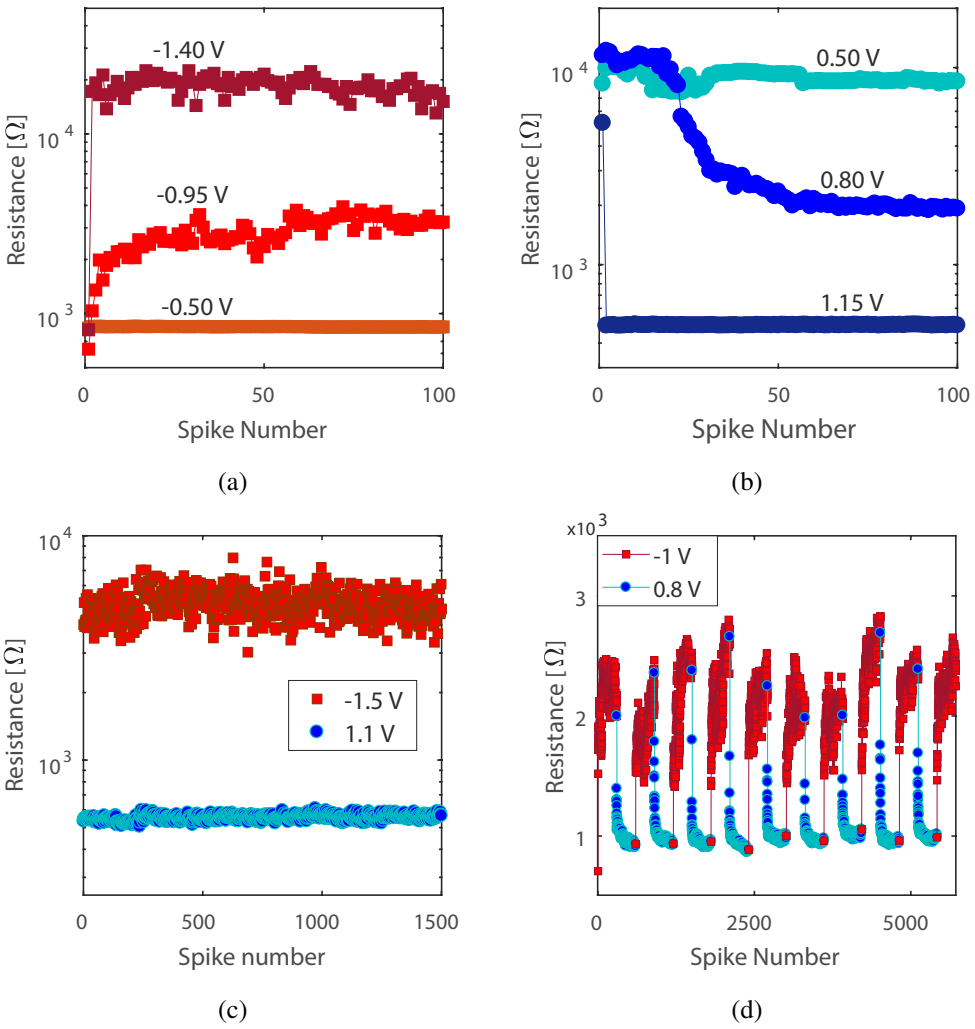


Figure 1: (a) Depression operation with trains of 100 spikes with time width of 100 μs and amplitudes (lower to higher) of -0.50 V, -0.95 V, and -1.40 V. (b) Potentiation operation with trains of 100 spikes of 100 μs and amplitudes (lower to higher) 0.50 V, 0.90 V, and 1.15 V.(c) Cycles of digital depression (red squares) and digital potentiation (blue circles) carried out using the automatic test equipment. Depression: voltage amplitude -1.5 V, time width 100 μs. Potentiation: voltage amplitude 1.1 V, time width 50 μs. (d) Cycles of analogue depression (red squares) and potentiation (blue circles) carried out using the automatic test equipment. Depression: 300 pulses, voltage amplitude -1 V, time width 100 μs. Potentiation: 300 pulses, voltage amplitude 0.8 V, time width 100 μs. Prior to operation of the neuromorphic chip, it is possible to set the voltage levels of the applied waveforms according to experimentally derived values, to achieve both depression and potentiation, with similar dynamic ranges.

synaptic neuron. The sensed current is low-pass filtered, thus modeling the EPSC, whose integral is the membrane potential V_{mem}. When an action potential is generated, the post-synaptic neuron fires a bi-phasic spike, i.e., a spike with a higher voltage phase followed by a lower voltage phase, at its output terminal.

Regarding plasticity mechanism, the change of the synaptic strength in biology is typically proportional to a variation in the number of neurotransmitter receptors in the post-synaptic neuron. The different spike rate and timing between pre- and post-synaptic

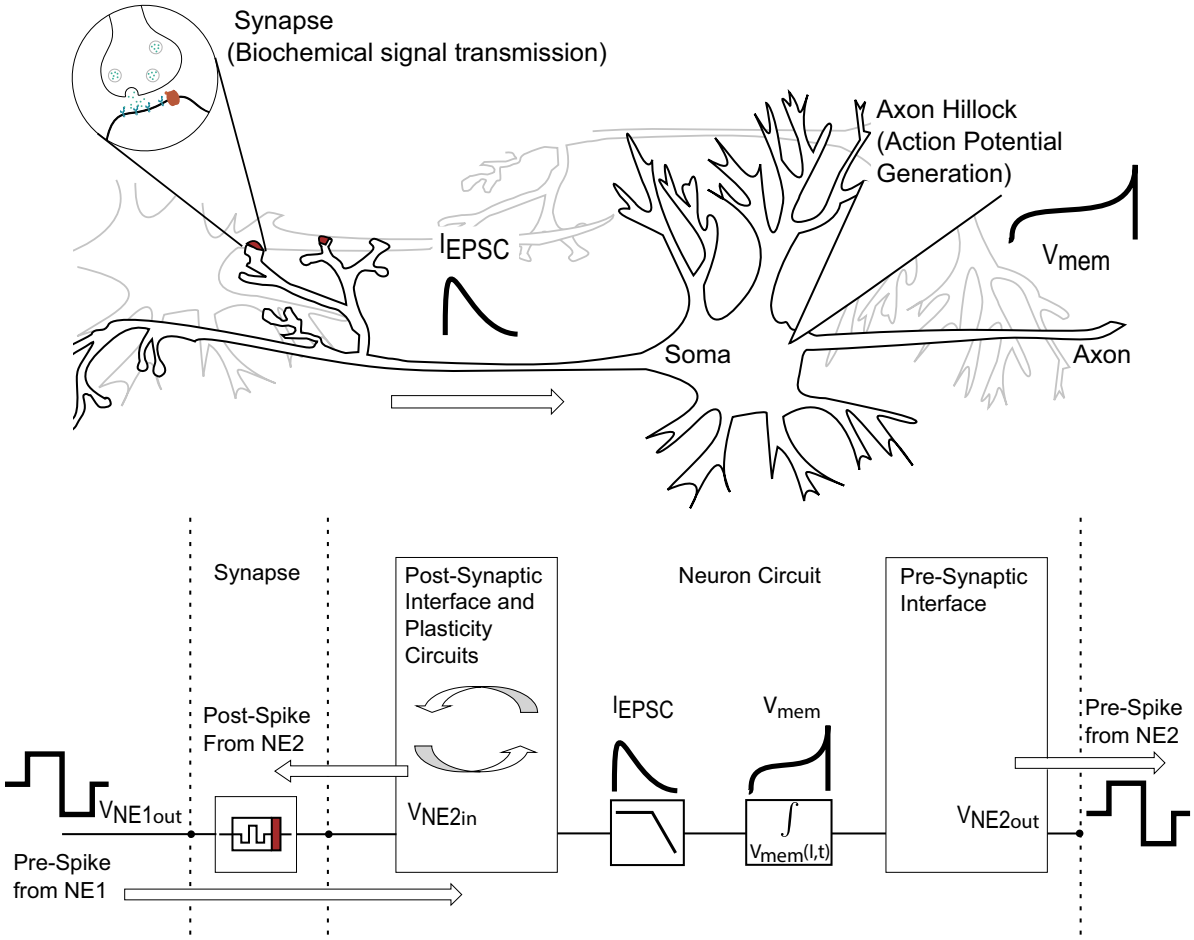


Figure 2: Comparison between biological neurons (top) and implemented silicon neurons (bottom).

neurons cause an alteration of the calcium intra-cellular concentration, which is believed to be responsible for the increase / decrease of the number of neurotransmitter receptors (and, consequently, of the potentiation / depression of the synapse). To implement the same phenomenon in hardware, a dedicated circuit integrates the neuron's spiking activity to produce a Calcium-concentration variable which keeps track of the recent post-neural activity. To enable synaptic weight updates the system produces overlapping spikes, one from the pre-synaptic neuron and one from the post-synaptic neuron, on the memristive device. Specifically, the post-synaptic neuron back-propagates a spike towards the memristive synapse at the onset of the bi-phasic pulse generated by the pre-synaptic neuron. Two interface circuits, namely the pre- and post-interface, enable the neuron to generate spikes able to drive the memristive synapse at both its terminals (i.e., to properly shape the pre- and the post-spike). The nature of the plasticity events (potentiating or depressing the synapse) is determined by the additional spike-driven learning circuits that sense the post-synaptic membrane potential and calcium concentration variable (see also [11] for details).Further details on the neuron model can be found in the Supplementary Information, Section S3.

3.3. Spike-driven threshold-based plasticity rule

The synaptic plasticity event is triggered every time a pre-synaptic spike is produced. The magnitude and direction of the weight change depend only on the state of the post-synaptic neuron at the time of the pre-synaptic spike arrival. The state variables playing a major role in the plasticity rule are post-synaptic neuron membrane potential V_{mem} and calcium variable $V_{Ca}(t)$. Figure 3a shows a measurement of the membrane potential of one of the silicon neurons on the chip, while Figure 3b shows a measurement of its calcium variable V_{Ca}. In Figure 3b, each spiking event from the neuron is indicated with an arrow. V_{Ca} naturally decays over time with slow time constant, unless a spiking event occurs to determine an increase of the calcium variable.

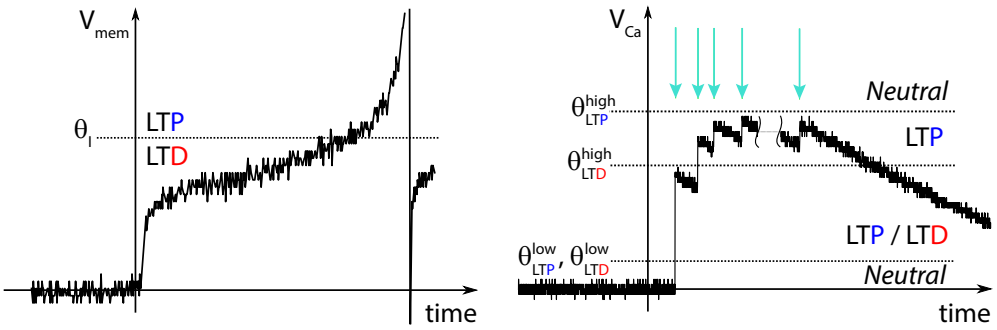


Figure 3: Example of (a) membrane potential and (b) Calcium variable. The thresholds used to determine the neuron state (LTP, i.e. potentiation, LTD, i.e. depression, or none, i.e. neutral) are indicated with the Greek letter θ. The arrows in (b) indicate the occurrence of a pre-synaptic spike.

Depending on the values of these two internal variables, the neuron is in one of three plasticity *states*: potentiation (LTP), depression (LTD), or neutral.More specifically, upon the arrival of a pre-spike, if the neuron is in its potentiation (depression) state, it will carry out a potentiating (depression) *operation*, i.e. it will fire a potentiation (depressing) spike, thus inducing a weight update in the synapse. If the neuron is in neutral state, instead, it will fire a low voltage spike (i.e. a neutral spike), thus inducing no weight update. The change in weight Δw is:

$$\Delta \mathrm{w} = \begin{cases} a & \text{if } V_{mem}(t|_{t_{pre}}) > \theta_I \quad \text{and} \quad \theta_{LTP}^{low} < V_{Ca}(t|_{t_{pre}}) < \theta_{LTP}^{high} \\ -b & \text{if } V_{mem}(t|_{t_{pre}}) \leq \theta_I \quad \text{and} \quad \theta_{LTD}^{low} < V_{Ca}(t|_{t_{pre}}) < \theta_{LTD}^{high} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where a and b are parameters which may in general depend on the programming requirements of the synaptic device; θ_I is the threshold on V_{mem}; $\theta_{LTP/LTD}^{high/low}$ define ranges, eventually overlapping, for the Calcium variable, to enable potentiation, depression or neutral plasticity operations.

The thresholds are set externally, therefore they can be adjusted so as to force the neuron to be in a determined state in certain circumstances, e.g. during system calibration.

3.4. Mixed memristive-neuromorphic electronic system

Figure 4a shows the system used to carry out plasticity experiments, which consists of a memristive device placed between two silicon neurons, namely a pre-synaptic (NE1) and a post-synaptic neuron (NE2). When NE1 is stimulated, it fires a bi-phasic spike at its output terminal. The spikes from NE1 are sensed by NE2, which fires back a potentiating, a depressing or a neutral spike at its input terminal, depending on its state variables. Figures 4b and 4c show (top to bottom) some representative experimental curves of the membrane potential of NE1 (V_{mem}), of the pre-spike from NE1 ($NE1_{out}$), of the post-spike from NE2 ($NE2_{in}$) during a depression (Figure 4b) and a potentiation (Figure 4c), and of the voltage ΔV across the synaptic device, where ΔV is defined as the difference between the voltage at $NE2_{in}$ minus the one at $NE1_{out}$. When V_{mem} overcomes a threshold, NE1 fires a bi-phasic pre-spike at its output terminal. The bi-phasic spike triggers a response at $NE2_{in}$ according to (1), i.e. a neutral spike (low voltage spike which does not affect the device state, not shown), a depressing spike (Figure 4b), or a potentiating spike (Figure 4c). Measurements of membrane potential and Calcium concentration in NE2 can be seen in [43], Figure 8. The pre- and post-spikes overlap and the memristive device changes its resistance once the setting of the neuron spike generators are set to provide the correct voltage drop on the device. As evident in Figures 4b and 4c, the pulse width of the post-spike is half the pulse width of the pre-spike. Therefore, the post-spike overlaps only with one phase of the bi-phasic pre-spike,and this phase is named *writing phase*. In the other phase, the synaptic weight is read, hence the name of *read phase*. The memristor conductance is read only in the read phase by measuring the current flowing through it. The read phase follows the write phase when NE2 is in potentiation or neutral state, whereas it precedes the writing phase when NE2 is in depression state.

3.5. Synaptic weight evaluation

During the read phase, NE2 integrates the current flowing through the memristive synapse, which is proportional to the memristor conductance, and generates an excitatory post-synaptic current, which is then used to increase its membrane potential. When the membrane potential overcomes its firing threshold, an action potential is generated, which results in NE2 firing a bi-phasic pulse at its output. The firing rate at $NE2_{out}$ is proportional to the current injected in NE2 and, consequently, to the synaptic device resistance. We therefore can use the neural activity at $NE2_{out}$, monitored under specific and controlled system conditions, as a way to measure the synaptic weight without using the ATE. To this end, we explore the response at $NE2_{out}$ under different synaptic weight conditions and relate it with the resistive state of the device in order to be able to measure the device resistance through the monitoring of the firing rate at $NE2_{out}$. Firstly, we use the ATE to set the synaptic device in a known resistive state. Then, we connect the device to the neuromorphic chip via the switch PCB. Post-synaptic neural activity on the $NE2_{out}$ terminal is measured as the pre-synaptic neuron NE1 is forced to fire for 1 s length at an inter-stimulus-interval of 2.5 ms. Meanwhile, we force the post-synaptic neuron to stay in neutral mode in order not to affect the device resistance while measuring. The procedure to estimate the synaptic weight is repeated for a series of known

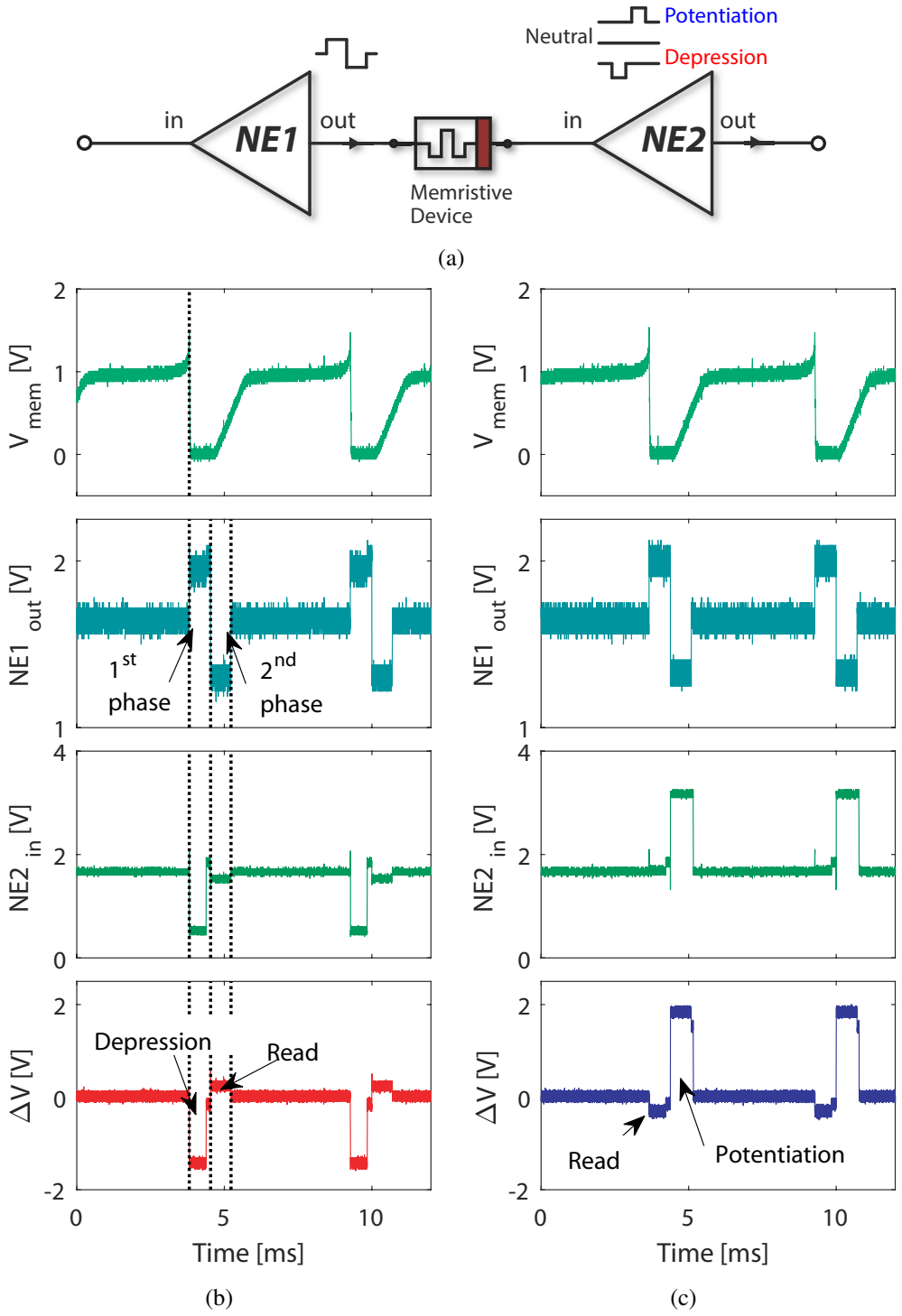


Figure 4: (a) Illustration of the experimental configuration used in this work: A memristive device connects the pre-synaptic terminal NE1$_{out}$ of neuron NE1 to the post-synaptic terminal NE2$_{in}$ neuron NE2. The neuron symbols are the ones used on neuroscience to represent biological pyramidal cells, with input on the left of the triangle and output on the right. External stimuli are used to elicit spikes on neuron NE1, the propagated spikes cause neuron NE2 to fire in response, observable on terminal NE2$_{out}$. (b)-(c) Membrane potential of NE1 (first panel) and bi-phasic spikes fired by NE1 (second panel) and NE2 (third panel) during a depression and a potentiation, respectively. Fourth panel: Voltage drop across the device. Vertical dashed lines evidence the simultaneity of V_{mem} overtaking the firing threshold (top panels in (b-c)), the bi-phasic spike generation by NE1 (second panels in (b-c)) and the response of NE2 (third panels in (b-c)). All the curves are acquired via oscilloscope.

resistive states (acquired through the ATE).

Figure 5a shows an example of neuronal activity during the stimulation protocol when the memristive synapse is in its potentiated or depressed state. The pre-synaptic neuron NE1 (top panel) is stimulated at a constant rate. In turn, the post-synaptic neuron NE2 both responds at $NE2_{in}$with a neutral pulse, thus generating on the device the voltage difference ΔV shown in the second panel of Figure 5a, and fires bi-phasic spikes at $NE2_{out}$ (third and fourth panels) with a rate proportional to the synaptic weight. When the synapse is in its potentiated (depressed) state, as in the third (fourth) panel of Figure 5a, the connection between the neurons is facilitated (inhibited) and the frequency of the activity at $NE2_{out}$ is about 100 Hz (50 Hz). The potentials applied during the read phases of the neutral pulses lie below400 mV, as shown in the second panel of Figure 5a, which was set as the safe operation range in which parts of the programming spikes, that are not involved in the actual depression or potentiation operation (see Figure 1), should lie.

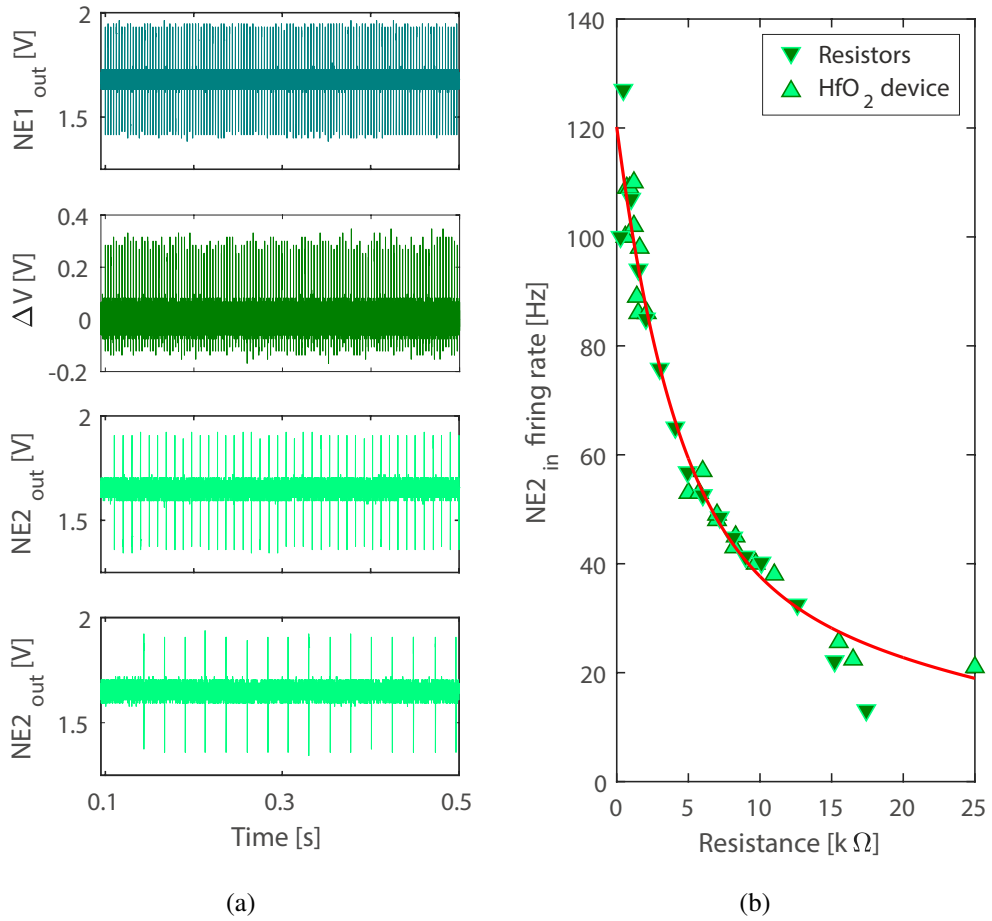
In order to show that the neutral pulses do not alter the resistive state of the memristive device, a control with a series of resistors, replacing the memristive device in the experiment, was performed. Figure 5b displays the firing rates obtained with control resistors and with the memristive device. Moreover, an exponential-like calibration curve is drawn to fit the non-linear dependence between device resistance and neuronal firing activity. The close overlap of the fit to the acquired data suggests that indeed the measurement protocol does not affect the memristive device. Furthermore, the obtained fitting equation is used in the experiments to directly acquire a resistance estimate through the observation of post-synaptic firing.

3.6. Synaptic weight update

After establishing the relationship between the synaptic weight and the firing rate at $NE2_{out}$ in a quantitative manner, we can trigger subsequent potentiating/depressing events in the memristive device with the neuromorphic chip to demonstrate the plasticity rule.

Potentiation (depression) occurs when the overlapping of pre- and post-spikes results in a voltage drop across the device high enough to allow for a change in the device resistance.

Figure 6 shows the plasticity experiments resulting from the stimulation of the memristive synapse by the CMOS neurons using the spike-driven threshold-based plasticity rule. Figure 6a shows few cycles of potentiation/depression where the device is operated in digital fashion. After each plasticity event, the memristive device state is read by setting the system in neutral mode and measuring the firing rate at $NE2_{out}$ terminal. The left y axis of Figure 6a indicates the resistance of the device extracted from the characterization curve (see Figure 5b), whereas the right y axis indicates the measured firing rate of $NE2_{out}$. After a potentiating event, the device is in its LRS (corresponding to a firing rate of about 110 Hz), whereas after depression it is in its HRS (corresponding to a firing rate of about 50 Hz). The results are in line with previous characterizations of the device (Figures 1c and 5b).In the following, the amplitude of both the bi-phasic spikes and of the potentiating and depressing spikes were selected according to the results in Figure 1 in order to obtain resistance ranges compatible in both the plasticity operations.

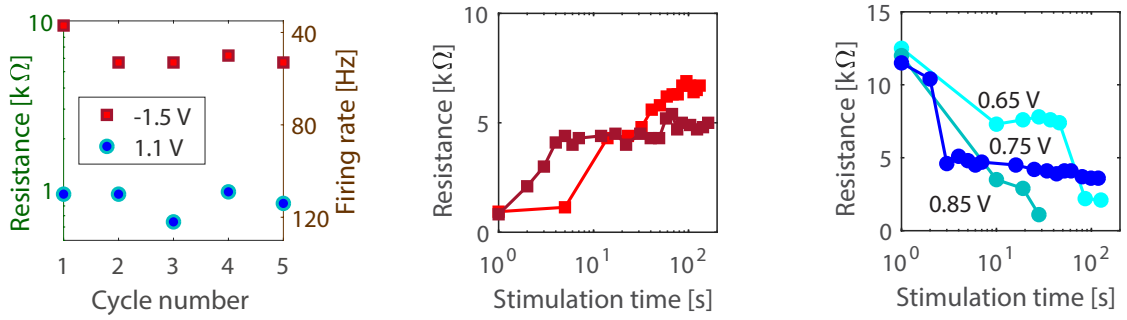


(a) (b)

Figure 5: (a) Example of firing rate of NE2 in neutral state. First panel: NE1$_{out}$ firing rate; second panel: ΔV on the memristive device; third and fourth panels: NE2$_{out}$ firing rate when the synapse is potentiated, i.e. high current integrated by NE2, and depressed, i.e. low current integrated by NE2, respectively. (b) Firing frequency of the post-synaptic neuron, in response to stimulation of the pre-synaptic neuron, as a function of resistivity of the synaptic connection between both. Red line: Exponential curve, fitted to the data acquired with the memristive device.

Figure 6b and 6c demonstrate analogue depression and potentiation, respectively. Figure 6b shows two depression operation carried out with spikes of 100 μs and -1 V. The experiments were carried out stimulating NE1 for several seconds and then reading the resistance of the synapse with the ATE. The device gradually switches from its LRS to its HRS and several intermediate states are detectable. The comparison between the two curves in the Figure evidences some variability in the weight update steps. This behaviour is nonetheless in line with previous characterisations shown in Figure 1d, where the HRS is shown to suffer from higher variability than LRS, and is in agreement with the inherent variability of filamentary memristive devices [5, 9, 12, 38, 45, 46]. However, it has been demonstrated [47] that variability is not expected to be an issue in larger systems because of the intrinsic parallel and fault tolerant nature of spike-based neuromorphic networks. Figure 6c demonstrate analogue plasticity with spikes of 100 μs and different amplitudes. Here as well, the resistance change is gradual and the voltage amplitude of the spike affects

the speed of the resistance decrease. Indeed, after 10 s stimulation, we observe a higher resistance change when the device is stimulated with higher voltage amplitudes. Moreover, it can be noted that in both Figures 6b and 6c, the evolution of the resistance is faster in the first seconds of the stimulation, in accordance with Figures 1a and 1b.



(a) (b) (c)

Figure 6: (a) Cycles of digital depression (red squares) and digital potentiation (blue circles). Depression: voltage amplitude -1.5 V, time width 87 μs. Potentiation: voltage amplitude 1.1 V, time width 46 μs. (b) Analogue depression: voltage amplitude -1 V, time width 100 μs. (c) Analogue potentiation: voltage amplitude 0.65 V, 0.85 V, and 0.75 V, time width 100 μs.

4. Discussion

In this work we provide a hardware demonstration of the spike-driven threshold-based learning rule proposed in [34]. The learning rule is hebbian in that the plasticity event occurs depending on the firing activity of the pre-synaptic neuron. The direction of the weight update (increasing or decreasing) depends on the post-synaptic neuron state, which is defined on the basis of the value of two state variables, the membrane potential and the calcium intra-cellular concentration. The latter includes the information on the neuron's recent firing activity. The ruleimplementation is asynchronous and behaves as the STDP, but the timing dependence between the pre- and post-synaptic neuron's firing activities is not explicit because there is no directcoding of temporal relationship between the two spikes. The implicit relationship with time allows the system to fire a plasticity event even in presence of a single pre-spike. Unlike the conventional pair-wise STDP [23, 24, 48, 49], the spike-driven threshold-based plasticity rule exploits the calcium variable to regulate the direction of the synaptic weight. The calcium variabletracks the recent neuronal activity at the post-synaptic neurons and has, therefore, a fundamental role in the evolution of learning dynamics. Moreover, since the decay time constant of this variable is slow,the present implementation can support the sensitivity to temporal correlation in the real-time (ms) domain through μs-long pulses with the synchronisation provided by the digital neuron/synapse interface as described with reference to Figures 4b-4c. Conversely, many literature proposal for STDP implementation involve overlapping pulses whose lengths provide the time window of the sensitivity to temporal correlation between incoming events. [23, 24] Despite many literature

reports in this direction aim only at providing a proof of principle, the use of long overlapping pulses poses fundamental practical problems, e.g. in case of (the likely) occurrence of two simultaneous plasticity events on the same neuron. Indeed, since a long spike pulse would activate an entire row / column of synapses, the shared spike address line cannot be re-used for new plasticity events until the previous spike/plasticity operation has concluded.

The neuron/synapse interface, of which we demonstrate the functionality is this paper, provides a solution for an additional aspect often neglected in memristor-based STDP implementation, e.g. the accurate reading of the memristive conductance. Indeed, one neuron must integrate current contributions that are proportional to the conductances of each afferent synapse. This fact requires that the current flowing through the memristive device must be driven by always the same voltage drop on it and for the same time and, therefore, it is necessary to define a precise read voltage an read time, which must not overlap with the programming operation of the device. This functionality is well demonstrated by our system as discussed with reference to Figures 4b-4c. The achievement of sensitivity to real-time temporal correlations through short pulses is made possible at the expenses of neuron complication, i.e. by the use of additional electronics for spike integration into the Calcium variable (see Figure 3b). However, this is realised with a compact log-domain integrator employing CMOS transistors in the sub-threshold regimes, [1] which allows the use of relatively small VLSI-compatible capacitors [50].

To summarize, the present work demonstrates, in a simplified case two neurons-one memristor, the possibility to use HfO_x based memristors implementing a spike-driven threshold based learning rule which is compatible with asynchronous CMOS circuits able to work also in a biological compatible time scale, and pushing for future applications directed to efficient circuits for online learning. The scaling up of the present system to a real application-oriented product requires the arrangement of synapses in an array. While passive and selector free cross bar have been proposed and demonstrated for some applications [51, 52], the co-integration of selectors and memristive devices might be required for very high density systems. In terms of selectors, despite two-terminal bipolar selector devices constitute the optimum solution in term of area occupancy, the level maturity of their technology is still not fully compatible with industrial exploitation and co-integration with CMOS circuit and memristor. In this respect, it is worth noticing that the neuron/synapse interface of the CMOS chip used in this work is already equipped with digital asynchronous control signals [11] which can be possibly exploited, through a suitable re-design of the chip, to drive also select transistors in active cross-bars, which can be foreseen as a short/mid-term solution for spiking neural networks.

To conclude, the present work, despite its small scale, provides the solution for a certain number of issues raising from the attempts of interfacing memristive devices with an event-based neuromorphic processor. The addressed technical issues take on a trailblazing role in view of the VLSI of memristive devices with CMOS neuromorphic processors.

5. Conclusion

In this work we presented the hardware demonstration of a spike-driven threshold-based learning rule in a system centered around a neuromorphic processor that uses the firing activity of analogue neuron circuits to trigger plasticity events in memristive synapses interfaced to silicon neurons. We demonstrate that the rule allows asynchronous plasticity events and weight dependent post-synaptic activity, thus improving the conventional STDP learning rule and paving the road towards the development of integrated mixed neuromorphic / memristive systems for real-time applications.

Acknowledgments

This work is partly supported by the project 'NeuRAM3' (EU Horizon2020, grant n. 687299), by the project 'RAMP' (EU FP7, grant no. 612058), and by the EPSRC project 'CHIST-ERA' ERA-Net, EPSRC EP/J00801X/1, EP/K017829/1. E.C. was partially funded by COST Action IC1401 - MemoCiS (Memristors - Devices, Models, Circuits, Systems and Applications) in the framework of the Short Term Scientific Mission (STSM) program.

References

[1] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri. Neuromorphic Electronic Circuits for Building Autonomous Cognitive Systems. *Proceedings of the IEEE*, 102(9):1367–1388, Sept. 2014.

[2] Steve Furber. Large-scale neuromorphic computing systems. *Journal of Neural Engineering*, 13(5):051001, 2016.

[3] C. Merkel, R. Hasan, N. Soures, D. Kudithipudi, T. Taha, S. Agarwal, and M. Marinella. Neuromemristive Systems: Boosting Efficiency through Brain-Inspired Computing. *Computer*, 49(10):56–64, Oct 2016.

[4] Hesham Mostafa, Ali Khiat, Alexander Serb, Christian G. Mayr, Giacomo Indiveri, and Themis Prodromakis. Implementation of a spike-based perceptron learning rule using TiO2x memristors. *Frontiers in Neuroscience*, 9(357), 2015.

[5] Thilo Werner, Elisa Vianello, Olivier Bichler, Daniele Garbin, Daniel Cattaert, Blaise Yvert, Barbara De Salvo, and Luca Perniola. Spiking Neural Networks Based on OxRAM Synapses for Real-Time Unsupervised Spike Sorting. *Frontiers in Neuroscience*, 10:474, 2016.

[6] International Technology Roadmap for Semiconductors. `http://www.itrs2.net/`.

[7] S. Ambrogio, S. Balatti, V. Milo, R. Carboni, Z. Q. Wang, A. Calderoni, N. Ramaswamy, and D. Ielmini. Neuromorphic Learning and Recognition With One-Transistor-One-Resistor Synapses and Bistable Metal Oxide RRAM. *IEEE Transactions on Electron Devices*, 63(4):1508–1515, April 2016.

[8] Doo Seok Jeong, Kyung Min Kim, Sungho Kim, Byung Joon Choi, and Cheol Seong Hwang. Memristors for Energy-Efficient New Computing Paradigms. *Advanced Electronic Materials*, 2(9):1600090–n/a, 2016. 1600090.

[9] S. Yu. Neuro-inspired computing with emerging nonvolatile memory. *Proceedings of the IEEE*, PP(99):1–26, 2018.

[10] M. Payvand, L. K. Muller, and G. Indiveri. Event-based circuits for controlling stochastic learning with memristive devices in neuromorphic architectures. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, May 2018.

[11] H. Mostafa, C. Mayr, and G. Indiveri. Beyond spike-timing dependent plasticity in memristor crossbar arrays. In *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 926–929, May 2016.

[12] Duygu Kuzum, Shimeng Yu, and H-S Philip Wong. Synaptic electronics: materials, devices and applications. *Nanotechnology*, 24(38):382001, 2013.

[13] G. Indiveri and S. C. Liu. Memory and Information Processing in Neuromorphic Systems. *Proceedings of the IEEE*, 103(8):1379–1397, Aug 2015.

[14] H. Wu, X. H. Wang, B. Gao, N. Deng, Z. Lu, B. Haukness, G. Bronner, and H. Qian. Resistive random access memory for future information processing system. *Proceedings of the IEEE*, 105(9):1770–1789, Sept 2017.

[15] Peng Yao, Huaqiang Wu, Bin Gao, Sukru Burc Eryilmaz, Xueyao Huang, Wenqiang Zhang, Qingtian Zhang, Ning Deng, Luping Shi, H.-S. Philip Wong, and He Qian. Face classification using electronic synapses. *Nature Communications*, 8:15199, 2017.

[16] Jacopo Frascaroli, Stefano Brivio, Erika Covi, and Sabina Spiga. Evidence of soft bound behaviour in analogue memristive devices for neuromorphic computing. *Scientific Reports*, 8(1):7178, 2018.

[17] Jean-Pascal Pfister and Wulfram Gerstner. Beyond pair-based stdp: A phenomenological rule for spike triplet and frequency effects. In *Advances in neural information processing systems*, pages 1081–1088, 2006.

[18] Julijana Gjorgjieva, Claudia Clopath, Juliette Audet, and Jean-Pascal Pfister. A triplet spike-timingdependent plasticity model generalizes the bienenstockcoopermunro rule to higher-order spatiotemporal correlations. *Proceedings of the National Academy of Sciences*, 108(48):19383–19388, 2011.

[19] F. L. Maldonado Huayaney, S. Nease, and E. Chicca. Learning in silicon beyond stdp: A neuromorphic implementation of multi-factor synaptic plasticity with calcium-based dynamics. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 63(12):2189–2199, Dec 2016.

[20] Manu V Nair, Lorenz K Muller, and Giacomo Indiveri. A differential memristive synapse circuit for on-line learning in neuromorphic computing systems. *Nano Futures*, 1(3):035003, 2017.

[21] Vishal Saxena, Xinyu Wu, Ira Srivastava, and Kehan Zhu. Towards spiking neuromorphic system-on-a-chip with bio-plausible synapses using emerging devices. In *Proceedings of the 4th ACM International Conference on Nanoscale Computing and Communication*, NanoCom '17, pages 18:1–18:6, New York, NY, USA, 2017. ACM.

[22] Guo-Qiang Bi and Mu-Ming Poo. Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type. *The Journal of Neuroscience*, 18(24):10464–10472, 1998.

[23] E. Covi, S. Brivio, A. Serb, T. Prodromakis, M. Fanciulli, and S. Spiga. Analog Memristive Synapse in Spiking Networks Implementing Unsupervised Learning. *Frontiers in Neuroscience*, 10:482, October 2016.

[24] Stefano Ambrogio, Nicola Ciocchini, Mario Laudato, Valerio Milo, Agostino Pirovano, Paolo Fantini, and Daniele Ielmini. Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses. *Frontiers in Neuroscience*, 10(56), 2016.

[25] Walter Senn and Stefano Fusi. Learning Only When Necessary: Better Memories of Correlated Patterns in Networks with Bounded Synapses. *Neural Computation*, 17(10):2106–2138, 2005.

[26] P. Jesper Sjöström, Ede A. Rancz, Arnd Roth, and Michael Husser. Dendritic excitability and synaptic plasticity. *Physiological Reviews*, 88(2):769–840, 2008. PMID: 18391179.

[27] John Lisman and Nelson Spruston. Questions about stdp as a general model of synaptic plasticity. *Frontiers in Synaptic Neuroscience*, 2:140, 2010.

[28] B. Rajendran and F. Alibart. Neuromorphic Computing Based on Emerging Memory Technologies. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 6(2):198–211, June 2016.

[29] V. Milo, G. Pedretti, R. Carboni, A. Calderoni, N. Ramaswamy, S. Ambrogio, and D. Ielmini. Demonstration of hybrid CMOS/RRAM neural networks with spike time/rate-dependent plasticity. In *2016 IEEE International Electron Devices Meeting (IEDM)*, pages 16.8.1–16.8.4, Dec 2016.

[30] G. Pedretti, V. Milo, S. Ambrogio, R. Carboni, S. Bianchi, A. Calderoni, N. Ramaswamy, A. S. Spinelli, and D. Ielmini. Memristive neural network for on-line learning and tracking with brain-inspired spike timing dependent plasticity. *Scientific Reports*, 7(1):5288, 2017.

[31] Per Jesper Sjöström, Gina G Turrigiano, and Sacha B Nelson. Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, 32(6):1149 – 1164, 2001.

[32] Wei He, Kejie Huang, Ning Ning, Kiruthika Ramanathan, Guoqi Li, Yu Jiang, JiaYin Sze, Luping Shi, Rong Zhao, and Jing Pei. Enabling an Integrated Rate-temporal Learning Scheme on Memristor. *Scientific Reports*, 4(4755):1–6, 2014.

[33] Yi Li, Yingpeng Zhong, Jinjian Zhang, Lei Xu, Qing Wang, Huajun Sun, Hao Tong, Xiaoming Cheng, and Xiangshui Miao. Activity-Dependent Synaptic Plasticity of a Chalcogenide Electronic Synapse for Neuromorphic Systems. *Scientific Reports*, 4(4906):1–7, 2014.

[34] M. Brader, W. Senn, and S. Fusi. Learning Real-World Stimuli in a Neural Network with Spike-Driven Synaptic Dynamics. *Neural Computation*, 19(11):2881–2912, 2007.

[35] Michael Graupner and Nicolas Brunel. Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location. *Proceedings of the National Academy of Sciences*, 109(10):3991–3996, 2012.

[36] Robert Urbanczik and Walter Senn. Learning by the dendritic prediction of somatic spiking. *Neuron*, 81(3):521 – 528, 2014.

[37] S. Brivio, E. Covi, A. Serb, T. Prodromakis, M. Fanciulli, and S. Spiga. Experimental study of gradual/abrupt dynamics of HfO_2-based memristive devices. *Applied Physics Letters*, 109:133504, 2016.

[38] Jacopo Frascaroli, Flavio Giovanni Volpe, Stefano Brivio, and Sabina Spiga. Effect of Al doping on the retention behavior of HfO_2 resistive switching memories. *Microelectronic Engineering*, 147:104–107, 2015.

[39] S. Brivio, G. Tallarida, E. Cianci, and S. Spiga. Formation and disruption of conductive filaments in a HfO_2/TiN structure. *Nanotechnology*, 25(38):385705, 2014.

[40] S. Brivio, J. Frascaroli, and S. Spiga. Role of metal-oxide interfaces in the multiple resistance switching regimes of Pt/HfO_2/TiN devices. *Applied Physics Letters*, 107(2):023504, 2015.

[41] E. Covi, S. Brivio, M. Fanciulli, and S. Spiga. Synaptic potentiation and depression in Al:HfO_2-based memristor. *Microelectron. Eng.*, 147:41–44, 2015.

[42] E. Covi, S. Brivio, J. Frascaroli, M. Fanciulli, and S. Spiga. (Invited) Analog HfO_2-RRAM Switches for Neural Networks. *ECS Transactions*, 75(32):85–94, January 2017.

[43] Ning Qiao, Hesham Mostafa, Federico Corradi, Marc Osswald, Fabio Stefanini, Dora Sumislawska, and Giacomo Indiveri. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128k synapses. *Frontiers in Neuroscience*, 9(141), 2015.

[44] Romain Brette and Wulfram Gerstner. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of neurophysiology*, 94(5):3637–3642, 2005.

[45] Jacopo Frascaroli, Stefano Brivio, Federico Ferrarese Lupi, Gabriele Seguini, Luca Boarino, Michele Perego, and Sabina Spiga. Resistive Switching in High-Density Nanodevices Fabricated by Block Copolymer Self-Assembly. *ACS Nano*, 9(3):2518–2529, 2015.

[46] D. Garbin, E. Vianello, O. Bichler, Q. Rafhay, C. Gamrat, G. Ghibaudo, B. DeSalvo, and L. Perniola. HfO_2-Based OxRAM Devices as Synapses for Convolutional Neural Networks. *IEEE Transactions on Electron Devices*, 62(8):2494–2501, 2015.

[47] S. Yu, P. Y. Chen, Y. Cao, L. Xia, Y. Wang, and H. Wu. Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect. In *2015 IEEE International Electron Devices Meeting (IEDM)*, pages 17.3.1–17.3.4, Washington, DC, Dec 2015.

[48] Johannes Bill and Robert Legenstein. A compound memristive synapse model for statistical learning through STDP in spiking neural networks. *Frontiers in Neuroscience*, 8(412), 2014.

[49] Nan Du, Mahdi Kiani, Christian G Mayr, Tiangui You, Danilo Brger, Ilona Skorupa, Oliver G. Schmidt, and Heidemarie Schmidt. Single pairing spike-timing dependent plasticity in $BiFeO_3$ memristors with a time window of 25ms to 125μs. *Frontiers in Neuroscience*, 9(227), 2015.

[50] Giacomo Indiveri, Bernab Linares-Barranco, Robert Legenstein, George Deligeorgis, and Themistoklis Prodromakis. Integration of nanoscale memristor synapses in neuromorphic computing architectures. *Nanotechnology*, 24(38):384010, 2013.

[51] Sangsu Park, Myonglae Chu, Jongin Kim, Jinwoo Noh, Moongu Jeon, Byoung Hun Lee, Hyunsang Hwang, Boreom Lee, and Byung-geun Lee. Electronic system with memristive synapses for pattern recognition. *Scientific Reports*, 5(10123):1–9, 2015.

[52] F. Merrikh Bayat, M. Prezioso, B. Chakrabarti, H. Nili, I. Kataeva, and D. Strukov. Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits. *Nature Communications*, 9(1):2331, Jun 2018.