# CMOS and Memristive Hardware for Neuromorphic Computing

**Mostafa Rahimi Azghadi[1,*], Ying-Chen Chen[2], Jason K. Eshraghian[3], Jia Chen[4], Chih-Yang Lin[3], Amirali Amirsoleimani[5], Adnan Mehonic[6], Anthony J Kenyon[6], Burt Fowler[2], Jack C Lee[2], Yao-Feng Chang[2,*]**

[1]College of Science and Engineering, James Cook University, Townsville, QLD 4811, Australia

[2]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712, United States

[3]Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48109-2122, United States

[4]Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China

[5]Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada

[6]Department of Electronic and Electrical Engineering, University College London, Torrington Place, London, United Kingdom

[*]mostafa.rahimiazghadi@jcu.edu.au, yfchang@utexas.edu

Author Manuscript

**ABSTRACT**

The ever-increasing processing power demands of digital computers cannot continue to be fulfilled indefinitely unless there is a paradigm shift in computing. Neuromorphic computing, which takes inspiration from the highly parallel, low power, high speed, and noise-tolerant computing capabilities

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/aisy.201900189.

This article is protected by copyright. All rights reserved

of the brain, may provide such a shift. To that end, various aspects of the brain, from its basic building blocks, such as neurons and synapses, to its massively parallel in-memory computing networks have been being studied by the huge neuroscience community. Concurrently, many researchers from across academia and industry have been studying materials, devices, circuits, and systems, to implement some of the functions of networks of neurons and synapses to develop bio-inspired (neuromorphic) computing platforms. These are being designed using various hardware technologies, including the well-established Complementary Metal-Oxide Semiconductor (CMOS), and emerging memristive technologies such as SiOx-based memristors. In this work, we review recent progress in CMOS, SiOx-based memristive, and mixed CMOS-memristive hardware for neuromorphic systems. We provide new and published results from various devices that we have developed to replicate selected functions of neurons, synapses, and simple spiking networks. In particular, we present results from CMOS neural and synaptic circuits that mimic some functions of their biological counterparts. We also show experimental results from SiOx-based memristive synaptic and neuron-mimicking devices and present simulation results from memristive neurons replicating known functions of the Hodgkin-Huxley (HH) and Morris-Lecar (ML) neuron models. Furthermore, we present results showing the integration of CMOS and memristor technology in two neuromorphic platforms. The first hybrid system shows spike-based image sensing, while the second system simulates rate-based synaptic plasticity using a CMOS-memristor synapse. Next, we show that the proposed CMOS and memristive devices can be assembled in different neuromorphic learning platforms to perform simple cognitive tasks such as classification of spike rate-based patterns or MNIST handwritten digits. We hope this paper will be useful to the neuromorphic and unconventional computing research communities by providing insights into recent advances in both established and emerging hardware technologies for neuromorphic computing.

# 1 Introduction

This article is protected by copyright. All rights reserved

Author Manuscript

Conventional computing that involves storing data in, and retrieving it from, memory results in numerous interactions between the slow memory and the fast processors. Besides, the data movement between memory and processor produces a bottleneck that is limited by the available bandwidth in the computing platform used[1–3]. This significantly affects the computation speed, in particular when a large amount of data is to be processed in a short time. Furthermore, the excessive data movement in conventional sequential- and parallel-processing computers, results in very high power consumption. Even though supercomputers with highly parallel processing capabilities are presently used to satisfy the high throughput of computationally complex tasks, they are extremely power hungry[4].

The prevalence of big data and a need for low-power and high-speed processing in everyday life and edge computing devices demands a shift to computers with elevated capabilities but with low power consumption. This may be achieved through unconventional computing methods[5], among which neuromorphic computing is a promising approach, where engineering inspirations are taken from the noise-tolerant, parallel, low-power, and high-speed signal processing abilities of biological brains[6].

Neuromorphic computing was coined by Carver Mead in the early 90s[7], when he envisioned that exploiting the similarities between semiconductor physics and biological neural systems, one may develop brain-inspired computing platforms. Ever since, neuromorphic research has evolved and researchers are implementing various technologies, from conventional semiconductors as proposed by Mead[8–12], to memristive systems[2, 13, 14], to hybrid CMOS-memristive designs[15, 16] to develop neuro-mimicking platforms for replicating experimental results observed in biology[14, 17] or for neuro-inspired platforms used in computing systems[18–20].

In this paper, as shown in Figure 1, we provide an overview of neuromorphic computing with CMOS, SiOx-based memristive, and mixed CMOS-memristive technologies. Our work covers neuro-mimicking designs, which are able to replicate some known aspects of biological neurons and synapses. We also use these neuro-mimicking components to show they can be used in simple cognitive tasks such as spike-based pattern classification, or image sensing. In addition, we use

This article is protected by copyright. All rights reserved

Author Manuscript

memristive devices to perform more complex classification tasks such as classifying MNIST hand-written digits using networks developed mainly from memristive weight elements, where these elements show good performance despite device variations and non-idealities.

Author Manuscript

## 2 Neuromorphic Components Design

In this section we discuss some previously published as well as some new results on CMOS and memristive neuromorphic neurons and synapses. These designs are toward one of the main goals of neuromorphic engineering, i.e. replicating the underlying principles of neural systems, with the hope to understand them better and to discover the use of technology for artificial neural components. Using these components, we will then, in the next section, design systems that perform neuromorphic learning and computation, which are essential purposes of neuromorphic systems.

### 2.1 CMOS Synapse and Neuron Design

Silicon technology that has reached maturity in the past 40 years has been widely used in neuromorphic computing, from the early works of Mead's team such as the Mahowald and Douglas silicon neuron[21], to cooperation between Mead's students and the wider neuromorphic community in constructing CMOS implementations of various neuron models[22]. In addition, synaptic plasticity, which is believed to play an essential role in learning and memory in the brain, has also been implemented in CMOS in various forms. These range from detailed biophysical models[23, 24] to computational rate- and timing-based synapse models[25–27].

When implementing CMOS-based neuronal and synaptic models, one could design a circuit that utilizes the above-threshold[23] or sub-threshold[28] region of operation of transistors to implement various mathematical expressions, which are mainly devised by computational neuroscientists[29]. One of the main advantages of analog CMOS designs for neuron and synapse circuits is that the designer

This article is protected by copyright. All rights reserved

can produce almost any behavioral characteristics of the neuron or synapse, which has been approximated and mapped through a mathematical model. In that case, the designer will route together transistors, and in some cases capacitors (which can be realized using transistors), to realize those rules.

In the following sections we discuss the design and implementation of some previous neuron and synapse circuits implemented using CMOS technology and discuss their behaviors, which closely mimic the observations in biological experiments.

### *2.1.1 CMOS Synapse Design*

Here, we describe the design of an analog synaptic circuit that closely mimics some of the known behavioral characteristics of synapses in the visual cortex and hippocampus[30, 31]. The circuit that is shown in Figure 2(a) has been designed and fabricated in CMOS to implement the triplet Spike Timing Dependent Plasticity (STDP) rule proposed in[32]. STDP is a recognized synaptic plasticity rule that alters the synaptic weight based on the timing differences between pre- and post-synaptic spikes. If a pre-synaptic spike arrives before a post-synaptic one, it can result in potentiation, while a reverse spiking order may cause depression. Triplet STDP, on the other hand, considers not only the timing differences between pre- and post-synaptic spikes, but it takes into account the timing difference between a pair of pre- or a pair of post-synaptic spikes in the presence of a post- or pre-synaptic spike, respectively[32].

Our newly proposed STDP circuit that resembles, but is simpler than a previous design presented in[33] is capable of reproducing the outcome of a wide range of synaptic plasticity biological experiments including spike pair, triplet, and quadruplet performed in the hippocampus as shown in[30]. In addition, it can mimic frequency-dependent pairing experiments performed in the visual cortex as shown in[31]. There are three different circuit parts depicted in Figure 2(a) that each

This article is protected by copyright. All rights reserved

Author Manuscript

corresponds to a different feature of a computational model which represents the triplet STDP rule proposed by[32]. Here, the red middle part implements the pair-based dynamics of STDP. With the addition of the two side parts, shown in blue, extra pre-pre or post-post interaction dynamics are added up to the base-line pair-based dynamics (shown in red). For more details about the circuit and its functionality, please refer to[33].

Figure 2(b-c) shows measurement results from the circuit in (a), which was fabricated using a 1-poly 6-metal 0.18 $\mu$ m AMS CMOS process. In this circuit, the synaptic weight is represented by the charge stored in the weight capacitor, $C_W$ . Here, the top purple trace in (b) shows the change in the synaptic weight due to a pre-post-pre triplet of spikes, when only the pair-based STDP (red part) of the circuit is active. On the other hand, the bottom trace demonstrates the synaptic weight change in the results of the same spike combination and timings, but when the complete triplet circuit, i.e. all red and blue parts, is active. Similarly, in the top purple trace in (c), the synaptic weight potentiation as a result of the pair-based pre-post interaction is shown to be smaller compared to the triplet-based weight change. The higher potentiation of the triplet-based STDP is due to the additional post-pre-post interaction, which is activated through the blue triplet parts of the circuit. This figure presents the first physical implementation of a Very Large Scale Integration (VLSI) synaptic device that accounts for higher order STDP rules.

### *2.1.2 CMOS Neuron Design*

A CMOS-based circuit design of an adaptive exponential Integrate and Fire (IF) neuron circuit [22, 34] is shown in Figure 3(a). The circuit is composed of several parts including an input differential pair integrator low-pass filter (ML1-ML3), a second low-pass filter (MG1-MG6) which implements spike frequency adaption, a non-inverting amplifier with current-mode positive feedback for Address Event Representation (AER) (MA1-MA6), and a reset block (MR1-MR6) for resetting the neuron and implementing the required refractory period. Figure 3(b) demonstrates measurement results from

This article is protected by copyright. All rights reserved

Author Manuscript

the neuron fabricated in a 0.35 um AMS CMOS process[18]. Here, the silicon neuron's membrane voltage stored on $C_{mem}$ is shown in response to a constant current injection ($I_{in}$). This closely resembles the spiking behavior of cortical neurons measured in response to somatically injected currents[35]. As the figure depicts, the neuron can be controlled to exhibit behaviors similar to biology.

## 2.2 SiOx Memristive Synapse and Neuron Design

Memristive and Resistive Random Access Memory (RRAM) devices have been extensively used to implement synapse and neuron circuits for neuromorphic devices, circuits, and systems[36]. In this paper, we focus mainly on Silicon Oxide (SiOx) memristive devices for neuromorphic computing.

SiOx is commonly used as a gate dielectric for Metal-Oxide-Semiconductor Field Effect Transistors (MOSFET) due to its stable physical and chemical characteristics, i.e. relative dielectric constant: 3.9, energy gap: 9 eV, dielectric strength 13.5 MV/cm, and thermal stability > 1050 oC[37]. In addition to its excellent insulating properties, CMOS compatibility, and controllability, SiOx-based resistive switching phenomena have recently been demonstrated in a vacuum[38] and in ambient atmospheric conditions[39] indicating that this traditionally passive material can be utilized as an active memory element (memristor), controlled by an external electrical field[40, 41].

Furthermore, the microstructure plays a crucial role in air stable resistance switching in pure silicon oxide. The columnar structure in sputtered silicon oxide films provides preferential parts for filaments formation[42]. The control of the interface roughness between the bottom electrode and the switching layer affects both the switching voltages and endurance[43]. Apart from the intrinsic CMOS compatibility, switching properties of silicon oxide-based RRAMs compare favorably against other more commonly used RRAM materials. We direct readers to the extensive review[39]. These features make SiOx an attractive material for neuromorphic computing.

Many different mechanisms could govern the resistance switching in silicon oxide. Generally, the filamentary resistance switching mechanisms are classified into those that are intrinsic to the

This article is protected by copyright. All rights reserved

Author Manuscript

oxide material – commonly known as valance change or intrinsic switching mechanisms, and those that involve metallic diffusion from electrochemically active electrodes (e.g. silver or copper) or metal doping to form metallic filaments – commonly known as electrochemical metalization, conductive bridge, or extrinsic switching mechanisms[44].

Here, we only focus on intrinsic switching in silicon oxide that can further be subdivided into air-stable switching and air-sensitive switching[39]. The first type has been studied with devices with exposed oxide surfaces, and also in bulk and porous oxides[45–47]. Air-sensitive switching occurs only in a vacuum as the oxidation of conductive filaments occurs in an oxidizing ambient. The second type, air-stable switching is possible either in oxygen-rich or non-oxidizing environments, as conductive filaments form far from surfaces and are more critically affected by the oxide microstructure[42, 48, 43]. The taxonomy of resistance switching in SiOx-based RRAMs is shown in Figure 4.

### *2.2.1 SiOx Memristive Synapse Design*

SiOx memristive devices are used to implement both STDP[49, 50] and synaptic weights in physical implementations of artificial neural networks[15, 51]. STDP is often regarded as a key local learning rule in biological systems, modulating synaptic weights in accordance with the degree of temporal overlap between pre- and post-synaptic action potentials. It has been shown that memristive devices can implement a pulse timing dependent change of resistance by suitably overlapping voltage pulses[52]. Most realizations use bipolar resistance switching, but unipolar SiOx RRAMs can perform STDP plasticity by controlling set and reset processes.

In this case, the STDP response is generated by using identical square voltage pulses, where the post-synaptic pulse is modified by a capacitor and converted into a triangular pulse before being applied to an RRAM device. The purpose of the pulse modification is to achieve the desired device response to leading and trailing pulse edge slopes. The concept is described in detail in[50]. Figure 5

This article is protected by copyright. All rights reserved

Author Manuscript

shows the results of the implementation of STDP-like response in unipolar SiOx-RRAM devices. In this figure, (a) shows a non-identical STDP-mimicking pulse set up. If a square pulse and a triangular pulse are below threshold there is no change in device resistance if a single pulse is applied. However, if the sum of these two pulses is above the threshold it is possible to adjust device resistance. In these examples the square pulse is a pre-synaptic spike and the triangle pulse is a post-synaptic spike. The pre-synaptic spike arrives earlier than the post-synaptic spike; the resulting sum is a slow leading edge above the threshold. This leads to a decrease in resistance (increase in conductance–SET process). The post-synaptic spike arrives earlier than the pre-synaptic spike; the resulting sum is a slow trailing edge above the threshold. This causes an increase in resistance (decrease in conductance–RESET process). Figure 5(b) shows the percentage of the expected successful operation, i.e. if a conductance decrease or increase is expected for a specific $\Delta t$, does that happen or not on the actual device. Figure 5(c-d) show the resulting STDP synaptic weight (conductance) changes in response to $\Delta t$ values in the range -600 to +800 $\mu s$.

All the above memristive synapse results are for air-stable devices. However, one-diode-one-resistor (1D-1R) air-sensitive test structures can also be used for vacuum-type SiOx-based synapse device demonstration, as shown in Figure 6. In this figure, (a) demonstrates a secondary electron microscopy (SEM) image of the top-down view of the fabricated device, while (b) shows a cross-section image of the 1R device, including its layer information. The synapse device structure and characteristics of this 1D-1R design have been described in detail in[53].

To fabricate this device, the active SiOx memory layer is deposited to a thickness of 40 nm using Plasma-Enhanced Chemical Vapor Deposition (PECVD). The Reactive Ion Etch (RIE) step then clears out the SiOx layer inside the hole, and creates a SiOx sidewall where the memory device is formed (Figure 6(b)). The active memory area of the 1R device is $2 \times 2$ $\mu m^2$ and the overall size including metal interconnects is $21.9 \times 21.9$ $\mu m^2$. The overall size of the 1D device is $41 \times 19$ $\mu m^2$

.

This article is protected by copyright. All rights reserved

Author Manuscript

Figure 6(c) shows an endurance of sequential Long Term Potentiation (LTP) and Long Term Depression (LTD) behaviors using non-identical pulses in 1D-1R architecture with different voltage increment steps for potentiation and depression: 0.1 V (top), 0.2 V (middle), and 0.3 V (bottom). For depression, the pulse height modulates from 11 V to 17 V with 10 $\mu s$ pulse width; for potentiation, the pulse height changes from 4.0 V to 10 V with 10 $\mu s$ pulse width. Such flexible artificial control built with synaptic devices could provide a suitable platform for a broad weight range of computing applications. Some of the advantages that SiOx-based synaptic devices provide over other resistive switching materials include a higher dynamic range ($10^4$) and the potential to achieve as many as 10-60 multi-level states (dependant on the stability, e.g. retention and endurance) in both LTP and LTD by changing the increment/decrement of the voltage step, as shown here.

Figure 6(d)-(e) demonstrate that the SiOx-based 1D-1R synaptic device can mimic STDP. These figures show a total of 10 different conductance levels of the device, when positive and negative spike timing differences, as well as spike widths are used, to emulate the potentiation and depression window behavior of STDP, observed in experimental[30, 49] and computational experiments[32].

### *2.2.2 SiOx Memristive Neuron Design*

Memristive and RRAM devices are more commonly used to implement synaptic weight plasticity in Spiking Neural Networks (SNNs) or to represent weights in artificial neural networks. However, unipolar SiOx-RRAMs have also been considered for modeling aspects of the electrical activity of the neuron. It has been demonstrated that specific operational procedures could lead to the generation of controlled voltage transients that resemble spike-like responses seen in biological neurons. Additionally, the integration and thresholding capabilities that are crucial for neuronal functionality have also been demonstrated. Further, redox-based resistance switching models can be related to the Hodgkin-Huxley (HH)[54] conductance model by analyzing the equivalent electrical circuits[55], and are

This article is protected by copyright. All rights reserved

Author Manuscript

found to be very similar.

Figure 7 demonstrates the thresholding, spiking, and integration capabilities of SiOx RRAM-based neurons. The main idea is to control the competing set and reset processes (the former being field-driven; the latter current-driven) by stressing devices with appropriate current inputs. A constant current bias is used to test thresholding and spiking functionality, while current pulses are used to implement integration.

The unipolar switching in SiOx-RRAM devices is utilized to generate voltage transients (resembling voltage spikes) by applying a constant current bias. By applying currents larger than the reset current, the device is put in the metastable state, fluctuating between the LRS and the HRS. As a result, the voltage spiking is measured at the output of the device. If the input current is lower than a threshold (reset) current, the resistance states are stable, therefore, no voltage spiking is observed. Figure 7(a) demonstrates threshold spiking behaviour. Furthermore, the integration functionality is demonstrated by applying a train of excitatory current pulses while changing the timing between the pulses. A much smaller current pulse senses the voltage across the device. Figure 7(b-f) show the obtained results. The frequency of input current pulses controls the rate of voltage spiking. The results resemble the Leaky Integrate-and-Fire (LIF) model. A more in-depth analysis of the results and the experimental setup can be found in[55].

In addition to the neuron spiking behavior achieved in Figure 7, a simple one-resistor-one-resistor (1R-1R) test structure can be used for air-stable type SiOx-based neuron device demonstration, as shown in Figure 8. The detailed neuron device structure and characteristics have been described in[38, 56]. For this neuron design, two device structures have been fabricated; (i) SiOx/HfOx stacking bilayer as shown in Figure 8(a-b), and (ii) vanadium electrode (V)/SiOx single layer demonstrated in Figure 8(d). For SiOx/HfOx stacking bilayer structure, SiO2 layer of 30 nm thickness has been sputtered and HfO2 of 1 nm thickness (confirmed by TEM image, as shown in Figure 8(a)) has been deposited using Atomic Layer Deposition (ALD) at 250 C. The TaN layer of

This article is protected by copyright. All rights reserved

Author Manuscript

170 nm that serves as a top electrode has been deposited also through sputtering. For the V/SiOx single layer structure, vanadium electrode of 200 nm and SiO2 layer of 6 nm have been deposited by sequentially sputtering without a vacuum break.

The SiOx/HfOx stacked structure exhibits excellent reliable threshold switching in air with ultra-low operation voltage (< 0.5V) (Figure8(b)). The non-linearity/selectivity of this device is about 180, which can stably operate at 0.3 V threshold voltage, while the holding voltage is 0.1 V, as shown in Figure 8(c). The self-compliance current is also observed here, which is due to the internal filament limitation[57]. Figure 8(d) shows the other SiOx-based threshold switching structure formed by V/SiOx single layer stacking. The non-linearity and selectivity of this device is about 102, which can stably operate at 1.2 V, where the device threshold voltage is 1.1 V. Note that, an electroforming process is necessary before any operation[38, 56].

One of the promising features of these two threshold switching devices is that they are able to demonstrate the all-or-none principle in the neuronal behavior. It means that, if a neuron responds at all, then it must respond completely. Also, it means that a greater intensity of stimulation does not produce stronger spiking, but can increase firing frequency[58, 59]. Figure 8(e) shows the schematic of the simple 1R-1R test circuit and its actual experimental setup to realize the all-or-none principle bio-mimetically. By applying the green input pulse shape as shown in Figure 8(e) and measuring the voltage between the resistance and the measurement instrument (Agilent DSO9254A), we can distinguish the all-or-none principle, as shown in Figure 8(f) in blue color. When the applied voltage (CH2, 2 V) is below the threshold voltage (note to the voltage divider setting, which is 0.4 V for SiOx/HfOx stacked structure and 1.1 V for V/SiOx structure), there are no spikes firing. However, when the applied voltage is at or above the threshold voltage (3.7 V), a sequential and repeatable 1.2 V output spike (peak-to-peak voltage) with 25 $\mu$ s period, is generated.

The above memristive neurons only replicate the threshold-based spiking integrate and fire behavior of neurons. However, biophysical neurons are considered to provide richer dynamical

This article is protected by copyright. All rights reserved

Author Manuscript

behavior and more computational complexity in comparison with their bioplausible counterparts such as integrate and fire (IF) and LIF neurons. HH neuron[54] and its simplified version such as Morris-Lecar (ML)[60] are members of the biophysical neuron family and due to their neuro-computational properties, non-biomimetic CMOS circuits are not capable of reproducing their dynamics efficiently in terms of energy and area. The introduction of memristor devices has transformed the design of HH neuron circuits[61, 62], whose design progress was slowed due to the lack of built-in stochasticity in CMOS circuits for building more realistic brain-plausible neuromorphic platforms[63, 64].

We have previously shown that the known functionalities of HH[65] and ML[66] neuron models can be simulated using behavioral memristive models. Figure 9 shows two different panels, each describing one bio-inspired memristive neuron circuit diagram, memristive channel characteristics, and bifurcation behaviors at different frequencies, for a different neuron model. In the top panel of Figure 9, (a) shows the basic circuit schematic for HH neuron memristor-based circuit model with memristive behavior of sodium and potassium channels. A voltage-gated potassium (sodium) channel is emulated by positively (negatively) biased memristor devices coupled with a membrane capacitor. In this panel, (b) shows the tonic spiking behavior of memristive HH neuron and potassium (K+) and sodium (Na+) channels for HH model. In the bottom panel, (c) shows the Morris-Lecar neuron model circuit diagram and the behavior of potassium and calcium channels for ML memristive circuit model in response to a 2.5 V sinusoidal input voltage at different frequencies. The Potassium channel memristor i-v curve is shown at 5 Hz, 50 Hz and 5 kHz (from left-to-right). The calcium channel memristor i-v is displayed at 100 Hz, 500 Hz, and 50 kHz (from left-to-right). The scaled ML neuron bifurcation behavior for different current stimulus is shown in Figure 9(d). The membrane voltage, calcium channelâ€™s current, potassium channelâ€™s current and state variables (M, N) have been also displayed for the stimulus currents of 60 $\mu$ A and 100 $\mu$ A.

In the last two sections, we presented a number of designs implementing synaptic and

This article is protected by copyright. All rights reserved

Author Manuscript

neuronal functionalities such as LTP, LTD, and STDP for memristive synapses, and all-or-none and threshold spiking behavior for memristive neurons. Except for the simulation-based HH and ML memristive neuron models discussed, all other designs explained use SiOx structures. In addition to the design presented above, SiOx-based devices have been extensively used in other neuromorphic systems, specifically for synaptic behavior demonstration. Here, we provide a comparison of various SiOx-based memristive synaptic devices discussing their functionalities and physical/switching structures, as shown in Table 1.

As the Table shows, various techniques have been used to implement different devices with diverse layer stacks, deposition methods, switching polarity, switching type, and SiOx thicknesses. These designs that operate at different SET/RESET voltages, as low as 0.5 V and as high as 10-15 V offer various degrees of functionality and on/off ratios. Some of these devices also present a multilevel structure, which is useful in implementing neuromorphic computing systems. It is worth noting that, not all designs, except for two of our previous works presented in[51, 53], are able to implement the three investigated synaptic plasticity functionalities, i.e. LTP and LTD, and STDP.

Note that in Gao's study[67], a porous silicon oxide structure has been used. The main function of SiOx here is to facilitate lithium-ion migration in forming a conductive path, with well-controlled lithium ions in the SiOx layer. These features enhance neuromorphic computing performance, i.e. LTP and LTD, similar to those observed in the SiOx device developed in[68]. In the work by Wang et al.[69], an opposite SiOx functionality is used to enhance the resistive switching characteristics for analog resistance modulation. The SiOx here was found to have a denser microstructure compared with the TaOx layer. This microstructural attribute could lead to the retardation of oxygen diffusion in the SiOx layer. Also, the diffusion constant is 1–6 orders of magnitude lower than that in the sub-stoichiometric TaOx. Therefore, in their proposed device, the diffusion rate of oxygen ions is expected to decrease in SiOx. As a result, the resistive switching in the cell is likely to take place in the more stoichiometric and insulating SiOx layer that is in series with the sub-stoichiometric and more conductive TaOx layer (valence change-type memristor). Similarly, in Acevedo's study[70], it is

This article is protected by copyright. All rights reserved

Author Manuscript

reported that the SiOx drives the observed resistive switching behaviors by taking and releasing oxygen ions to and from the nearby manganite layer. These results confirm that the combination of two active oxides (stacked structure) improves the neuromorphic functionalities compared to the single-oxide structures, as shown in Table 1. However, some materials (such as LiCoO2 or LCMO) may not be fully CMOS compatible.

Author Manuscript

| | Device stacks | Deposition method | Switching polarity | SiOx thickness | Set (V) | Reset (V) | On/Off Ratio | Multi level | LTP/LTD | STDP |
|---|---|---|---|---|---|---|---|---|---|---|
| Gao et al.[67] | LiCoO2 /Porous SiOx | Thermal | Bipolar | 150 nm | 4 | –4 | $< 10^6$ | Yes | Yes | – |
| Zarudnyi et al.[50] | SiOx | Sputtering | Unipolar | 37 nm | 3.7 | 2.5 | $< 10^2$ | – | – | Yes |
| Wang et al.[69] | SiOx/TaOx | Ion beam | Bipolar | 1, 2, 4 nm | 0.5 | -0.5 | $< 10^3$ | – | Yes | Yes |
| Acevedo et al.[70] | LCMO/SiO2 | Native | Bipolar | 2 nm | 5 | -5 | $< 10^2$ | Yes | – | – |
| Chen et al.[68] | SiOx | PVD | Bipolar | 100 nm | 1 | -1 | $< 10$ | Yes | Yes | – |
| Chang et al.[53] | SiOx | PECVD | PECVD Unipolar | 40 nm | 4-10 | -15 | $< 10^2$ | Yes | Yes | Yes |
| Mehonic et al.[42, 51] | Mo/SiOx/Ti/Au | reactive sputtering | Bipolar | 35 nm | -1 | 1 | $< 10^4$ | Yes | Yes | Yes |

**Table 1.** Comparison of switching characteristics and parameters of SiOx-based resistive switching devices.

This article is protected by copyright. All rights reserved

## 2.3 Hybrid CMOS-Memristive Neuromorphic Designs

Although neuronal and synaptic behaviors have been replicated using memristors, as discussed in the previous sections, silicon and CMOS-based neuron designs have shown more convenient implementation processes and better functionalities[22]. Unlike neurons, synaptic behaviors and functionalities are more conveniently implemented using memristive devices. Therefore, one may choose to use memristive devices to implement synaptic learning in a neuromorphic system, while all other components of the system are implemented using analog and digital CMOS[15]. This requires seamless integration of CMOS and memristive technologies in a hybrid design, so that the benefits of both domains are achieved in the final system.

There have been many previous hybrid CMOS-memristor neuromorphic designs in the literature[71, 15, 72, 73]. In most of these systems, the neuron and interfacing circuitry are designed in CMOS, while synapses implementing targeted synaptic plasticity rules, such as STDP are realized using one or a few memristors, which are programmed through shared or individual CMOS circuits. Here, we show a hybrid CMOS-memristor neuromorphic synapse that implements rate-based synaptic plasticity in the form of Bienenstock Cooper Munro (BCM)[74] rule, as a neuromorphic component. In addition, in Section 3.3, we show a hybrid system implementing spike-based image sensing[16].

In a previous work[15], we proposed a hybrid CMOS-memristor neuromorphic synapse that was shown to be capable of reproducing a number of biological experiments including pair-based, triplet, and quadruplet, similar to the CMOS circuit shown in Figure 2(a). The hybrid synapse circuit that is connected to a pre-synaptic and a post-synaptic neuron is shown in Figure 10(a). Here, we show for the first time that this circuit can generate BCM-like behavior, which is observed in biological experiments[75]. Figure 10(b) shows the weight modifications achieved using the circuit in (a), driven by random pre and post-synaptic Poissonian spike trains. This experiment is composed of 10 trials. In

This article is protected by copyright. All rights reserved

Author Manuscript

each trial, a random pre-synaptic spike train with the rate, $\rho_{pre}$, along with a random post-synaptic spike train with the rate, $\rho_{post}$, both of a 10 s duration, are applied to the bi-memristive hybrid synapse. The rates of pre and post-synaptic spike trains are considered equal, and are swept over the range of 0-50 Hz. Figure 10(b) demonstrates the average weight change and their standard deviations over the 10 trials. This figure depicts the main characteristic of the BCM rule, which is turning LTD to LTP at a specific spike rate, i.e. the BCM threshold. Additionally, this figure shows the sliding threshold feature of the BCM rule, which can alter the frequency at which LTD turns to LTP. This turning threshold as shown in[32] can be controlled using the STDP parameters. Here we have utilized the triplet potentiation amplitude, $A_3^+$, to slide the threshold towards lower or higher depression[15].

Although previous CMOS circuits implementing triplet STDP has also shown the ability to realize BCM behavior, they occupy large silicon real estate, mainly due to the large capacitances required to store the synaptic weight and to provide time constants required for triplet STDP dynamics. However, the proposed memristor-based hybrid circuit, compared to its CMOS counterparts, occupies up to 10 times smaller chip area[15]. This feature makes it a promising component for large-scale neuromorphic systems with spike-based learning capabilities and encourage designers to benefit from hybrid designs in more applications, such as neuromorphic image sensing.

## 3 Neuromorphic Systems Design

In Section 2, we discussed the design and implementation of various CMOS, memristive, and hybrid CMOS-memristive neuromorphic circuits that were designed and implemented to realize neuronal or synaptic behaviors. We showed that the devices and circuits designed are able to replicate some known functionalities and behaviors of biological systems, such as LTP, LTD, STDP, triplet STDP, all-or-none spiking, threshold spiking, HH and ML neuron channel characteristics and bifurcation

This article is protected by copyright. All rights reserved

Author Manuscript

behaviors. In this section, we utilize several of the designed circuits developed and explained earlier to implement neuromorphic computing systems applied to spike-based as well as artificial neural network learning. We also discuss the design and implementation of a hybrid CMOS-Memristor neuromorphic sensing system.

Author Manuscript

## 3.1 CMOS Neuromorphic Computing

A neuromorphic system can be implemented completely using digital[76, 77] or using a mixed of analog and digital CMOS technology[27, 18]. If the system is implemented in digital, the behaviors of neurons and synapses are approximated and the spiking and Address Event Representation (AER) structure of the system is realized in digital architectures[76, 8, 78]. However, if the system is designed in mixed analog-digital domain, the synaptic and neuronal behavior are replicated closely in analog, while AER and interfacing is realized using digital technology[9].

Most of the previous CMOS-based neuromorphic systems include fixed learning and neuron circuits, where an array of analog synaptic cells implementing a fixed learning rule, such as STDP[28] or SDSP[25], is connected to an array of IF neurons. These implementations offer great biomimicing properties because their components closely mimic biological experiments, however, they can only be used with the learning rules implemented in their hardware. Some neuromorphic systems, such as[18] provide more flexibility by giving the user the ability to describe the network structure as well as the required learning rule in software, while the neuron and synapse components are implemented in analog hardware. The AER and spike transmission in these systems use digital technology and are usually implemented on Field Programmable Gate Arrays (FPGAs)[79] that facilitate programmability and reproduction[76].

Here, we describe the operation and use of such a hardware-software system[18] that uses CMOS circuits providing programmable synaptic learning. This programmability gives the user the freedom to explore any arbitrary spike-based learning rule. The design that is shown in Figure 11(a) is

This article is protected by copyright. All rights reserved

composed of (1) an asynchronous SRAM array that can be programmed as virtual synapses, connected to (2) an array of CMOS synapses, which integrate a current proportional to the weight stored in their corresponding SRAM cells, to feed into (3) IF neurons similar to the neuron explained in Figure 3, which receive/transmit spikes from/to other neurons/synapses in a set network, through (4) asynchronous control and interfacing circuits that manage the AER communication. All analog components on the chip are tuned by biases received from (5) a bias generator circuit.

Figure 11(b) shows an example spiking neural network that has been implemented on the chip shown in (a). The chip receives AER input spikes, which contain information on (1) the address of the post-synaptic neuron, (2) the address of the programmable SRAM-based virtual synapse connected to this neuron, (3) the type and address of the physical integrator synapse, and (4) the new 5-bit weight value that will be written to the addressed virtual synapse (SRAM cell). The AER output spikes, on the other hand, only show the address of the post-synaptic neuron that generated them.

Using the network shown in Figure 11(b), two UP and DOWN rate-based spike patterns (shown at the top of Figure 11(c)), with 20% correlation (6 fixed common input spike trains shown with red circles, from the total 30) are learned through triplet STDP[32], to be classified unsupervised. The four panels in the figure show the output neuron firing rate during learning. In the first trial, learning has not happened yet, hence the output neuron fires with a similar rate for both UP and DOWN patterns. As the learning progresses, at each trial either UP or DOWN spike-based pattern is randomly selected to be applied to the network. Also, a new Poissonian spike train is generated for each of the 18 active synaptic inputs for either UP or DOWN. It is seen that after 20 trials, the output neuron successfully distinguishes between the two patterns, showing a higher firing rate for UP patterns. The figure shows the results for 20 runs, each including 20 trials. For each run, the initial synaptic weights, the order in which UP and DOWN patterns are applied to the network over 20 trials, as well as the definition (spike-timing-intervals and distribution) of UP and DOWN patterns are randomized. Note that, the learning rule used for this classification, implemented in software to program the SRAM virtual synapses on the chip, provides the same functionality as the circuit shown

This article is protected by copyright. All rights reserved

Author Manuscript

in Figure 2(a). We have shown that, this classification task can achieve 100% accuracy, even in the presence of 87% (i.e. 26 common input spike trains) correlation between the two patterns.

## 3.2 SiOx Memristive Neuromorphic Computing

Spiking and artificial neural network systems that utilize memristors as their learning component or simply as a programmable element, has been researched extensively[15, 19, 80–83]. Here, we mainly focus on neuromorphic systems that utilize SiOx memristive devices for learning, where memristors are used as weight elements. We also show simulation results of neuromorphic learning using spike-based learning and STDP, realized using ideal memristor models. However, there exist challenges in the use of memristive devices as learning components, which should be considered in neuromorphic learning systems design. Here, we show and discuss some of these challenges that mainly arise from the unavoidable device non-idealities.

In addition to the ability to replicate STDP-like behavior for spike-based learning (shown in Figure 5) memristive and RRAM crossbars offer the execution of vector-matrix multiplication in the analog domain[19]. When voltage pulses are applied on one side of the programmed RRAM crossbar, and current is sensed on the orthogonal terminals, this system provides approximate vector-matrix multiplications (multiply and accumulate (MAC)) operations in constant time steps using Kirchhoff's and Ohm's laws[84]. This approach promises speed and power efficiency improvements of many orders of magnitude compared to conventional CMOS systems[19]. However, there exist obstacles on the way of utilizing memristive devices as STDP-enabling synaptic devices or MAC operation facilitators.

The main obstacle to realizing the full potential of RRAM crossbars is the number of both device-level and system-level non-idealities. These include device-to-device and cycle-to-cycle variabilities, a limited number of resistance states or a narrow operational range of resistance modulation, non-linearity of both voltage-current characteristics and pulse response of devices, and

This article is protected by copyright. All rights reserved

Author Manuscript

high line resistances in crossbar arrays. Despite these, SiOx-RRAM devices have been used to implement weights in physical neural networks, in which multiple weight values (memristive resistance states) are required[51].

Figure 12 (a-b) demonstrate the feasibility of achieving multiple resistance states in intrinsic SiOx RRAM devices by controlling the reset process (i.e. varying the reset voltage) during voltage sweeps. Stable resistance states are achieved as devices reset gradually from the Low Resistive State (LRS) to the High Resistive State (HRS). However, for practical applications, pulse operation is preferred. In this case, the gradual reset dynamics is highly dependent on voltage amplitude and pulse width, as clearly seen in Figure 12(c) and 12(d). Programming curves that gradually switch devices from the LRS to the HRS (see for instance Figure 12(d)) are typically called Long Term Depression (LTD), and are akin to synaptic functionality. As these figures show, the memristive devices could be challenging for implementation of distinct resistance levels required for learning and inference in artificial neural networks. In addition, other non-idealities such as device failure and fabrication yield are present that can affect the network performance.

We have studied extensively the effects of various memristors non-idealities on inference accuracy, using the MNIST handwritten digits data set[51], where the neural network weights are represented as memristive device resistant states as shown in Figure 13. Figure 14 shows examples of the impact of yield and device failure on inference accuracy, as well as that of varying the number of resistance states with different schemes to map device resistance onto ANN weights as shown in Figure 13.

From these figures, it is evident that higher device non-idealities result in lower inference accuracy in the implemented classification task for MNIST. The figure also suggests that, small device non-idealities (< 10%) does not significantly affect the performance of the implemented SiOx-based neuromorphic system, which uses memristors as programmable weights, while learning happens through supervised Back-Propagation (BP) [51].

This article is protected by copyright. All rights reserved

Author Manuscript

In a similar study to the one explained above, we developed another SiOx-based memristor neuromorphic learning platform, in which a three-layer neural network (shown in Figure 15(a)) was designed and trained for the classification of MNIST dataset (784x64x10) and Topology Patterns (TP) of the atmosphere effect (256x64x7) [38]. Note that, the experimental LTP and LTD property of the synaptic devices was used to devise a function of synaptic weight (W) with respect to the input pulse numbers [85]. During the training process, the feedback error of network weights obtained from the BP algorithm was modulated using the function. Then, the deviation between the output and the target signal was minimized in the learning process. We found that, the network learning performance relied heavily on the linearity and accuracy of the weight resistance tuning process that was developed for the memristive network.

Figure 15(b) shows that the best learning accuracies achieved for MNIST and TP are around 89% and 80%, respectively. The figure also depicts that there is no clear difference between various device-to-device variation cases. Furthermore, the accuracy correlation with LTP and LTD nonlinearity (as defined in[85]) is shown in Figure 15(c). It is shown that the implemented neuromorphic learning system that uses SiOx-based weight reaches simulated accuracy of up to 90%, but the accuracy decreases with the increase of nonlinearity by a slope of 6.

The promising neural network learning results shown in Figure 15 was achieved using simple SiOx-based Metal-Insulator-Metal (MIM) structure in place of the network weight elements, in a crossbar structure. In addition, it has been shown[86, 87] that such MIM devices simplify memory array design if used in a crossbar architectures. However, the leakage through the sneak paths inevitably induced while accessing MIM crossbar networks may cause weight variation issues detrimental to the development of reliable memristor-based neuromorphic learning and computation [88]. To mitigate the sneak paths currents, a diode or a selector device is usually positioned in series with a memristor cell to form a 1D-1R or 1S-1R structure[89, 90]. These configurations considerably increase fabrication

This article is protected by copyright. All rights reserved

Author Manuscript

and circuit design complexity and cost.

To address this problem, in Figure 16, we have developed a selectorless memristor, which does not require a 1D or 1S series connection. This is achieved by using simple high-k/low-k bilayer stacks (as shown in Figure 16(b)). The intrinsic selectorless property (or the nonlinear I-V characteristics, as shown in Figure 16(a)) can be realized by inserting an ultra low-k layer (e.g. SiOx layer or graphite oxide layer)[91]. The main benefit here is that the ultra low dielectric constant layer provides the nonlinearity intrinsically by carrier transport formulation design, specifically in Poole-Frenkel defect cases[92].

Here we use this selectorless memristor to demonstrate neuromorphic learning. Figure 16(a) shows typical bipolar resistive switching I-V characteristics during DC voltage sweeps for HfOx single layer (H11), and HfOx (7 nm)/graphite (5 nm) stacked (H7G5) bilayer selectorless memristive devices. The sneak path current in H7G5 device can be avoided by taking advantage of the nonlinearity in the self-rectifying I-V characteristics of this device. The I-V characteristics shows that the H11 has a higher sneak path current at low resistance state (LRS). Therefore, the H7G5 structure can be used that shows the self-rectifying behavior, where the LRS current is suppressed and sneak paths current can be eliminated.

Figure 16(b) shows the TEM image of the H7G5 stacked selectorless memristor device. The device structure and fabrication processes have been fully described in [93]. The immunity to the sneak path currents introduced through the nonlinearity of the selectorless devices introduced here, mitigates the potential weight variability in neuromorphic learning and computing applications. Moreover, the selectorless device introduced can result in larger crossbar sizes as shown in the figure. This results in a larger number of wordlines, which in turn means a bigger network suitable for extended computing demands.

Figure 16(c) shows LTP (top) and LTD (bottom) behaviors using identical pulses method to examine the conductance flexibility and modulation in H7G5 devices. For depression, the pulse height is -0.7 V with 50 $\mu$ s pulse width; for potentiation, the pulse height is 1.54 V lasting for 100

This article is protected by copyright. All rights reserved

Author Manuscript

$\mu$ s. Figure 16(d) shows the simulation accuracy results obtained using H11 and H7G5 selectorless devices mapped into the three-layer neural network shown in Figure 15(a) for classifying the handwritten digits MNIST dataset. As the figure depicts, the best classification accuracy achieved using the network composed of the H7G5 devices is around 85%, which is far better than that of the network using H11 devices (30%). This network uses selectorless RRAM devices as its weight elements, which suppress the sneak path current without the need to additional transistor compared to previous designs.

In all aforementioned learning experiments performed using SiOx devices, neural networks were simulated with realistic device models, where learning happened through machine learning approaches such as backpropagation. In these networks, the memristive devices were mainly used as weight elements, which are programmed to represent a certain resistance state. Therefore, no unsupervised learning through neural-inspired algorithms such as STDP, similar to that shown in Figure 11(c), happened. However, it has been shown that memristive devices can be used to perform unsupervised learning through synaptic plasticity mechanisms used in neuromorphic systems. In the next section, we provide simulation results of a memristive neuromorphic system that uses STDP for unsupervised learning of simple patterns.

## 3.3 Hybrid CMOS-Memristive Neuromorphic Sensing

The use of memristors as a processor of input sensory information has rapidly gained traction, but has only had limited success in the front-end generation of sensory data. For example, in[16, 94] both works use memristive meshes for spatiotemporal smoothing of information as a way to reduce noise through averaging. The resistive power dissipation, and lack of equal set/reset processing speeds has made it too difficult to justify the additional fabrication steps required to implement thin film metal-oxide memristors in the back-end-of-the-line (BEOL) for CMOS integration.

This article is protected by copyright. All rights reserved

Author Manuscript

In the circuit shown in Figure 17(a), we present a novel memristor-CMOS image circuit that utilizes temporal storage of information in RRAM to block signals in the absence of temporal change in light intensity. This circuit is an extension and improvement on what is presented in the image sensing circuit from[16], which relied on the use of a separate match line for voltage comparisons to a constant reference. A constant reference means that an output was generated at the source of transistor M3 when input intensity exceeded some threshold; that is to say, there was no detection of temporal difference of intensity, and therefore light adaptation was a mechanism of DC thresholding. We draw inspiration from dynamic vision sensors here, but rather than using a capacitance to block DC content, we rely on the identical signals generated from the application of a read pulse to a pair of identically programmed memristors at two consecutive points in time to block the output. The voltage and memristor state X read-out is shown in Figure 17(b), and the use of the match line alone in a spike-based programming scheme is in Figure 17(c). The output signal from the image sensing circuit is passed through a sub-threshold amplitude-to-frequency converter which generates voltage spikes, in a manner similar to that of the retinal ganglion cells[95, 96], the schematic of which is depicted in Figure 17(d), and array-based architecture of image sensor interfacing to subthreshold cell circuits in Figure 17(e). The final spiking output results are shown in Figure 17(f) for a given input of light intensity change. The chip is fabricated in the SK Hynix 180nm process, $V_{DD}$ = 1.8 V, $V_{th}$ = 0.4 V. Here, $V_{G_{OUT}}$ is the voltage converted current output, while $V_{A_{ON}}$ and $V_{G_{IN}}$ are the voltage converted inputs from the preceding stage. Note that the response is shifted negatively to replicate a biologically plausible resting membrane potential.

## 3.4 Simulation of Memristive Neuromorphic Computing

Biophysical neuron models, such as those developed for HH and ML neurons, have shown memristive behavior in their footprint[97, 98]. There exist only a few experimental implementations [62, 99]

This article is protected by copyright. All rights reserved

Author Manuscript

of memristor-based biophysical neurons due to the high complexity of the neuronal dynamics and bifurcation behaviors. To the best of our knowledge, there is no memristive biophysical neuron-based network, which has been tested for any type of applications. To address this challenge, one approach is to develop and utilize simulation frameworks for memristive neuromorphic networks. Significant research[64, 100–102] has been conducted on simulating memristive networks, where experimental implementation cannot be carried out due to the network size, complexity, technology limitations, etc. The simulation analysis can provide insight into the dynamical behavior of the equivalent memristive circuit model of various neuron models and the learning behavior and performance of their network.

We have developed and simulated simple cross-bar structured memristive networks to validate the functionality of the biologically-inspired ML and HH memrisitive neurons (that were shown in Figure 9), in a simple pattern classification task. Figure 18(a) illustrates STDP learning mechanism for two coupled HH neurons with memristive synapses in a small-scale crossbar structure. Figure 18(b) shows the membrane voltage of pre- and post-synaptic HH neurons connected by a memristive synapse for a 15 $\mu$A stimulus current. Figure 18(c) shows a 2x2 two-layer perceptron network with HH neurons and memristor synapses, where the inputs are two classes of two-pixel images. The classification results of the two-pixel input patterns performed by the 2x2 memristive network are displayed in Figure 18(d). Here, the membrane voltages of the pre-synaptic (black) and post-synaptic (blue) memristive HH neurons are shown. Initially, post-synaptic neurons spike without any specific pattern, and it takes some times for them to follow the input patterns. The learning is unsupervised and the winner neuron follows one of the patterns based on its initial weight vector. When the class 2 image is assigned to post-synaptic neuron 1, the weight of memristive synapse 1 increases due to STDP. Since post-synaptic neuron 1 spikes after pre-synaptic neuron 1 (a black pixel makes a neuron spike), the weight of memristive synapse 1 increases. Also, the weight of memristive synapse 2 decreases since post-synaptic neuron 2 is inactive while pre-synaptic neuron 1

This article is protected by copyright. All rights reserved

Author Manuscript

spikes.

Figure 18(e) shows a 4x2 two-layer perceptron network with ML memristive neurons and memristive synapses. In this case, inputs are two classes of four-pixel images. Four pre-synaptic neurons' membrane voltages are displayed in Figure 18(f), which show their spiking activity, after applying the input spike trains. LTD and LTP phenomena are displayed in Figure 18(g) during the classification in the memristive network on the synaptic devices. Finally, Figure 18(h) shows post-synaptic neurons' membrane voltages and the results show successful classification of the two classes of the input patterns, where after 5 s, neuron 1 fires for one pattern, while neuron 2 fires for the other pattern.

## 4 Discussion and Conclusion

Neuromorphic computing is a far-reaching field that encompasses the hardware implementation of basic artificial neural networks, through to biologically plausible spike-based systems. In both cases, learning and inference are the cornerstones of building functional large-scale networks that facilitate tasks that are fundamentally tied to neural cognition.

We have shown how CMOS and memristive systems have become highly pervasive in neuromorphic computing as they mimic the functional primitives of spiking and artificial neural networks alike, and reproduce neuronal and synaptic models at the device/physics-level. Whilst various RRAM switching mechanisms are available, we have focused on the use of filamentary-based switching and joule-heating modulation in SiOx memristors to construct substantially simplified neurons and synaptic circuits when compared to their purely CMOS counterparts. In essence, the integration of nanoscale non-volatile devices in the BEOL has completely removed reliance on physically disparate volatile DRAM or SRAM arrays, where information transfer is thwarted by restricted bus sizes.

The von-Neumann bottleneck in conventional computing systems is alleviated by leveraging

This article is protected by copyright. All rights reserved

Author Manuscript

analog domain in-memory computing in memristive crossbar systems capable of spike-based and non-spiking training and inference, both supervised and unsupervised. This was demonstrated in both oxygen migration of hafnium-oxide devices and filamentary growth within SiOx memristors, which may be integrated with active CMOS arrays or passive arrays by exploiting highly nonlinear characteristics that exhibit diode-like suppression of sneak paths currents. As cognitive neuroscientists better understand the learning mechanisms that occur within our neural systems, having a fundamental physical device-level primitive enables an optimized mode of implementation that can be used to successfully demonstrate more complex learning rules, and higher-order behaviors such as triplet-based, quadruplet-based STDP, and beyond.

In-memory processing has also been demonstrated in similar analog-domain form by implementing multi-bit flash cells, but these are bottlenecked by the lower limit technology nodes that floating gate transistors are fabricated in, and memristive devices appear to be the prime candidate for moving beyond such limitations. An alternative in both array-based computing and biological spike generation has been shown in the form of subthreshold analog circuits such as those shown in this paper, which take advantage of ultra-low power consumption, but suffer from extremely high RC delays which are introduced from high drain-to-source equivalent resistance values when the channel has not been fully inverted due to low voltages applied at the gate.

Neuromorphic engineering is not limited to drawing inspiration from the processing that occurs within the brain, and can be broadened to any biological process. We have demonstrated neuromorphic vision sensing by developing an RRAM-based adaptive image sensor, which is equally important as a method to optimize the front-end procurement of information that is transmitted to the processor. The notion that event-based processing strips input data of redundancy, and that the retina implements ultra-low power processing through adaptive vision is replicated in our sensory neuromorphic RRAM-CMOS integrated system. The realization of fully optimized systems may require similar approaches in integrating neuromorphic processing with front-end neuromorphic sensing.

This article is protected by copyright. All rights reserved

Author Manuscript

The advantages of RRAM-based neuromorphic computing are naturally counterbalanced by the requirement to convert read-out currents into digital-domain information for interfacing with computing systems, and this requires the use of Analog to Digital Converters (ADC) which partially offset the power, speed and area advantages of purely RRAM neuromorphic systems. Current integration techniques use active resistive switching layers above higher-level metal layers for any technology process, which enables us to take advantage of vertical real-estate of a chip and relaxing the burden of planar topological area. The same methodology has been applied in using FinFET technology to reduce the impact of electron traps in creating channels with larger cross-sectional area for CMOS scaling below 5 nm feature sizes. There is a continued need to explore the effect of noise associated with analog computing and quantization effects on precision, but with the silver-lining that neural networks can be designed to be robust to minor errors, signal fluctuations and non-idealities, making them optimal candidates for applications in soft-computing.

In addition, retention degradation (i.e. time-dependent RRAM state decay) and its impact should be considered in neuromorphic computing, as it is done in previous works such as[103]. At a temperature of 85° C, stacked TaOx devices have shown retention times of over 10 years[104]. Though when baking at higher temperatures, a 12% degradation in classification on the MNIST dataset is observed, and most degradation occurs after $10^4$ s[105]. As a result, a refresh operation is required in practical applications to maintain the accuracy.

To further improve retention, electrical design at the state-level (e.g. compliance current limiters) should be fine-tuned. From a structural and architectural perspective, oxygen exchange layer (OEL), and one transistor-one resistor (1T-1R) cells are needed to mitigate stochastic oxygen vacancy generation, annihilation, and overshooting pulses. The use of stacked-layer structures and thin-film plasma treatment are also helpful to eliminate the defective states recovery. At the board-level, a passive heat-sink is a common practical solution to prevent thermal throttling during heavy workload cycles.

It should also be noted that semi-volatile devices, such as WOx from[106, 107] which display

This article is protected by copyright. All rights reserved

Author Manuscript

millisecond retention characteristics may also be leveraged as leaky-integrate-and-fire neurons. This enables devices to reproduce time dependent spiking adaptations which have been observed in pyramidal neurons, and is useful in reservoir computing thus efficiently processing temporal information[108].

At this stage, researchers have a strong idea of the architecture of biological cells within sensory systems such as the retina, and to some extent, the brain as well. However, what is lacking is a unified account of the computation that takes place within biological sensors, and how they process, filter and store information before transmitting them to the brain. As neuroscientists converge towards a better understanding of the biological processing that takes place across the nervous system, CMOS and memristive researchers will be able to continue to cooperate in building large-scale systems to both better understand the cognitive circuits in performing functional and behavioral tasks such as pattern recognition and higher-order interpretation, but also to introduce the low-power and highly adaptive advantages of biological processes to conventional computing.

## Acknowledgments

Mostafa Rahimi Azghadi gratefully acknowledges the support of Prof Giacomo Indiveri and the Institute of Neuroinformatics at the University of Zurich and ETH Zurich for providing facilities to fabricate the chip and measure the results shown in Figure 2 and perform the experiments shown in Figure 11. He also acknowledges Dr Saber Moradi and Prof Indiveri for designing the circuit in Figure 3(a) and the chip, which its architecture is shown in Figure 11(a).

**References**

**1.** M. Le Gallo, A. Sebastian, R. Mathis, M. Manica, H. Giefers, T. Tuma, C. Bekas, A. Curioni, E. Eleftheriou, *Nat. Electron.* **2018**, *1*(4), 246–253

This article is protected by copyright. All rights reserved

Author Manuscript

**2.** D. Ielmini, H.-S. P. Wong, *Nat. Electron.* **2018**, *1*(6), 333–343

**3.** D. Strukov, G. Indiveri, J. Grollier, S. Fusi, *Nat. Commun.* **2019**, *10*(4838)

**4.** S. Lohr, *The New York Times.* **Retrieved 19 July 2018.**

**5.** A. Adamatzky, *Unconventional Computing: A Volume in the Encyclopedia of Complexity and Systems Science*, Springer, **2018**

**6.** G. Indiveri, S.-C. Liu, *Proc. IEEE* **2015**, *103*(8), 1379–1397

**7.** C. Mead, *Proc. IEEE* **1990**, *78*(10), 1629–1636

**8.** S. B. Furber, F. Galluppi, S. Temple, L. A. Plana, *Proc. IEEE* **2014**, *102*(5), 652–665

**9.** A. Neckar, S. Fok, B. V. Benjamin, T. C. Stewart, N. N. Oza, A. R. Voelker, C. Eliasmith, R. Manohar, K. Boahen, *Proc. IEEE* **2018**, *107*(1), 144–164

**10.** T. Wunderlich, A. F. Kungl, E. Müller, A. Hartel, Y. Stradmann, S. A. Aamir, A. Grübl, A. Heimbrecht, K. Schreiber, D. Stöckel, et al., *Front. Neurosci.* **2019**, *13*, 260

**11.** M. V. DeBole, B. Taba, A. Amir, F. Akopyan, A. Andreopoulos, W. P. Risk, J. Kusnitz, C. Ortega Otero, T. K. Nayak, R. Appuswamy, P. J. Carlson, A. S. Cassidy, P. Datta, S. K. Esser, G. J. Garreau, K. L. Holland, S. Lekuch, M. Mastro, J. McKinstry, C. di Nolfo, B. Paulovicks, J. Sawada, K. Schleupen, B. G. Shaw, J. L. Klamo, M. D. Flickner, J. V. Arthur, D. S. Modha, *Computer* **2019**,

This article is protected by copyright. All rights reserved

Author Manuscript

*52*(5), 20–29

**12.** N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, G. Indiveri, *Front. Neurosci.* **2015**, *9*, 141

**13.** Q. Xia, J. J. Yang, *Nat. Mater.* **2019**, *18*(4), 309–323

**14.** J. Chen, C. Lin, Y. Li, C. Qin, K. Lu, J. Wang, C. Chen, Y. He, T. Chang, S. M. Sze, X. Miao, *IEEE Electron Device Lett.* **2019**, *40*(4), 542–545

**15.** M. R. Azghadi, B. Linares-Barranco, D. Abbott, P. H. Leong, *IEEE Trans. Biomed. Circuits Syst.* **2017**, *11*(2), 434–445

**16.** J. K. Eshraghian, K. Cho, C. Zheng, M. Nam, H. H.-C. Iu, W. Lei, K. Eshraghian, *IEEE Trans. VLSI Syst.* **2018**, *26*(12), 2816–2829

**17.** R. Yang, H.-M. Huang, X. Guo, *Adv. Electron. Mater.* **2019**, *5*(9), 1900287

**18.** M. R. Azghadi, S. Moradi, D. B. Fasnacht, M. S. Ozdas, G. Indiveri, *ACM J. on Emerg. Technol. Comput. Syst. (JETC)* **2015**, *12*(2), 17

**19.** S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, G. W. Burr, *Nature* **2018**, *558*(7708), 60

**20.** Z. Wang, C. Li, W. Song, M. Rao, D. Belkin, Y. Li, P. Yan, H. Jiang, P. Lin, M. Hu, et al., *Nat. Electron.* **2019**, *2*(3), 115

This article is protected by copyright. All rights reserved

Author Manuscript

**21.** M. Mahowald, R. Douglas, *Nature* **1991**, *354*(6354), 515

**22.** G. Indiveri, B. Linares-Barranco, T. 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. SAÏGHI, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, K. Boahen, *Front. Neurosci.* **2011**, *5*, 73

**23.** G. Rachmuth, H. Z. Shouval, M. F. Bear, C.-S. Poon, *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*(49), E1266–E1274

**24.** J. Wang, G. Cauwenberghs, F. D. Broccard, *IEEE Trans. Biomed. Eng.* **2019**, 10.1109/TBME.2019.2948809

**25.** S. Mitra, S. Fusi, G. Indiveri, *IEEE Trans. Biomed. Circuits Syst.* **2009**, *3*(1), 32–42

**26.** M. R. Azghadi, O. Kavehei, S. Al-Sarawi, N. Iannella, D. Abbott, in *Proceedings of the 7th International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, **2011** S. 158–162

**27.** M. R. Azghadi, N. Iannella, S. Al-Sarawi, D. Abbott, *PLoS One* **2014**, *9*(2), e88326

**28.** G. Indiveri, E. Chicca, R. Douglas, *IEEE Trans. Neural Netw.* **2006**, *17*(1), 211–221

**29.** M. R. Azghadi, S. Al-Sarawi, N. Iannella, G. Indiveri, D. Abbott, *Proc. IEEE* **2014**

Author Manuscript

This article is protected by copyright. All rights reserved

**30.** H. Wang, R. Gerkin, D. Nauen, G. Bi, *Nat. Neurosci.* **2005**, *8*(2), 187–193

**31.** P. Sjöström, G. Turrigiano, S. Nelson, *Neuron* **2001**, *32*(6), 1149–1164

**32.** J. Pfister, W. Gerstner, *The Journal of Neuroscience* **2006**, *26*(38), 9673–9682

**33.** M. R. Azghadi, S. Al-Sarawi, D. Abbott, N. Iannella, *Neural Networks* **2013**, *45*, 70–82

**34.** S. Moradi, G. Indiveri, *IEEE Trans. Biomed. Circuits Syst.* **2014**, *8*(1), 98–107

**35.** C. Rossant, D. F. Goodman, B. Fontaine, J. Platkiewicz, A. K. Magnusson, R. Brette, *Front. Neurosci.* **2011**, *5*, 9

**36.** S. Green, J. B. Aimone, *Nat. Electron.* **2019**, *2*(3), 96

**37.** S. M. Sze, K. K. Ng, *Physics of semiconductor devices*, John wiley & sons, **2006**

**38.** Y.-F. Chang, B. Fowler, Y.-C. Chen, C.-Y. Lin, G. Xu, H.-C. Huang, J. Chen, S. Kim, Y. Li, J. C. Lee, *J. Mater. Chem. C* **2018**, *6*(47), 12788–12799

**39.** A. Mehonic, A. L. Shluger, D. Gao, I. Valov, E. Miranda, D. Ielmini, A. Bricalli, E. Ambrosi, C. Li, J. J. Yang, Q. Xia, A. J. Kenyon, *Adv. Mater.* **2018**, *30*(43), 1801187

**40.** T. Hickmott, *J. Appl. Phys.* **1962**, *33*(9), 2669–2682

**41.** J. Simmons, R. Verderber, *Proc. R. Soc. A* **1967**, *301*(1464), 77–102

This article is protected by copyright. All rights reserved

Author Manuscript

**42.** A. Mehonic, M. Munde, W. Ng, M. Buckwell, L. Montesi, M. Bosman, A. Shluger, A. Kenyon, *Microelectron. Eng.* **2017**, *178*, 98 – 103, special issue of Insulating Films on Semiconductors (INFOS 2017)

**43.** M. Munde, A. Mehonic, W. Ng, M. Buckwell, L. Montesi, M. Bosman, A. Shluger, A. Kenyon, *Sci. Rep.* **2017**, *7*(1), 9274

**44.** A. Mehonic, A. J. Kenyon, *Resistive Switching in Oxides*, S. 401–428, Springer International Publishing, Cham, **2015**

**45.** J. Yao, Z. Sun, L. Zhong, D. Natelson, J. M. Tour, *Nano Lett.* **2010**, *10*(10), 4105–4110, pMID: 20806916

**46.** Y.-F. Chang, B. Fowler, Y.-C. Chen, Y.-T. Chen, Y. Wang, F. Xue, F. Zhou, J. C. Lee, *J. Appl. Phys.* **2014**, *116*(4), 043708

**47.** G. Wang, Y. Yang, J.-H. Lee, V. Abramova, H. Fei, G. Ruan, E. L. Thomas, J. M. Tour, *Nano Lett.* **2014**, *14*(8), 4694–4699, pMID: 24992278

**48.** A. J. Kenyon, M. Singh Munde, W. H. Ng, M. Buckwell, D. Joksas, A. Mehonic, *Faraday Discuss.* **2019**, *213*, 151–163

**49.** G. Bi, M. Poo, *The Journal of Neuroscience* **1998**, *18*(24), 10464–10472

**50.** K. Zarudnyi, A. Mehonic, L. Montesi, M. Buckwell, S. Hudziak, A. J. Kenyon, *Front. Neurosci.*

This article is protected by copyright. All rights reserved

Author Manuscript

**2018**, *12*, 57

**51.** A. Mehonic, D. Joksas, W. H. Ng, M. Buckwell, A. J. Kenyon, *Front. Neurosci.* **2019**, *13*

**52.** S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, W. Lu, *Nano Lett.* **2010**, *10*(4), 1297–1301

**53.** Y.-F. Chang, B. Fowler, Y.-C. Chen, F. Zhou, C.-H. Pan, T.-C. Chang, J. C. Lee, *Sci. Rep.* **2016**, *6*, 21268

**54.** A. L. Hodgkin, A. F. Huxley, *The J. physiology* **1952**, *117*(4), 500–544

**55.** A. Mehonic, A. J. Kenyon, *Front. Neurosci.* **2016**, *10*, 57

**56.** C.-Y. Lin, P.-H. Chen, T.-C. Chang, K.-C. Chang, S.-D. Zhang, T.-M. Tsai, C.-H. Pan, M.-C. Chen, Y.-T. Su, Y.-T. Tseng, Y.-F. Chang, Y.-C. Chen, H.-C. Huang, S. M. Sze, *Nanoscale* **2017**, *9*, 8586–8590

**57.** Y.-F. Chang, B. Fowler, F. Zhou, Y.-C. Chen, J. C. Lee, *Appl. Phys. Lett.* **2016**, *108*(3), 033504

**58.** C. Koch, I. Segev, *Nat. Neurosci.* **2000**, *3*(11s), 1171

**59.** L. Gao, P.-Y. Chen, S. Yu, *Appl. Phys. Lett.* **2017**, *111*(10), 103503

**60.** C. Morris, H. Lecar, *Biophys. J.* **1981**, *35*(1), 193–213

Author Manuscript

This article is protected by copyright. All rights reserved

**61.** M. D. Pickett, G. Medeiros-Ribeiro, R. S. Williams, *Nat. Mater.* **2013**, *12*(2), 114

**62.** W. Yi, K. K. Tsang, S. K. Lam, X. Bai, J. A. Crowell, E. A. Flores, *Nat. Commun.* **2018**, *9*(1), 4661

**63.** Z. Wang, S. Joshi, S. Savel'ev, W. Song, R. Midya, Y. Li, M. Rao, P. Yan, S. Asapu, Y. Zhuo, et al., *Nat. Electron.* **2018**, *1*(2), 137

**64.** K. Yue, Y. Liu, R. K. Lake, A. C. Parker, *Sci. Adv.* **2019**, *5*(4), eaau8170

**65.** A. Amirsoleimani, M. Ahmadi, A. Ahmadi, M. Boukadoum, in *IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, **2016** S. 81–84

**66.** A. Amirsoleimani, M. Ahmadi, A. Ahmadi, in *International Joint Conference on Neural Networks (IJCNN)*, **2017** S. 3409–3414

**67.** Q. Gao, A. Huang, Q. Hu, X. Zhang, Y. Chi, R. Li, Y. Ji, X. Chen, R. Zhao, M. Wang, H. Shi, M. Wang, Y. Cui, Z. Xiao, P. K. Chu, *ACS Appl. Mater. Interfaces* **2019**

**68.** W. Chen, R. Fang, M. B. Balaban, W. Yu, Y. Gonzalez-Velo, H. J. Barnaby, M. N. Kozicki, *Nanotechnology* **2016**, *27*(25), 255202

**69.** Z. Wang, M. Yin, T. Zhang, Y. Cai, Y. Wang, Y. Yang, R. Huang, *Nanoscale* **2016**, *8*(29), 14015–14022

**70.** W. Román Acevedo, C. Acha, M. Sánchez, P. Levy, D. Rubi, *Appl. Phys. Lett.* **2017**, *110*(5), 053501

This article is protected by copyright. All rights reserved

Author Manuscript

**71.** M. R. Azghadi, S. Moradi, G. Indiveri, in *2013 IEEE 11th International New Circuits and Systems Conference (NEWCAS)*, IEEE, **2013** S. 1–4

**72.** F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, W. D. Lu, *Nat. Electron.* **2019**, *2*(7), 290–299

**73.** W. Jiang, B. Xie, C.-C. Liu, Y. Shi, *Nat. Electron.* **2019**, *2*(9), 376–377

**74.** E. Bienenstock, L. Cooper, P. Munro, *The J. Neurosci.* **1982**, *2*(1), 32

**75.** A. Kirkwood, M. Rioult, M. Bear, *Nature* **1996**, *381*(6582), 526–528

**76.** C. Lammie, T. J. Hamilton, A. van Schaik, M. R. Azghadi, *IEEE Trans. Circuits Syst. I* **2018**, *66*(4), 1558–1570

**77.** M. Heidarpur, A. Ahmadi, M. Ahmadi, M. Rahimi Azghadi, *IEEE Trans. Circuits Syst. I* **2019**, *66*(7), 2651–2661

**78.** 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*(6197), 668–673

**79.** D. B. Fasnacht, A. M. Whatley, G. Indiveri, in *2008 IEEE International Symposium on Circuits and Systems*, IEEE, **2008** S. 648–651

Author Manuscript

This article is protected by copyright. All rights reserved

**80.** M. Payvand, M. V. Nair, L. K. Müller, G. Indiveri, *Faraday Discuss.* **2019**, *213*, 487–510

**81.** M. A. Zidan, J. P. Strachan, W. D. Lu, *Nat. Electron.* **2018**, *1*(1), 22–29

**82.** C. Sung, H. Hwang, I. K. Yoo, *J. Appl. Phys.* **2018**, *124*(15), 151903

**83.** O. Krestinskaya, A. P. James, L. O. Chua, *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*(1), 4–23

**84.** A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, V. Srikumar, *ACM SIGARCH Comput. Archit. News* **2016**, *44*(3), 14–26

**85.** S. Yu, P.-Y. Chen, *IEEE Solid-State Circuits Mag.* **2016**, *8*(2), 43–56

**86.** S. Kim, H. Kim, S. Hwang, M.-H. Kim, Y.-F. Chang, B.-G. Park, *ACS Appl. Mater. Interfaces* **2017**, *9*(46), 40420–40427

**87.** O. Golonzka, U. Arslan, P. Bai, M. Bohr, O. Baykan, Y. Chang, A. Chaudhari, A. Chen, N. Das, C. English, et al., in *2019 Symposium on VLSI Technology*, IEEE, **2019** S. T230–T231

**88.** S. Deswal, A. Kumar, A. Kumar, *AIP Adv.* **2019**, *9*(9), 095022

**89.** L. Ji, Y.-F. Chang, B. Fowler, Y.-C. Chen, T.-M. Tsai, K.-C. Chang, M.-C. Chen, T.-C. Chang, S. M. Sze, E. T. Yu, et al., *Nano Lett.* **2013**, *14*(2), 813–818

**90.** Y. Deng, P. Huang, B. Chen, X. Yang, B. Gao, J. Wang, L. Zeng, G. Du, J. Kang, X. Liu, *IEEE Trans. Electron Devices* **2012**, *60*(2), 719–726

This article is protected by copyright. All rights reserved

Author Manuscript

**91.** Y.-C. Chen, C.-C. Lin, S.-T. Hu, C.-Y. Lin, B. Fowler, J. Lee, *Sci. Rep.* **2019**, *9*(1), 1–6

**92.** Y.-C. Chen, C.-Y. Lin, H.-C. Huang, S. Kim, B. Fowler, Y.-F. Chang, X. Wu, G. Xu, T.-C. Chang, J. C. Lee, *J. Phys. D Appl. Phys.* **2018**, *51*(19)

**93.** Y.-C. Chen, S.-T. Hu, C.-Y. Lin, B. Fowler, H.-C. Huang, C.-C. Lin, S. Kim, Y.-F. Chang, J. C. Lee, *Nanoscale* **2018**, *10*(33), 15608–15614

**94.** T. Prodromakis, C. Toumazou, in *2010 17th IEEE International Conference on Electronics, Circuits and Systems*, IEEE, **2010** S. 934–937

**95.** J. K. Eshraghian, S. Baek, J.-H. Kim, N. Iannella, K. Cho, Y.-S. Goo, H. H. Iu, S.-M. Kang, K. Eshraghian, *Int. J. Neural Syst.* **2018**, *28*(7), 1850004

**96.** K. Cho, S. Baek, S.-W. Cho, J.-H. Kim, Y.-S. Goo, J. K. Eshraghian, N. Iannella, K. Eshraghian, *IEEE Sensors J.* **2016**, *16*(15), 5856–5866

**97.** L. Chua, *Nanotechnology* **2013**, *24*(38), 383001

**98.** K. Usha, P. Subha, *Biosystems* **2019**, *178*, 1–9

**99.** H.-M. Huang, R. Yang, Z.-H. Tan, H.-K. He, W. Zhou, J. Xiong, X. Guo, *Adv. Mater.* **2019**, *31*(3), 1803849

**100.** M. V. Nair, L. K. Muller, G. Indiveri, *Nano Futures* **2017**, *1*(3), 035003

This article is protected by copyright. All rights reserved

Author Manuscript

**101.** I. Boybat, M. Le Gallo, S. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian, E. Eleftheriou, *Nat. Commun.* **2018**, *9*(1), 2514

**102.** D. Querlioz, O. Bichler, C. Gamrat, in *The 2011 International Joint Conference on Neural Networks*, IEEE, **2011** S. 1775–1781

**103.** H. Wu, M. Zhao, Y. Liu, P. Yao, Y. Xi, X. Li, W. Wu, Q. Zhang, J. Tang, B. Gao, H. Qian, in *2019 IEEE International Reliability Physics Symposium (IRPS)*, **2019** S. 1–4

**104.** Z. Wei, G. Yan, *Acta Opt. Sin.* **2011**, *31*(10), 1005002

**105.** P. Huang, Y. Xiang, Y. Zhao, C. Liu, B. Gao, H. Wu, H. Qian, X. Liu, J. Kang, in *2018 IEEE International Electron Devices Meeting (IEDM)*, IEEE, **2018** S. 40–4

**106.** J. K. Eshraghian, C. Lammie, M. R. Azghadi, in *IEEE International Symposium on Circuits and Systems, accepted for publication*, **2020**

**107.** C. Du, F. Cai, M. A. Zidan, W. Ma, S. H. Lee, W. D. Lu, *Nat. Commun.* **2017**, *8*(1), 2204

**108.** J. Moon, W. Ma, J. H. Shin, F. Cai, C. Du, S. H. Lee, W. D. Lu, *Nat. Electron.* **2019**, *2*(10), 480–487

Author Manuscript

This article is protected by copyright. All rights reserved

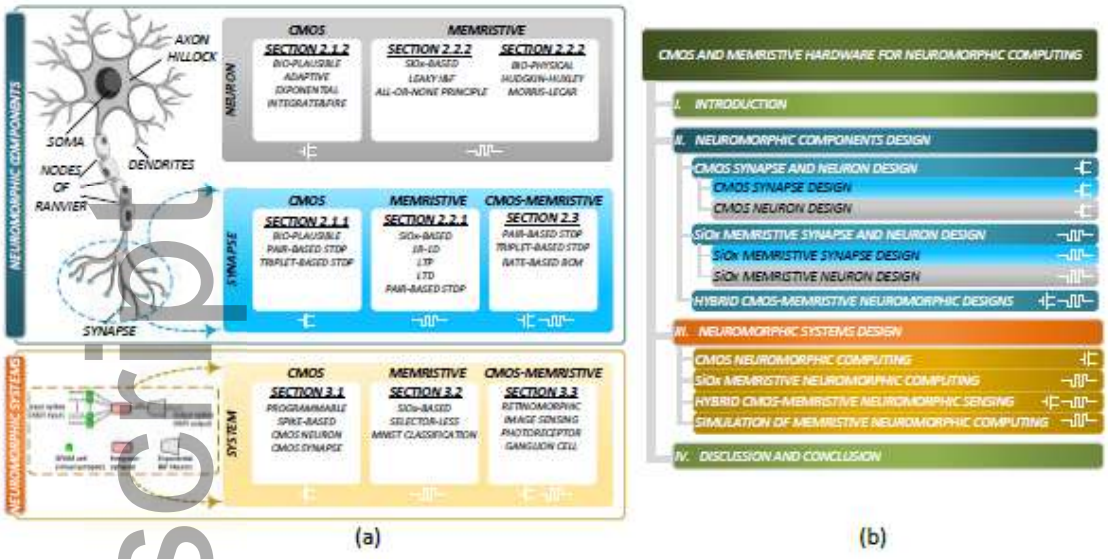



**Figure 1.** The structure of the paper at a glance. (a) The paper is composed of discussions on CMOS, Memristive, and hybrid CMOS-Memristive neuromorphic components and systems. (b) The tree-structure of the paper outline is shown. (Neuron figure designed by brgfx/Freepik)

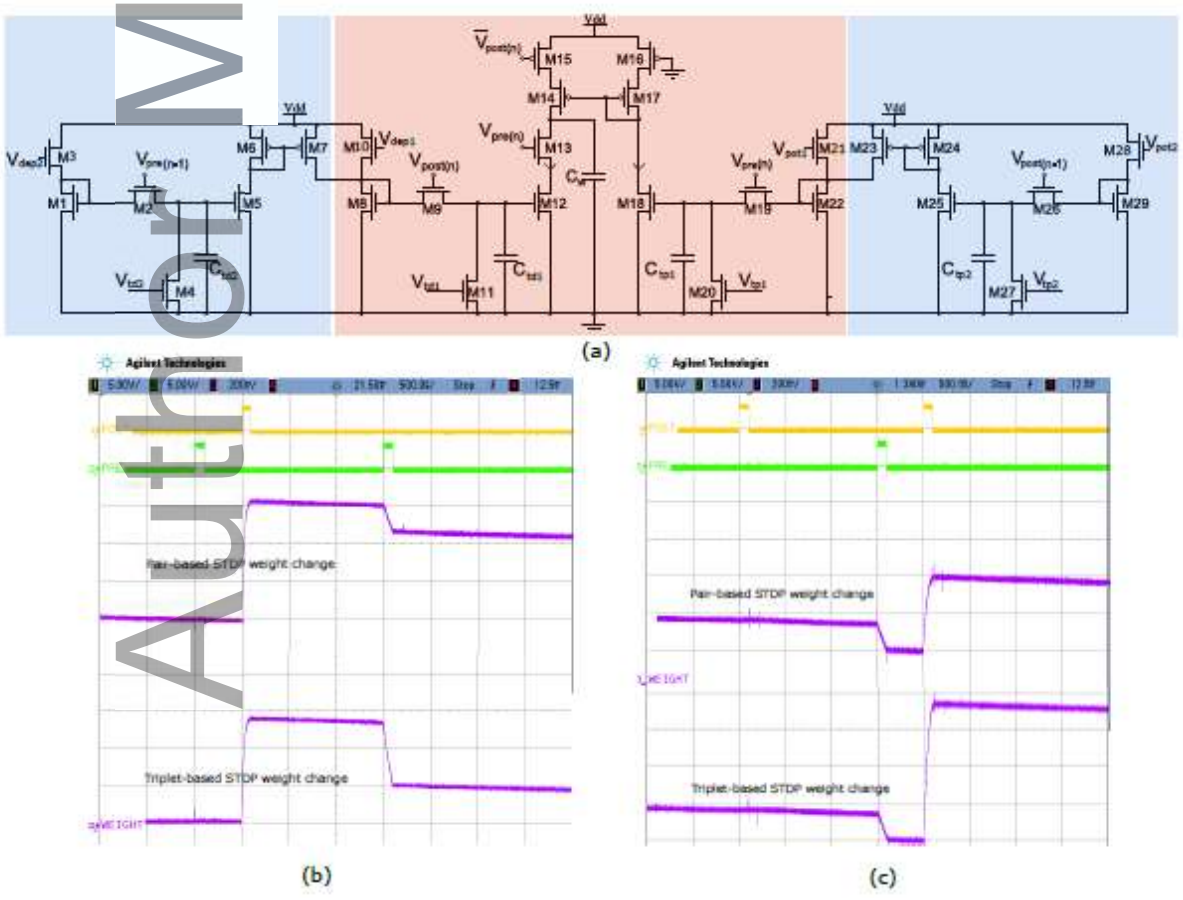



**Figure 2.** A synapse circuit implemented in CMOS to realize triplet-based STDP (a) can modify

This article is protected by copyright. All rights reserved

Author Manuscript

synaptic weight based on pair-based STDP (top traces in (b-c)) when only its pair-based (red) part is activated, while it shows triplet-based weight change dynamics (bottom traces in (b-c)) when all its parts, i.e. the triplet depression (left blue) and triplet potentiation (right blue) work in conjunction with the pair-based (red) part. Note that, a pre-post-pre spike triplet (b) can result in higher synaptic depression in triplet STDP interactions (bottom trace), while it shows a lower depression (top trace) in the pair-based spike interaction, due to considering only the post-pre pair, and not the pre-post-pre triplet. Similarly, in the case of post-pre-post spiking in (c), higher potentiation is expected when considering the triplet combination (bottom trace), while lower potentiation is elicited when only a pre-post pair is considered (top trace).

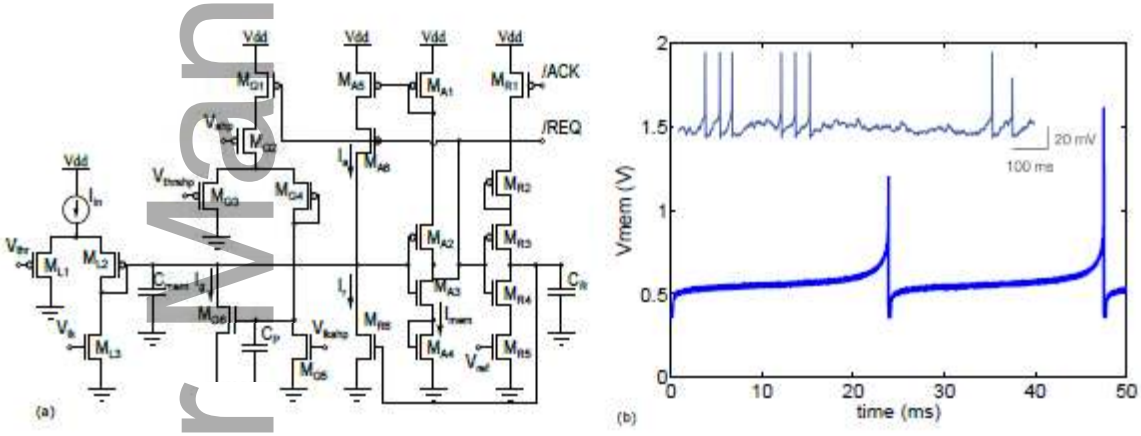



**Figure 3.** (a) A neuron circuit implemented in CMOS (Adapted with permission.[34] 2014, IEEE.) is capable of reproducing (b) the spiking behavior similar to that of biological neurons (Adapted with permission.[18] 2015, ACM). The inset shows the spiking behavior of neurons measured in response to somatically injected currents, while the spiking behavior of the neuron is also in result of increasing $I_{in}$.

Author Manuscript

This article is protected by copyright. All rights reserved

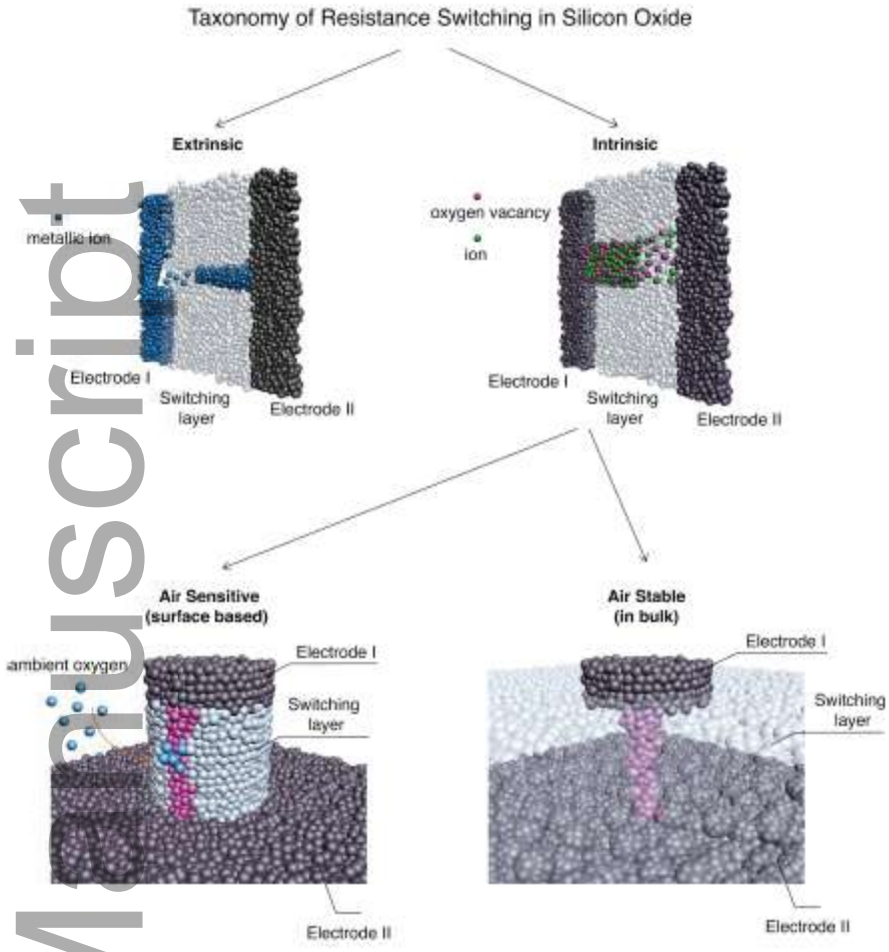



**Figure 4.** Taxonomy of resistance switching in silicon oxide. Upper LHS: Schematic of extrinsic resistance switching. Upper RHS: Schematic of intrinsic resistance switching. Lower LHS: Air-sensitive resistance switching: typically, electroforming and resistance switching occur only in devices with an exposed oxide surface and not in bulk devices. This is attributed to re-oxidation of surface-based silicon filaments by an oxidizing ambient. Lower RHS: Air-stable resistance switching. This type of switching occurs in ambient (oxidizing) conditions and is defined by the microstructure of the oxide material. Switching voltages are typically lower for air-sensitive switching. (Reproduced with permission.[39] 2018, John Wiley and Sons.)

Author Manuscript

This article is protected by copyright. All rights reserved

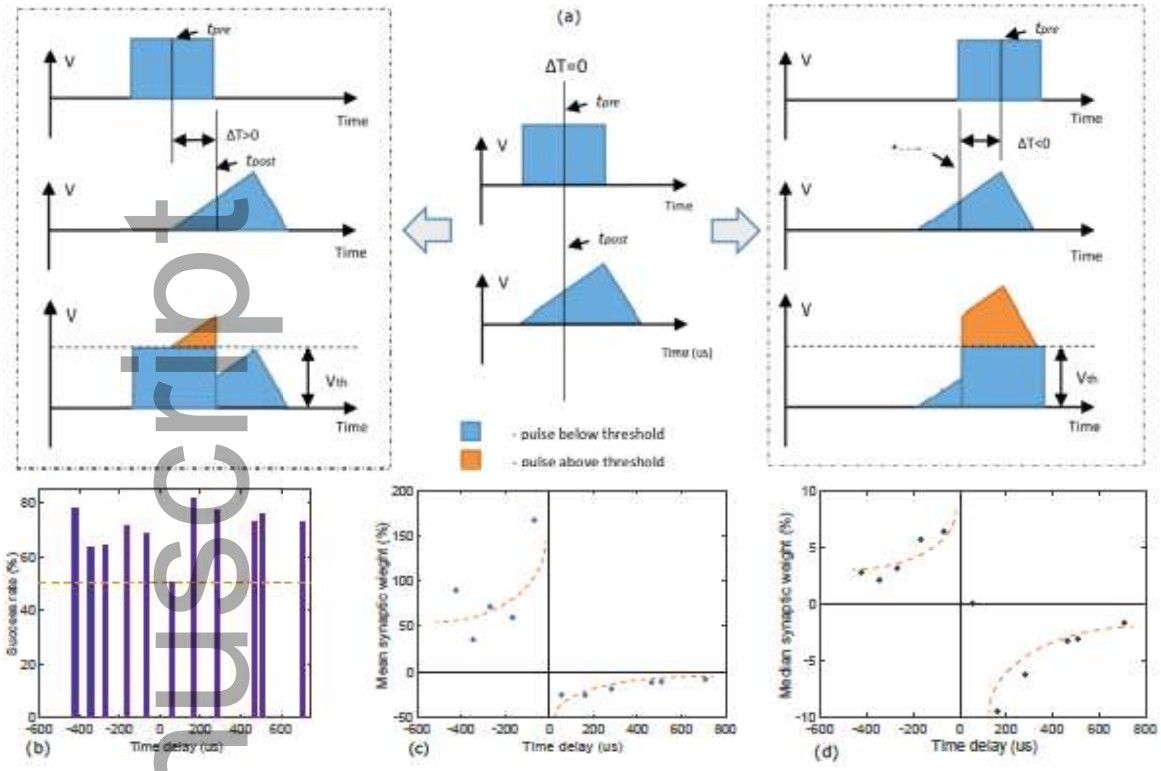



**Figure 5.** (a) Illustration of the experimental set up to test STDP-like behavior: Non-identical pre- and post- voltage pulses and the resulting overlapping voltage across the device. (b) The plot demonstrates successful device operations (conductance increase or decrease in dependence on time delays). Plots demonstrate STDP-like behavior for (c) mean and (d) median change in device conductance. Dotted lines serve as a visual guide. (Adapted under the terms of the CC BY license.[50] 2018, Zarudnyi et al.)

Author Manuscript

This article is protected by copyright. All rights reserved

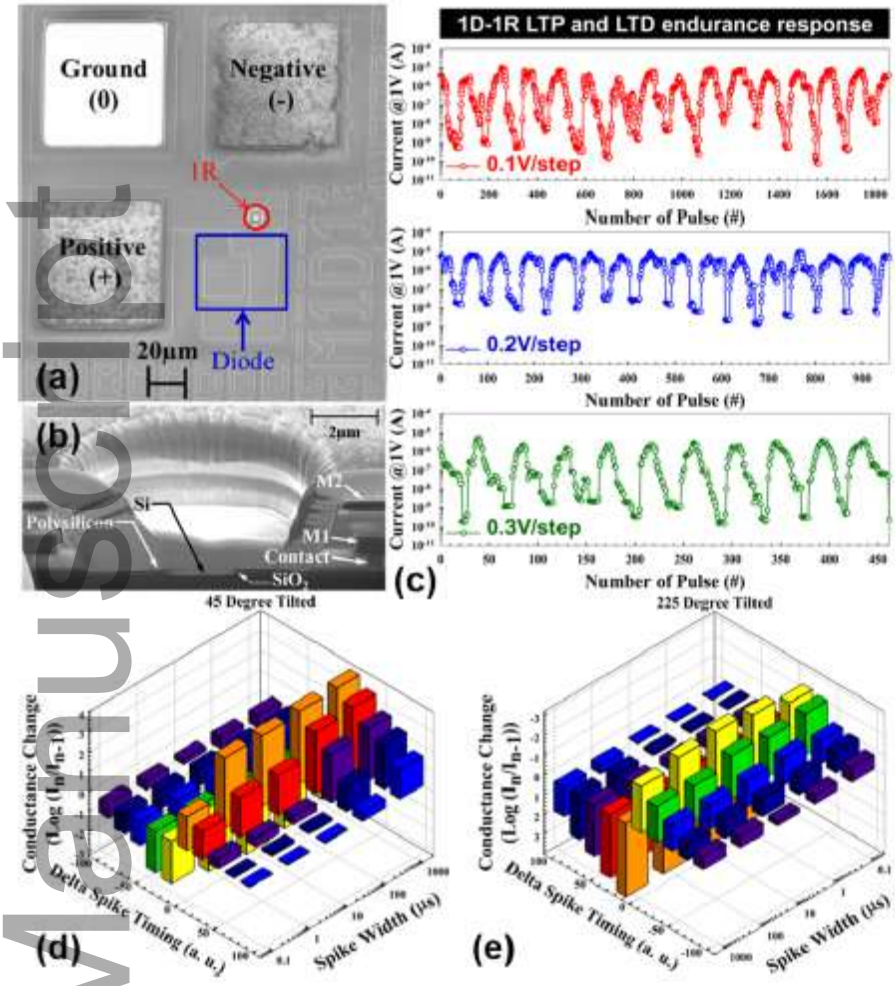



**Figure 6.** SiOx-based 1R-1D synaptic device. The device structure is shown in SEM images shown in (a) for top view, and (b) cross sectional view. The device shows great controllability to increase and decrease its conductance in response to pulse amplitude modulation as shown in (c). The device is used to demonstrate 10 different conductance change behavior as shown in (d) in result of an STDP experiment, where positive and negative spike timing difference, as well as spike widths are used. Here, (e) shows the same data as (d) but in the reversed order. (Parts a, b, d, and e are reproduced under the terms of the CC BY license.[53] 2016, Chang et al.)

Author Manuscript

This article is protected by copyright. All rights reserved

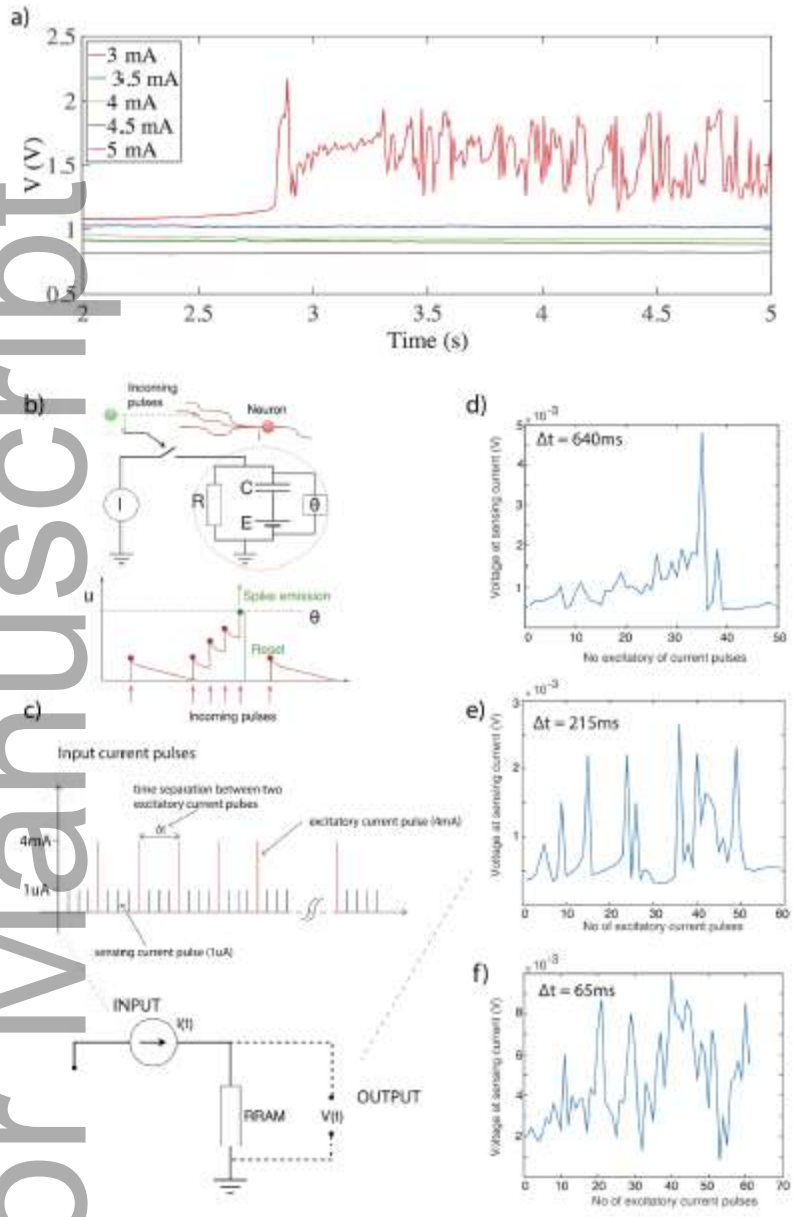



**Figure 7.** (a) Voltage transients (spikes) do not occur until the constant current of 5 mA is applied (current threshold). (b) Schematic of the leaky integrate-and-fire neuronal model. The lower figure illustrates the integration of input current pulses over time. Theta is the voltage threshold for spiking. (c) Excitatory current pulses (4 mA) are applied to the device, and small current pulses (1 $\mu$ A) are used for sensing the voltage across the device. Device voltage response is shown for pulse time separations of (d) 640 ms, (e) 215 ms, and (f) 65 ms. These demonstrate that spiking occurs for the smaller number of pulses if the separation between input current pulses is shorter. (Adapted under the terms of the CC BY license.[55] 2016, Mehonic and Kenyon.)

This article is protected by copyright. All rights reserved

Author Manuscript

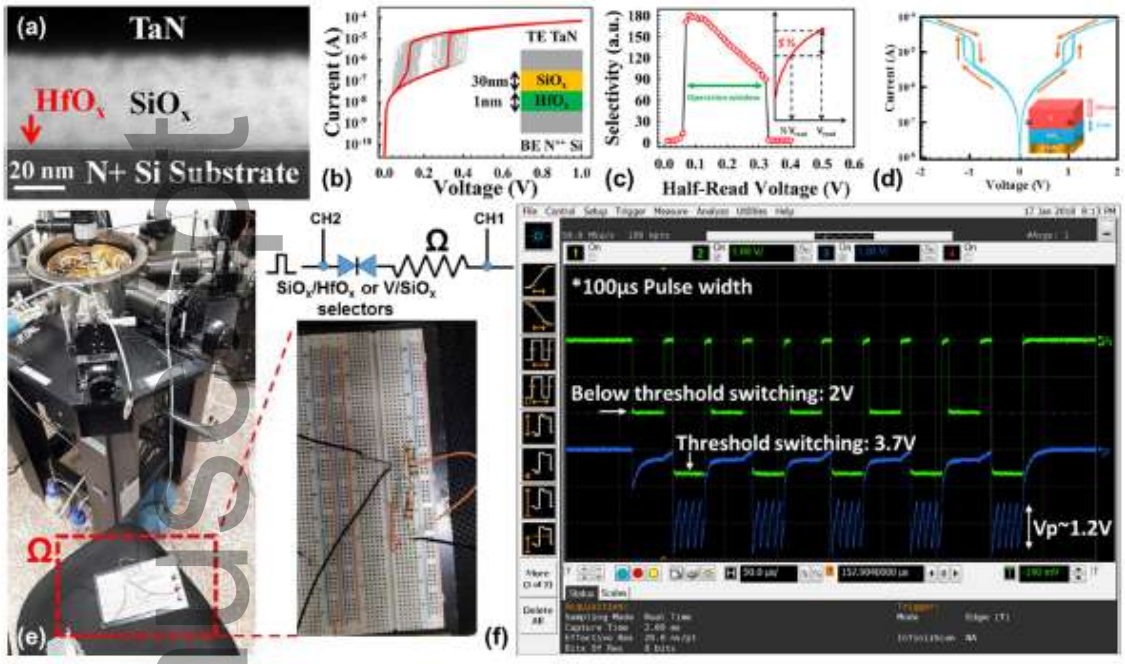



**Figure 8.** SiOx-based 1R-1R neuron device. (a) TEM cross-section image of (Metal Insulator Semiconductor) MIS structure: TaN/30 nm-SiOx/1 nm-HfOx/N+ Si substrate. (b) 100 times I–V threshold switching of MIS device in air. (c) Threshold switching in the air for selectivity and nonlinearity of the stacking bilayer structure shown in (a). (d) Typical I–V threshold switching of vanadium electrode/6 nm-SiOx/TiN structure. (e) Experimental setup for a simple neuron circuit by SiOx-based threshold switching selector. (f) The screenshot of measurement instrument (Agilent DSO9254A) and measured voltage results in CH1 (green line) and CH2 (blue line). Pulses with 100 $\mu$ s width are used here. When CH1 is below the firing threshold (e.g. at 2V), there is no firing observed; while if threshold switching happened (e.g. input pulse amplitude > 3.7V), the neuron firing begins on CH2 (damping Vp 1.2V). (Parts a and b are adapted with permission.[38] 2018, The Royal Society of Chemistry.)

Author Manuscript

This article is protected by copyright. All rights reserved

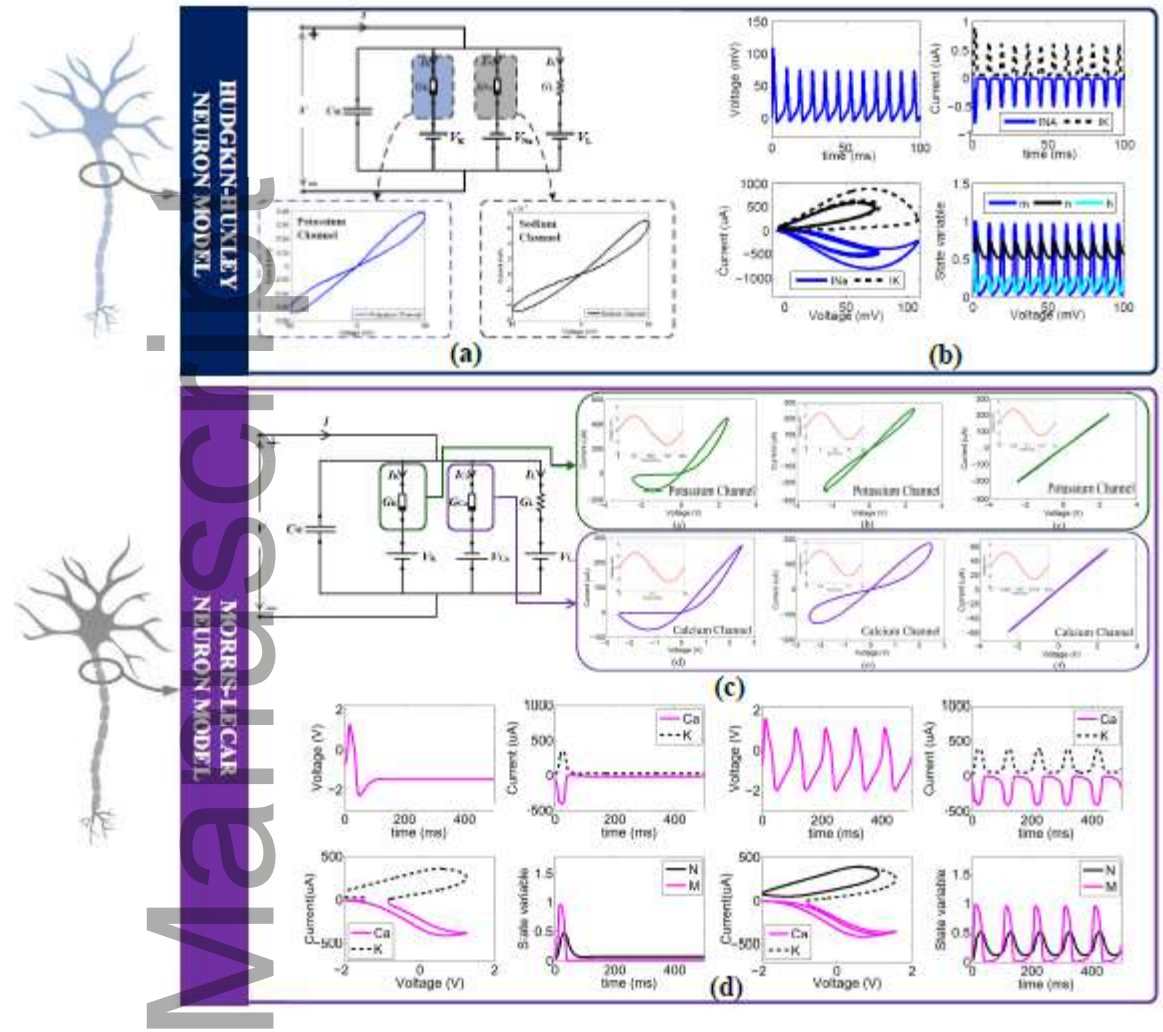
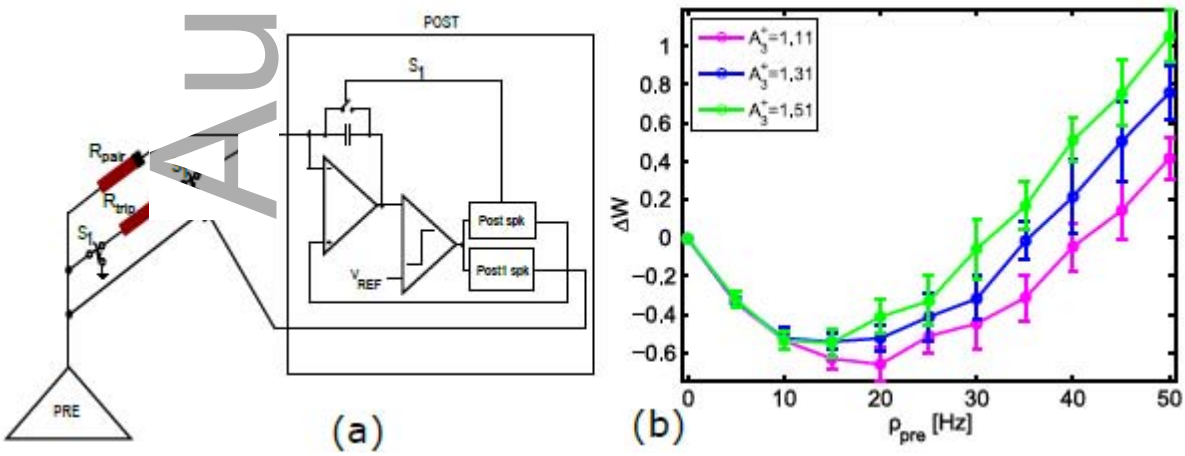



**Figure 9.** Bio-inspired memristive neuron circuit diagrams, memristive channel characteristics and bifurcation behaviors at different frequencies, for HH (top panel) and ML (bottom panel) neuron models. (Adapted with permission.[65, 66] 2016 and 2017, IEEE.)




This article is protected by copyright. All rights reserved

**Figure 10.** (a) A hybrid CMOS-memristor neuromorphic synapse connected to CMOS pre- and post-synaptic devices (Adapted with permission.[15] 2017, IEEE). The crossed square in the synapse circuit is representative of a multiplication/rectification circuit. In addition, $S_1$ is a digital signal that activates two switches in the synapse and one in the neuron, to implement the correct timings required for the STDP rule[15]. (b) The hybrid synapse is able to reproduce a BCM-like weight modification behavior[74, 75], by modifying $A_3^+$, a triplet STDP potentiation parameter, which dictates the amplitude of potentiation due to the interaction between a post-synaptic spike, with its two immediate previous post-synaptic and pre-synaptic spikes[32].

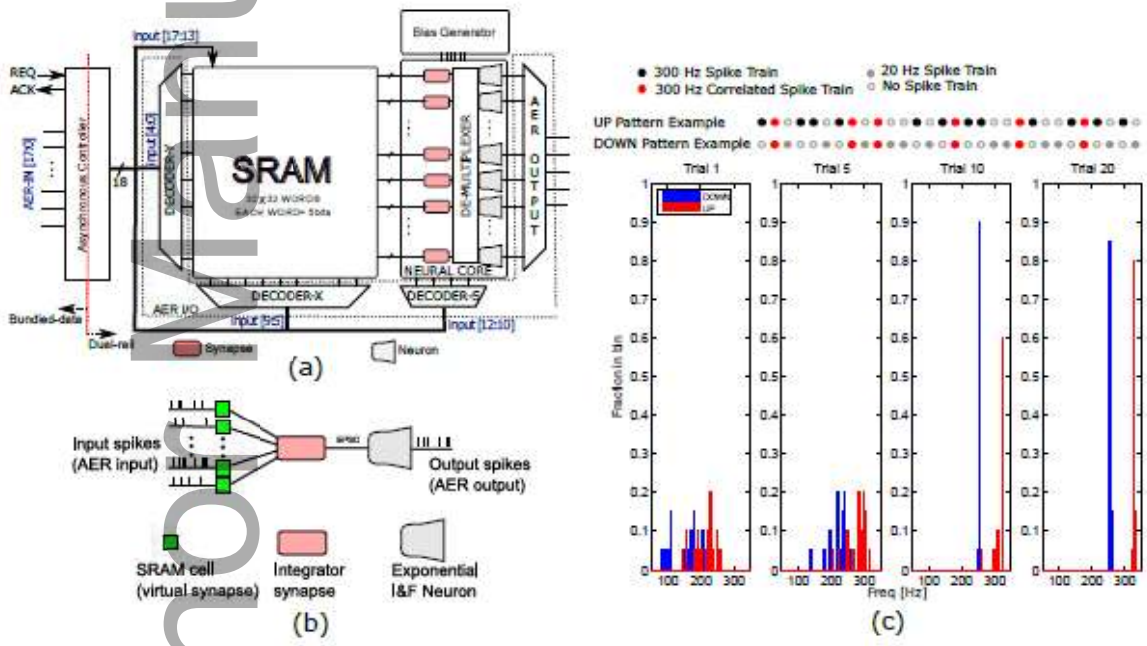



**Figure 11.** A programmable neuromorphic chip shown in (a) implements a simple perceptron network shown in (b) to learn unsupervised to detect 20% correlated rate-based patterns as depicted in (c). (Adapted with permission.[34, 18] 2014 and 2015, IEEE and ACM.)

Author Manuscript

This article is protected by copyright. All rights reserved

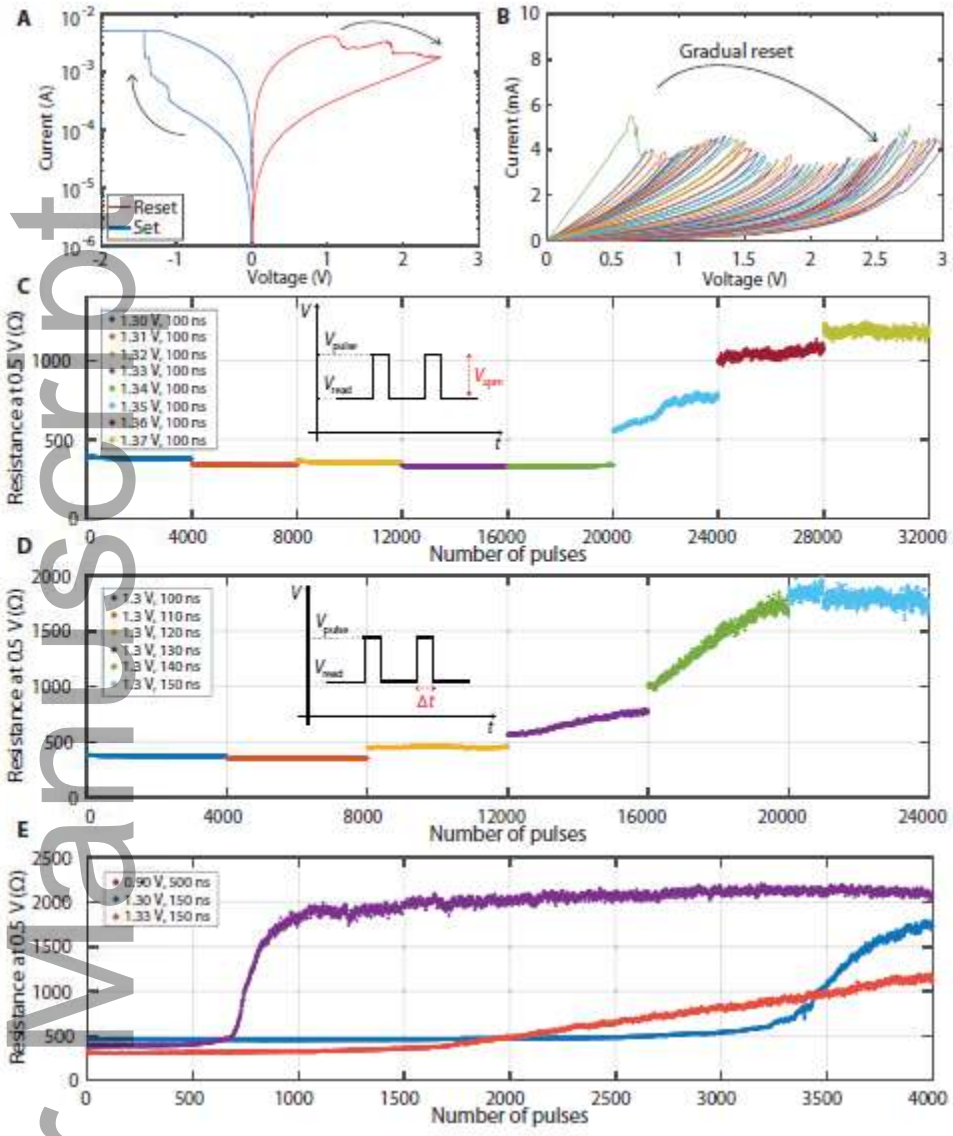



**Figure 12.** (A) IV curves demonstrate a typical bipolar resistance switching. (B) Adjustment of reset voltage shows a gradual reset process achieving multiple stable resistance states. Dynamics of resistance increase (LTD) is highly dependent on both pulse amplitude and pulse width. This is demonstrated in figures (C) and (D), receptively. (E) Three very different LTD curves are obtained from the same RRAM device by changing the pulse amplitude and width. (Reused under the terms of the CC BY license.[51] 2019, Mehonic et al.)

Author Manuscript

This article is protected by copyright. All rights reserved

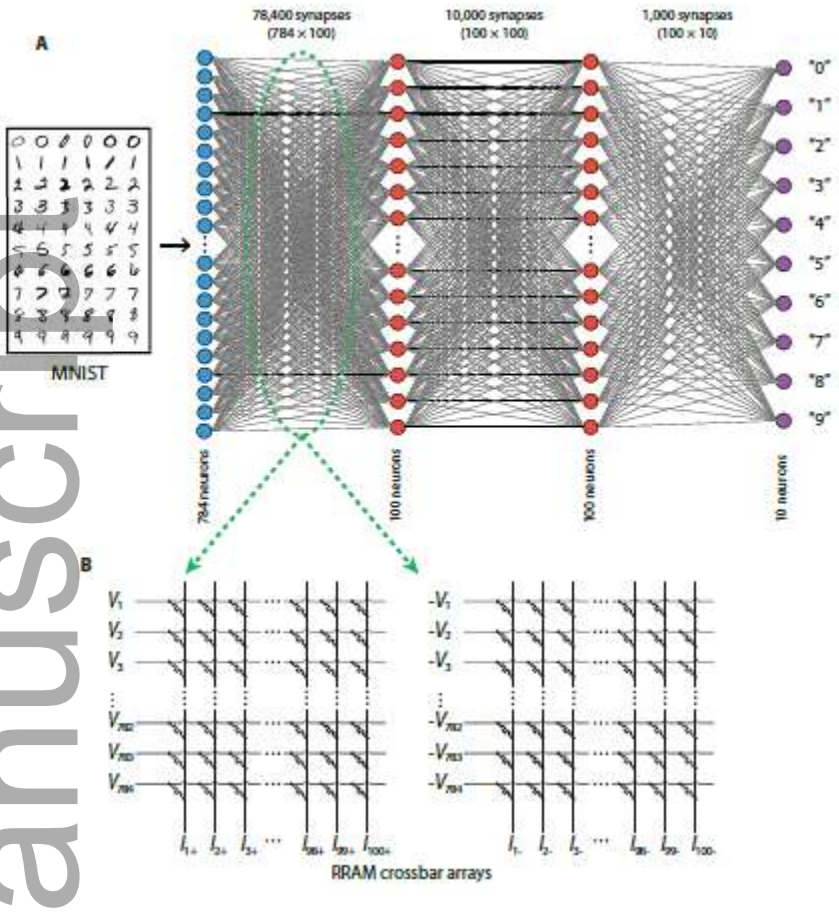



**Figure 13.** (A) The structure of the implemented ANN for MNIST classification task. (B) The network weights at each layer are mapped to a two-xbar structure. Here, the network has two hidden layers. (Reused under the terms of the CC BY license.[51] 2019, Mehonic et al.)

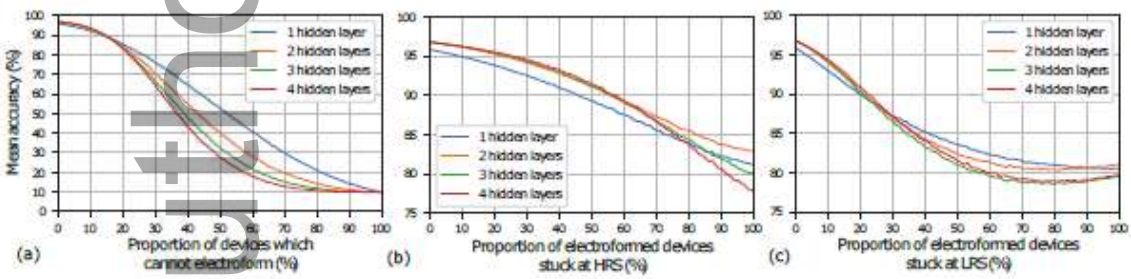



**Figure 14.** The decrease of inference accuracy of RRAM-based ANNs from (a) the effect of un-electroformed devices, (b) devices failed at HRS, and (c) devices failed at LRS. (Adapted under the terms of the CC BY license.[51] 2019, Mehonic et al.)

This article is protected by copyright. All rights reserved

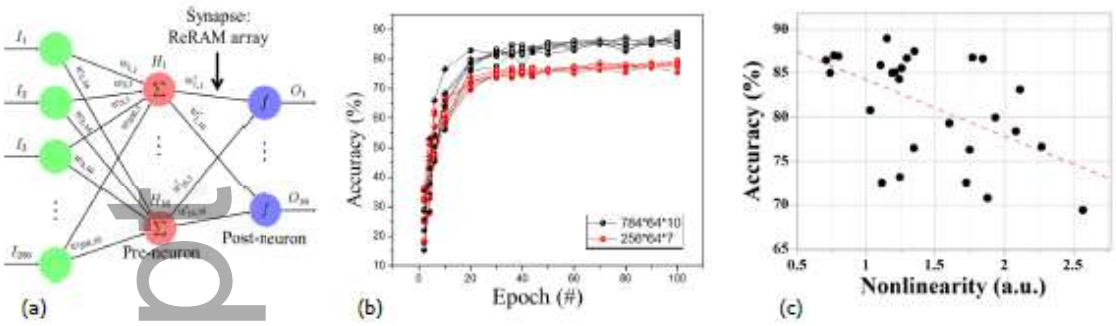



**Figure 15.** (a) The structure of the simulated neural network is shown. It learns through back-propagation algorithm and supervised training. (b) The actual network sizes are 256 x 64 x 10 and 256 x 64 x 7, which achieved different accuracy levels after 100 learning epochs for various device-to-device variation levels (each curve), averaged around 89% and 80%, for the two network sizes for MNIST and TP classification. Note that the convergence speed was similar in both network sizes and for various device variation levels. (c) Correlation between LTP and LTD nonlinearity effect and the simulated accuracy is shown.

Author Manuscript

This article is protected by copyright. All rights reserved

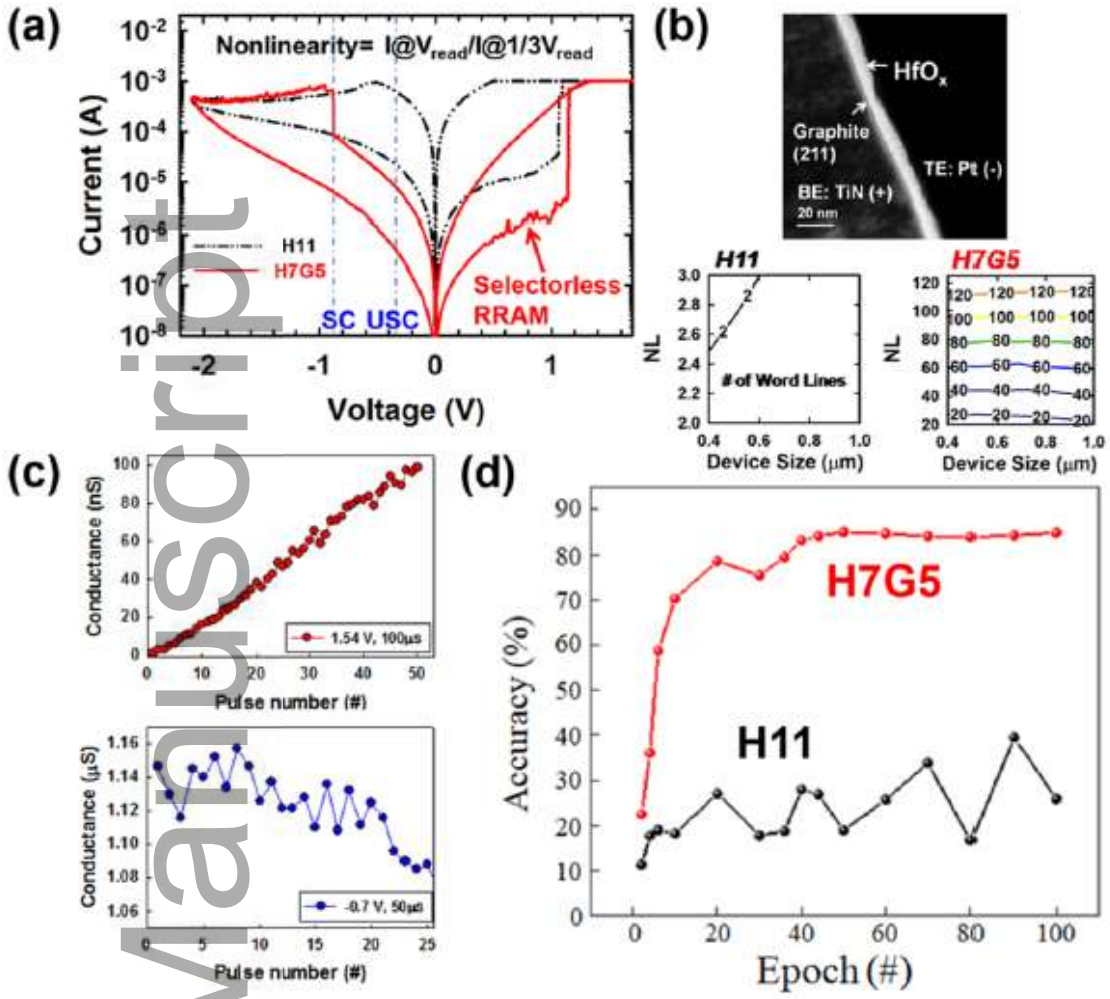



**Figure 16.** (a) I-V characteristics of bipolar resistive switching operation in HfOx single layer memristor (H11) and selectorless HfOx (7 nm)/graphite (5 nm) stacked memristor (H7G5) (Adapted under the terms of the CC BY license.[91] 2019, Chen et al.) (b) TEM image of a sample H7G5 stacked device is shown at the top (Adapted with permission.[93] 2019, The Royal Society of Chemistry.), and crossbar array size calculation based on measured nonlinearity and memory window with various device sizes (0.4, 0.6, 0.8, 1 um) on H11 and H7G5 devices are shown at the bottom. (c) The LTP/LTD behaviors using identical pulses are shown for H7G5 devices. (d) Neural network simulation accuracy using back-propagation algorithm and supervised training is shown. The accuracy is around 85% for H7G5 and 30% for H11. A similar network to that shown in Figure 15(a) was used.

Author Manuscript

This article is protected by copyright. All rights reserved

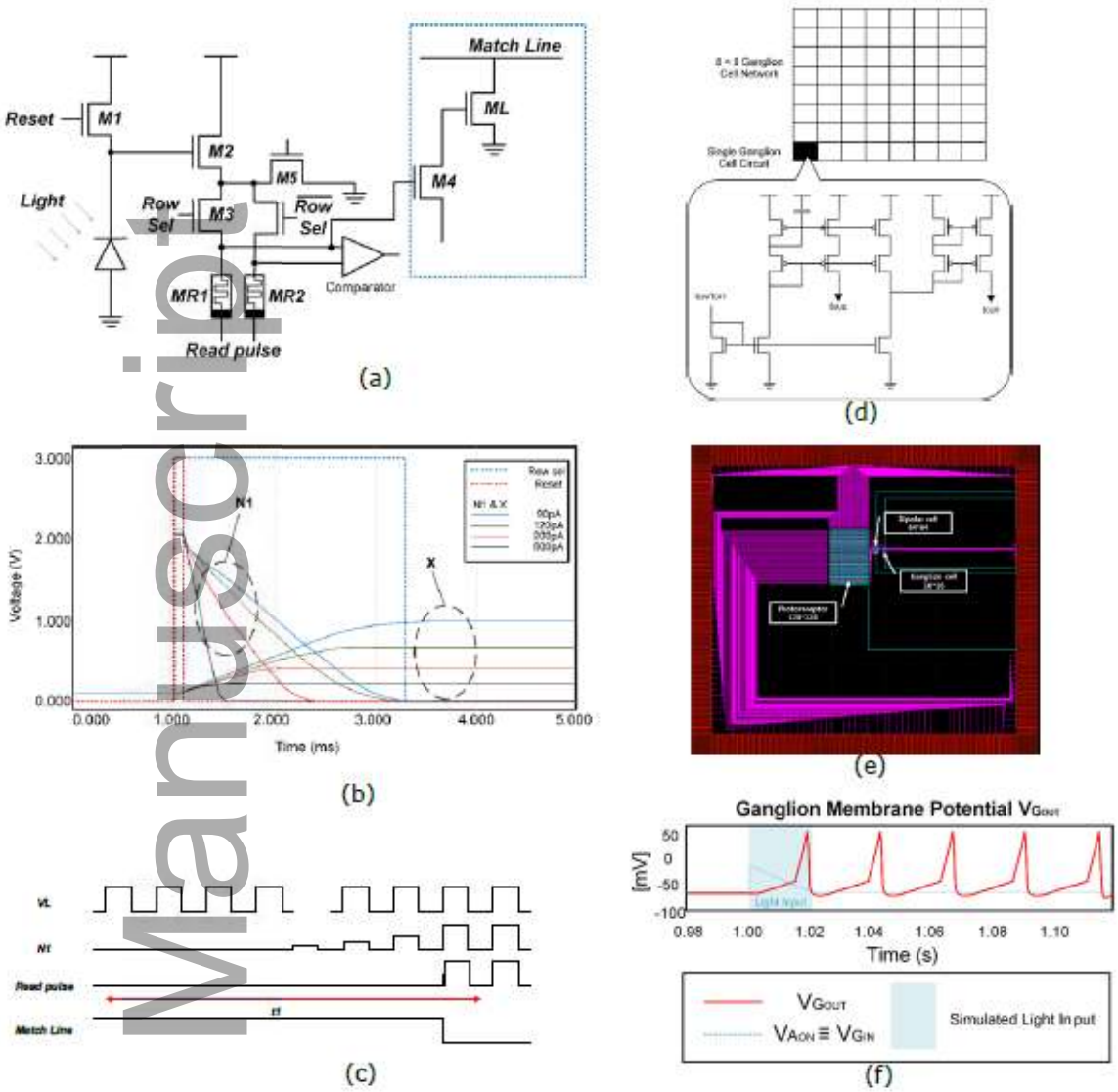



**Figure 17.** (a) The schematic structure of the CMOS-RRAM photoreceptor. (b) Photoreceptor cell output and state response. (c) Photoreceptor cell pulsing. (d) Circuit schematic of a ganglion cell-inspired subthreshold CMOS amplitude to frequency converter. The ganglion cell circuit accepts current as input from the previous image sensing stage, and will only generate an output where there has been a change in intensity, to achieve retina-like power optimization. (e) Retinal Network CMOS Chip Layout. The full flow of photocurrent commencing at a 128 x 128 array of photoreceptor cells, which are averaged to a 64 x 64 bipolar cell network to produce a graded action potential, and are subsequently converged to a 16 x 16 retinal ganglion cell network for fast spike generation. (f) Experimental results of the RRAM-CMOS retinomorphic architecture. (Parts e and f are reused with permission.[16] 2018, IEEE)

This article is protected by copyright. All rights reserved

Author Manuscript

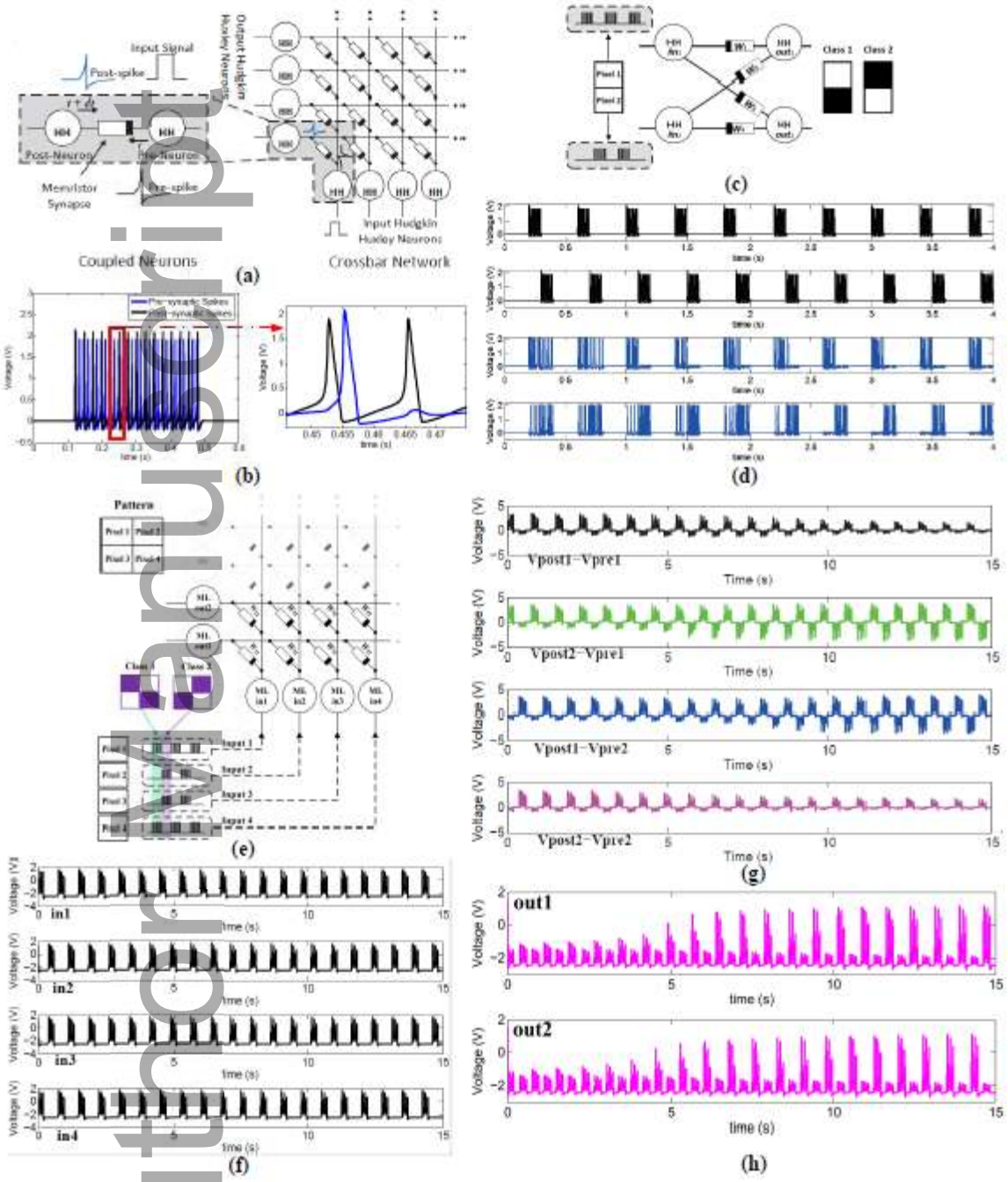



**Figure 18.** Simple patterns are classified by HH and ML memristive neurons, connected through crossbar structures of memristive synapses, following STDP learning, through overlapped input spikes. (Adapted with permission.[65, 66] 2016 and 2017, IEEE.)

This article is protected by copyright. All rights reserved

**Dr. Yao-Feng Chang** is a Memory Engineer at Intel. He received his B.S., M.S., and Ph.D. degrees in electrical engineering from National Sun Yat-sen University in 2007, in electronics engineering from National Chiao Tung University in 2009, and in the Department of Electrical and Computer Engineering from The University of Texas at Austin in 2015, respectively. His research topics include $SiO_x$- and $SiN_x$-Based ReRAM fabrication and characterization; Stateful Boolean logic and synaptic Post-Moore computing, and bio-inspired biomimetic systems in electronic devices; and Selector integration for large-scale low power array design.

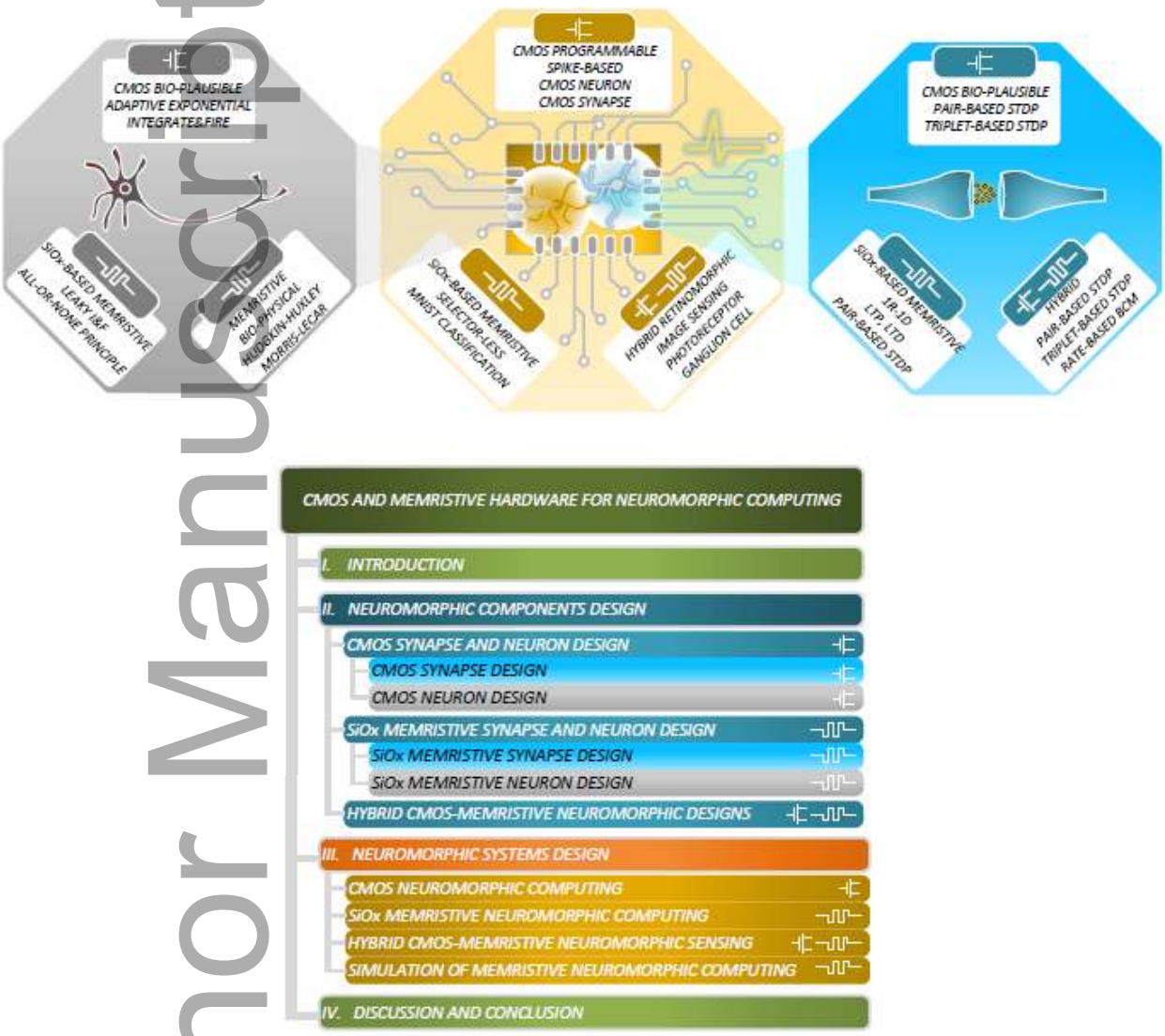
**Mostafa Rahimi Azghadi** completed his Ph.D. in 2014 in neuromorphic VLSI design at Adelaide University, Australia, achieving two doctoral research medals. His main research interests include neuromorphic computing, and hardware implementation of learning for spiking and deep neural architectures. Dr Rahimi was a 2017 QLD Young Tall Poppy Scientist and is currently a tenured Senior Lecturer with James Cook University, Australia.

<qry>Author: Publication and editorial histories have been deleted, in accordance with journal style.</qry>

Author Manuscript

This article is protected by copyright. All rights reserved

Herein, recent advances in CMOS, SiOx-based memristive, and mixed CMOS-memristive hardware for neuromorphic systems are discussed, along with the challenges and opportunities. New and published results are provided from various devices that have been developed to replicate selected functions of neurons, synapses, and simple spiking networks, which have been used for MNIST and pattern classification. <qry>Author: The ToC text has been rewritten in the impersonal form in accordance with journal style.</qry>




M. R. Azghadi,* Y.-C. Chen, J. K. Eshraghian, J. Chen, C.-Y. Lin, A. Amirsoleimani, A. Mehonic, A. J Kenyon, B. Fowler, J. C. Lee, and Y.-F. Chang*

**CMOS and Memristive Hardware for Neuromorphic Computing**

<qry>Author: The ToC image is too detailed to be seen clearly. Please update the image to fit the dimensions, either 55 mm × 50 mm (w × h) or 110 mm × 20 mm (w × h).</qry>

This article is protected by copyright. All rights reserved

Author Manuscript