A CMOS-Memristor Hybrid System for implementing Stochastic Binary Spike Timing Dependent Plasticity

Javad Ahmadi-Farsani[1], Saverio Ricci[2],Shahin Hashemkhani [2], Daniele Ielmini[2], Bernabé Linares-Barranco[1], and Teresa Serrano-Gotarredona[1]
[1]Instituto de Microelectrónica de Sevilla, IMSE-CNM (CSIC and Universidad de Sevilla), Av. Américo Vespucio 28, 41092, Sevilla, Spain
[2]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza L. da Vinci 32, 20133 Milano - Italy

February 14, 2022

Abstract

This paper describes a hybrid system in which a 4×4 memristive crossbar Spiking Neural Network (SNN) was assembled using custom high resistance state (HRS) memristors with analog CMOS neurons fabricated in 180nm CMOS technology. The custom memristors used NMOS selector transistors, made available on a second 180nm CMOS chip. One drawback is that memristors operate with currents in the micro-amperes range, while analog CMOS neurons may need to operate with currents in the pico-amperes range. One possible solution was to use a compact circuit to scale the memristor-domain currents down to the analog CMOS neuron domain currents by at least 5 to 6 orders of magnitude. Here, we proposed using an on-chip compact current splitter circuit based on MOS ladders to aggressively attenuate the currents by over 5 orders of magnitude. This circuit was added before each neuron. This paper describes the proper experimental operation of an SNN circuit using a 4×4 1T1R synaptic crossbar together with four post-synaptic CMOS circuits, each with a 5-decade current attenuator and an integrate-and-fire neuron. It also demonstrates one-shot Winner-Takes-All (WTA) training and stochastic-binary Spike-Timing-Dependent-Plasticity (STDP) learning using this small system.

1 Introduction

In recent years, neuromorphic engineering has attracted great interest in both academia and industry because of its potential for providing energy-efficient artificial cognitive sensory and processing systems that imitate brain functions. Neuromorphic computing and engineering is highly multidisciplinary. It encompasses high level computational neuroscience for unraveling the computing and

learning principles used in biological brains, novel hardware-friendly parallel architectures capable of mapping brain computing principles on fast, efficient hardware platforms, novel circuits that imitate event-driven brain computations, and new nanoscale devices that can be used directly as synaptic or neuron primitives. One key difference between classic computers and neuromorphic computing hardware is the latter's circumvention of the Von Neumann bottleneck [1]. In classic computers, processing elements and memory storage are physically separated and a great amount of energy is consumed in massive data transfers between processors and the different hierarchical levels of memory. In neuromorphic hardware, however, it is possible to co-locate knowledge (stored dynamically as synaptic weights) and information processing (typically performed jointly by synapses and neurons), thus reducing such continuous massive data transfers. One way of co-locating memory and processing is to use memristor devices, which can be fabricated monolithically on top of CMOS neurons [2]. In this regard, many researchers have proposed building hybrid CMOS-memristor neuromorphic computing systems. However, one issue that arises when trying to interconnect memristor-based synapses with compact CMOS neurons is the big difference between their operating currents. Memristor devices typically have an ON-resistance in the range of 2-20$k\Omega$, so synaptic currents flowing through each of them could be of hundreds of micro-amps. On the other hand, compact CMOS neurons integrate synaptic current pulses on small capacitors in the range of tens to hundreds of femto-farads, thus presupposing current pulses in the range of few nano-amps, pico-amps, or even less.

This problem is not normally highlighted in the literature. Researchers have often performed electrical measurements and characterizations on isolated memristor devices or crossbars and then extrapolated their extracted models to numerically simulated full systems [**?**, 3, 4], circumventing the physical problem of current scaling. In other studies, memristor crossbar currents are sensed by on-chip analog-to-digital converters [5, 6], or driven off-chip and integrated by analog integrators with operational amplifiers and large off-chip capacitors [7]. Other reported solutions downscale the memristor current in each synaptic circuit, resulting in costly area overheads [8], or use the memristor crossbar as a digital memory, reading it out with sense amplifiers and using the read digital words to activate correctly scaled, dedicated, digitally- controlled current injectors [9].

Here we propose a solution based on current splitting using compact MOS ladder circuits [10], inserted between the memristor crossbar and the analog CMOS neurons, with one splitter circuit per neuron. We also built a 1T1R memristor crossbar by combining isolated custom-made special high-on-resistance memristors, available on custom chips, with NMOS selector transistors fabricated in CMOS technology. This method was used to assemble a full analog memristor+CMOS multi-chip system. We demonstrate this system's functionality with examples of its use in winner-takes-all (WTA) one-shot training and stochastic binary STDP (Spike Timing Dependent Plasticity) learning [11, 12]. In summary, the contributions/innovations in the present paper are the following:

- **Ladder circuit:** the use of one ladder circuit per neuron allows for an efficient, low power and low area means of interfacing memristor synapses and CMOS neurons (Sections 2-3).

- **Modified neuron circuit:** we present a modification of a previously reported neuron circuit that allows to tune its threshold voltage, depending on the application (Section 4).
- **New memristor stack:** in this paper we use a new memristor stack (with Ti/C/Au top electrode) aimed at reducing the set/reset voltage to below 3V (Section 5), while providing an OFF resistance in the mega-ohm range.
- **Hybrid Memristor-CMOS multi-chip architecture:** we present a new practical low-cost setup for interfacing custom-lab made memristors with 1T MOS transistors made with main-stream CMOS technologies (Section 6).

2 The Problem of Big Differences in Current Domains

Most currently-available memristor devices have low-resistance-states (LRS) in the range of one to ten kilo-ohms from about $R_{ON} \simeq 2k\Omega$ to about $20k\Omega$, and high-resistance-states (R_{OFF}) typically above $100k\Omega$ but with higher degrees of variability [13]. When these memristors are used in a crossbar configuration for performing computational inference (for example, a vector-matrix multiplication) each active memristor is subject to a relatively small amplitude read pulse in the range of $V_{Read} \simeq 100mV - 300mV$, to avoid alteration of the stored resistance state. This read pulse is typically applied for a time T_P in the range of hundreds of nano-seconds or a few micro-seconds. As a result, the charge packet delivered by an individual LRS memristor when stimulated by one single read pulse could be in the order of

$$\begin{aligned} &\delta q_{memr} = I_{ON} \times T_P \simeq 50\mu A \times 100ns = 5pC \\ &\text{with} \qquad I_{ON} = V_{amp}/R_{ON} \simeq 100mV/2k\Omega = 50\mu A \end{aligned} \tag{1}$$

The post-synaptic neurons on the other side of the memristive crossbar integrate all the charge packets produced by dynamically arriving spiking input patterns. When implemented on-chip, these neurons should have minimum area. To properly recognize complex features, they also need to integrate spikes coming from a large number of synapses. These neurons typically comprise a compact integration capacitor C_{memb}, which integrates the charge packets δq coming from the different synapses. The integration is typically leaky, so incoming synaptic charge packets have to coincide within a time window and allow the integrated capacitor voltage to reach a given threshold voltage V_{th} fast enough to counteract the leakage. When the neuron reaches this threshold, its capacitor voltage is reset to the resting level V_{rest}. As a rule of thumb, a neuron in a large-scale neural network can on average be expected to fire after receiving between a hundred and a thousand incoming spikes. This means that the increment or decrement induced in the integrating capacitor's voltage by a single "average" synaptic spike should be about $\Delta V_{spk} \simeq (V_{th} - V_{rest})/n_{spk}$ with n_{spk} typically in the range of 10^2 to 10^3. In analog CMOS neurons $V_{th} - V_{rest}$ is typically in

the range of 1V [14], so ΔV_{spk} would be a round 1 mV to 10 mV. For compact CMOS neurons, capacitance C_{memb} should be kept around $100fF$ or less. Therefore, one individual synaptic charge packet δq_{neur} feeding the membrane capacitance should, on average, satisfy

$$\delta q_{neur} \simeq C_{memb} \times \Delta V_{spk} = 100fF \times 1mV = 0.1fC \quad (2)$$

This charge packet is about a five orders of magnitude smaller charge packet than in Eq. (1). Even for smaller scale proof-of-concept systems with n_{spk} in the order of 10-100 and $\Delta V_{spk} \simeq 100mV$, we would still need to scale the charge packets down by about 3 orders of magnitude. One possibility is to use very fast read pulses with T_P in the range of $100ps$ or less. This would require the use of fast deep-submicron technologies and fast crossbar driver circuits. Alternatively, one may think of using a $10^3 - 10^5$ larger capacitor, but this implies multiplying by the same factor its area, making it prohibitive (for example, in our neuron design as explained later, C_{memb}'s area is 20% of the neuron area; increasing it 10^3 times would increase the neuron area by about 200 times). If very fast pulses are not feasible, the only alternative is to provide some kind of mechanism for scaling or mapping from the memristor-domain current/charge-packet levels to the neuron-domain current/charge-packet levels.

One solution is to use ADC converters to collect the memristor crossbar currents [15, 16, 5]. The information is then switched to the digital domain, where the rest of the computation can be performed. Alternatively, other researchers have proposed techniques to scale from the memristor-domain to the neuron-domain current levels inside each synaptic circuit [8]. This, however, results in a high overall chip area penalty. In this study, we decided to use a compact ladder-based circuit at each neuron input, capable of scaling down the memristor-domain current by several orders of magnitude. This also results in a highly energy-efficient technique, due to circuit simplicity, as will be highlighted later in the experimental results (see Table 5). For now, let us define the energy consumption by one LRS memristor as $E_{LRS} = V_{DD} \times \delta q_{memr}$, where V_{DD} is the power supply voltage.

3 Compact Ladder-Based Circuit for Current Down-Scaling

MOS-based ladder circuits have been known since 1992 [10] and have proven capable of downscaling currents from hundreds of micro-amps to a few femto-amps [17]. Fig. 1 illustrates the MOS ladder-based current-splitting technique

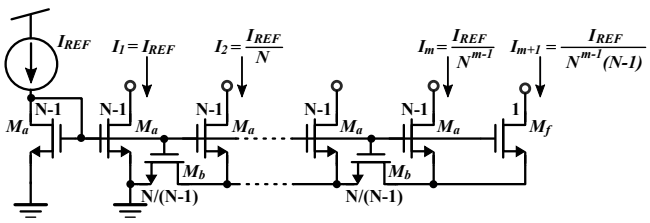


Figure 1: Circuit schematic for generic current splitting ratio N

for a generic branch-to-branch scaling factor N. The transistors have a size ratio of either $W/L = N - 1$, $W/L = N/N(N - 1)$, or $W/L = 1$. Normally, ladder circuits are used with $N = 2$, thus providing binary-weighted currents which are very convenient in, for example, digital to analog converters (DACs). Here, however, we needed to downscale the current aggressively with a reduced number of transistors. We therefore used current ladders with $N = 10$, while minimizing transistor dimensions.

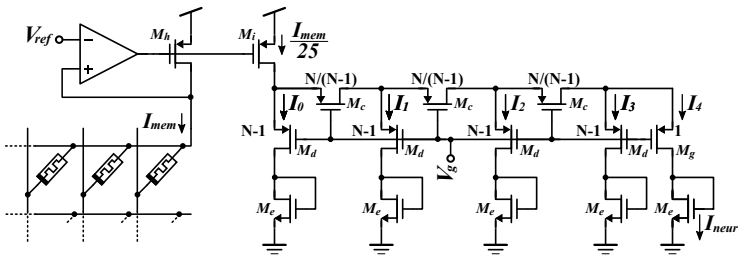


Figure 2: Compact ladder-based MOS current splitter circuit used to downscale the memristor-domain current I_{mem} (in the range of $10 - 100\mu A$) by five orders of magnitude to neuron-domain current levels. Input mirror M_h-M_i provides an extra 25 factor attenuation.

Fig. 2 shows the specific circuit used in this study. It used a ladder branch-to-branch scaling factor $N = 10$, with transistor sizes as shown in Table 1, downscaling currents by about 5 orders of magnitude from the memristor-domain to the neuron-domain. The area consumed by this atteanuator circuit is $68 \times 36\mu m^2 = 2448\mu m^2$. Depending on the scale of the neural network, the average number of synapses connecting to a neuron, the pulse width stimulating the memristors, the value of the memrsistor ON (LRS) resistance, and the desired average number of incoming spikes that should trigger a post-synaptic neuron output spike, one may need a different down-scaling factor for the current attenuator. In our fabricated prototype we connected branch I_4 in Fig. 2 to provide the current for the neuron I_{neur}. However, any other branch could have been selected, from I_4 to I_0, to feed current I_{neur} giving the possibility of scaling down by 5 to one orders of magnitude, respectively.

Table 1: PMOS Ladder Transistor Sizes

	M_c	M_d	M_e	M_g	M_h	M_i
W (μm)	1.1	9	1	1	25	1
L (μm)	1	1	1	1	4	4

4 CMOS Neuron Circuit

Fig. 3 shows the neuron circuit employed in the study. The schematic is separated into six conceptual blocks, A to F. This neuron is a leaky integrate-and-fire neuron, with positive feedback to sharpen spikes, a frequency adaptation mechanism, and a refractory period mechanism. It is based on the CMOS neuron reported by Qiao et al. in 2015 [18] in which we simplified some parts, but added

block E so that the neuron threshold voltage can be freely adjusted. Block A feeds the input spikes, block B provides constant leakage, block C provides the frequency adaptation mechanism, block D controls the positive feedback spike-sharpening mechanism, block F provides the refractory mechanism, and block E is a comparator that activates blocks C, D, and F. Transistor M1 in block A mirrors the output current pulses coming from the current-attenuator I_{neur} into the neuron circuit, where they are integrated by membrane capacitor C_{memb}. Transistor M2 tends to isolate the neuron from the incoming synaptic circuit during spike generation to minimize crosstalk. Concurrently, transistor M3 in block B introduces a comparatively small but continuous leakage current I_{leak}, which slowly discharges the membrane capacitor. Capacitor C_{memb}'s top terminal (which is the neuron's output terminal *Output*) is also connected to the negative input in the comparator (block E). When the *Output* voltage exceeds the reference voltage Vth_V of the comparator's positive input, the comparator's output voltage at node V_n starts falling. As a result, transistor M12 activates block D and injects a positive feedback current I_{fead} into C_{memb}, leading to the generation of a sharper spike at *Output*. Simultaneously, transistor M5 activates block C, causing a charging current to be injected into recovery capacitor C_{rec}. Block C is in charge of the spike-frequency adaptation mechanism, which serves to progressively lower the neuron's firing rate in response to a continuous input stimulation [19]. This way, block C adds an additional leakage current to membrane capacitor C_{memb} when the neuron's spiking output activity increases. Block F implements the refractory mechanism. Transistor M21 is activated on the rising edge of a spike (falling edge at node V_n) and charges refractory capacitor C_{ref}. This leads to the discharge of C_{memb} through M13, in order to hold *Output* close to rst_V. After a spike, capacitor C_{ref} is discharged with a small current controlled by voltage $refractory\_I$ via transistors M22-M24. The refractory period lasts until V_{ref} falls below transistor M13's threshold voltage.

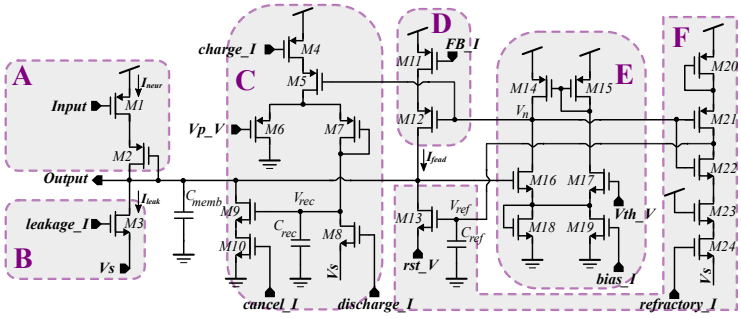
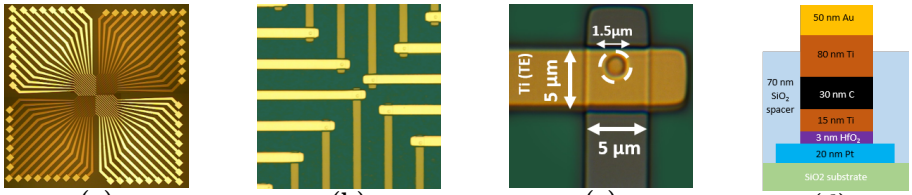


Figure 3: Circuit level schematic of the neuron circuit

The three capacitances of the neuron circuit in Fig. 3 were designed with $C_{memb} = 150fF$, $C_{ref} = 100fF$, and $C_{rec} = 100fF$. The total area consumed by the neuron is $57.7 \times 15.5\mu m^2 = 863\mu m^2$, including all 3 capacitors. This CMOS technology allows placing capacitors over the transistors, while their density is about $1\mu m^2/fF$. Consequently, the area of the three capacitors is about $350\mu m^2$ and could be fit above the rest of the neuron circuitry.

5 RRAM Devices

In this paper we report for the first time experimental results using a new memristor device aimed at reducing the set/reset voltages to below 3V, while presenting an OFF resistance of about $1M\Omega$ or higher. Figs. 4a-4c show SEM images of the RRAM cells used in this study. The RRAM devices comprised stacks are made up of a Pt bottom electrode (BE), a HfO_2 active layer and a Ti/C/Au top electrode (TE). The fabrication procedure was as follows. First, a SiO_2 layer was deposited by chemical vapor deposition (CVD) on a heavily p-doped silicon wafer serving as substrate. The $20nm$-thick Pt bottom electrode was then deposited using ultra-high vacuum ($10^{-7}mbar$) e-beam evaporation and patterned using lithography and lift-off. All the lithographic steps were performed using a Heidelberg MLA100 UV laser writer. A $70nm$-thick SiO_2 spacer layer was then deposited to isolate the BE lines from the TE lines. Holes with diameters of $1.5\mu m$ were then made through the spacer layer by reactive ion etching (RIE). The holes were opened in correspondence of the BEs with regular lattice spacing, to serve as cell active regions in the array. The $3nm$ thick HfO_2 film was then deposited, again by e-beam evaporation, followed by the TE stack, without breaking the vacuum between the different layers. A $15nm$-thick Ti cap layer was deposited on HfO_2, followed by a $30nm$ C layer to increase the series resistance and thus reduce the possible overshoot effects at forming and set transitions. A thick Ti/Au layer was finally deposited as an electrical contact. The oxide/TE stack was patterned using lithography and lift-off.



(a) **(b)** **(c)** **(d)**

Figure 4: Microphotographs of (a) the memristor die, (b) the top and bottom interconnects, and (c) the memristor active area. (d) The RRAM device structure

The RRAM device structure can be seen in Fig. 4d. Possible short circuits between the TE and BE layers are prevented by the SiO_2 spacer layer. The Ti cap layer has been reported to act as an oxygen scavenger, leading to the formation of a TiOx oxygen-exchange layer at the Ti-HfO_2 interface [20]. This mechanism leads to an increase in the local concentration of oxygen vacancies in HfO_x, which in turn enhances the leakage current in the pristine state and reduces the devices' forming voltage [21].

The RRAM devices were first characterized using an Agilent 4156C Precise Semiconductor Parameter Analyzer. Fig. 5 shows the multiple-cycle DC characteristics indicating repeatable set and reset transitions when positive and negative voltages, respectively, were applied to the TE. Also shown is the first cycle, with the forming transition taking place at around $3V$. This is relatively high compared to the typical set voltage of about $1.5V$. A curve was collected

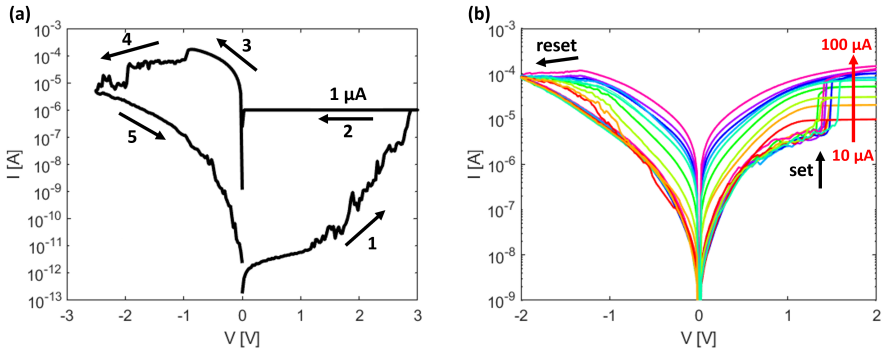


Figure 5: I-V characteristics of the RRAM device. (a) Forming characteristics. The device is initialized by the forming operation with a relatively small compliance current of 1μA (1-2). The first reset operation (3) shows a relatively large reset current of about 150μA, due to the parasitic currents during the high voltage forming. Extending the reset operation to large negative voltages (4) allows to reach a deep reset state with relatively large resistance (5). (b) set/reset characteristics. The following set/reset curves show a relatively tight distribution of set voltages and controllable resistance via the compliance current.

for each increment in compliance current I_C from $10\mu A$ to $100\mu A$. As I_C increased, the low resistance state (LRS) conductance increased almost linearly with it, while the high resistance state (HRS) conductance remained almost constant.

The pulsed operation of the RRAM devices was studied using a TTi-TGA12102 Arbitrary Waveform Generator and a LeCroy WaveRunner 640Zi oscilloscope. The compliance current was applied by an external transistor. Fig. 6(a) shows the oscilloscope traces for the applied voltage and the measured current. Positive and negative triangular pulses with pulse-widths of $20ms$ were applied for set and reset transitions, respectively. From the measured current and voltage, we obtained the pulsed I-V curves displayed in Fig. 6(b). Each curve represents the averaged current between 100 characteristics. The results indicate a non-linear characteristic where the resistance window increases with I_C. Fig. 6(c) shows the cumulative distributions of the measured LRS conductance G at increasing I_C, indicating a standard deviation of about $12\mu S$. Fig. 6(d) shows the measured HRS and LRS conductance values as a function of I_C, supporting the increase in the resistance window at increasing compliance current. Conductance increased with a slope of approximately one, indicating a linear relationship with I_C, except for the relatively low G, where the LRS collapsed with the HRS level.

For our setup and experiments described in the rest of this paper, we wanted to have a large resistance window, maximizing the OFF resistance (HRS). We therefore used high compliance current levels and used the memristors as binary memories with maximally separated LRS and HRS.

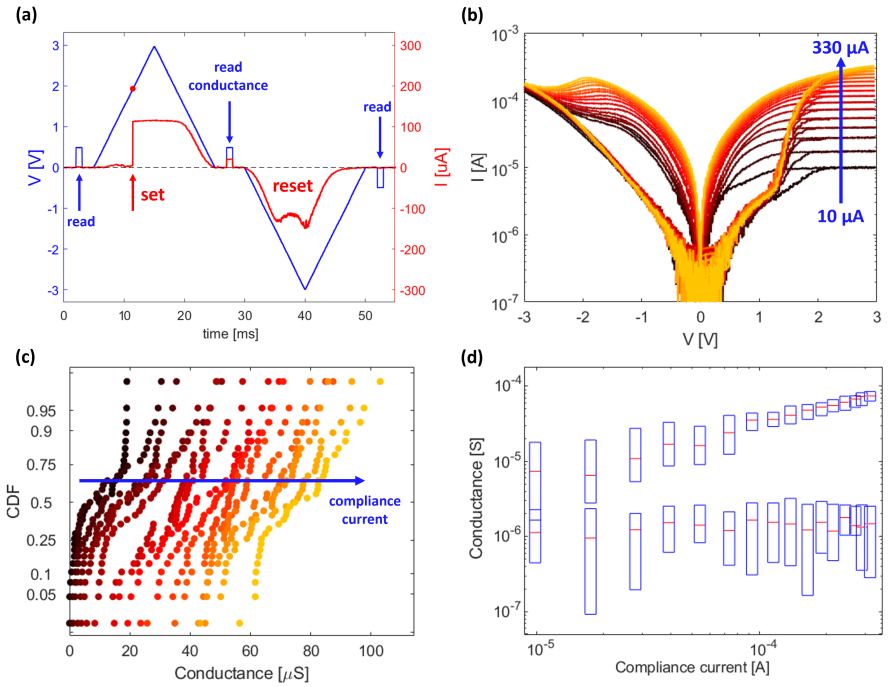


Figure 6: Pulse characterization of the memristors: (a) measured current (red) and applied voltage (blue) versus time, (b) I-V curves obtained by applying triangular pulses, (c) cumulative distributions of the measured conductance, and (d) measured HRS and LRS conductance values as a function of compliance current I_C.

6 System Architecture

In this paper we showcase a small 4×4 1T1R synaptic memristor crossbar with CMOS analog neurons performing learning and inference. Our system illustrates how to use a set of separate CMOS chips together with custom made memristive devices to build a hybrid CMOS-memristor system operated with an auxiliary custom PCB and controlled by an FPGA. The interesting novelty about this approach is that it shows a simple way for memristor researchers to assemble 1T1R arrays by combining custom memristor-only chips with standard CMOS ASICs. The system's overall architecture is shown in Fig. 7. The memristors were on a separate chip (colored green in Fig. 7). As can be seen in Fig. 4a, a total of 32 separate memristors were on the chip. Due to yield issues, however, not all 32 were functional. Here we used a total of 16 memristors, each having a separate pin for its bottom plate while the top plates were shared by the four memristors in the same row (see Fig. 7). The 16 NMOS selectors were fabricated on a separate CMOS chip (colored blue in Fig. 7). These shared their gates row-wise and their source terminals column-wise. Their drains were connected individually to each of the memristors' bottom plates. This way, a full 4×4 1T1R synaptic array could be assembled using custom memristors.

The post-synaptic CMOS circuits (shown in orange in Fig. 7) were allocated on a separate chip. They each included one current attenuator (See Section 3) and one CMOS neuron circuit (see Section 4). The other elements in Fig. 7 were allocated on a custom PCB. This custom PCB had some external digital control signals ("Row Active Select", "Column Active Select", and "Inf"), set by an additional FPGA control board running a state machine. The system in Fig. 7 could be configured in two operation modes, "Inference" mode or "Element-Wise" mode, using a digital control signal Inf.

6.1 Inference Mode

When $Inf = 1$ the inference mode is activated and the synaptic crossbar will perform parallel inference. In this mode, all 1T gates are connected to a gate inference voltage bias V_{Ginf}, each of the four memristor row top-plates is connected to one post-synaptic circuit, and each of the four 1T source columns is connected to one pre-synaptic circuit (Pre_i in Fig. 7). The pre-synaptic circuits are physically implemented on the custom PCB and are simple switches connecting the column to either a V_{HIGH} or V_{LOW} voltage, depending on digital input In_i (either '0' or '1'), as illustrated in the inset in Fig. 7. In this "Inference Mode", all post-synaptic rows are set to voltage V_{HIGH}, while active digital input In_i will set the corresponding column to V_{LOW}. This way, the current flowing from a post-synaptic circuit will be given by

$$I_j = \sum_{i=1}^{4} R_{ij}^{-1} \Delta V_{Read} In_i \tag{3}$$

where R_{ij} is the resistance of the 1T1R synapse at column i and row j, and $\Delta V_{Read} = V_{HIGH} - V_{LOW}$. In this "Inference-Mode", post-synaptic currents I_j will be collected by the current attenuator circuits and scaled down before being integrated in the neurons. The neuron outputs Out_j are monitored by the FPGA state machine.

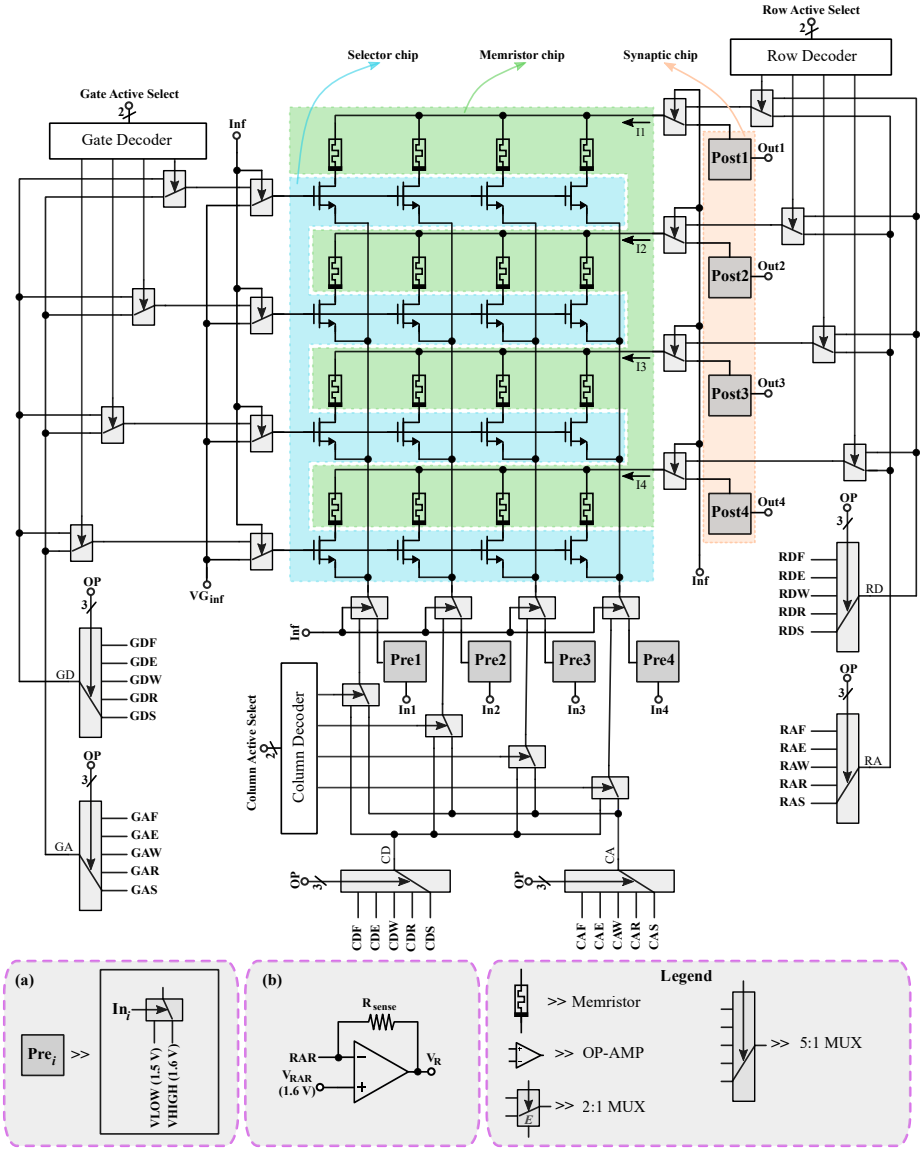


Figure 7: System level architecture of the memristor-based spiking neural network

6.2 Element-Wise Mode

When $Inf = 0$, the "Element-Wise Mode" is activated. In this mode, only one column and only one row at a time are set as "active", so only one 1T1R element is selected to perform an individual "Forming", "Write", "Erase", or "Read" operation on them. The 2-bit digital control "Row Active Select" sets one of the rows as the active row, while the 2-bit digital control "Column Active Word" sets one of the columns as the active column. The other, non-active, rows and columns are set as "default". For the active row, the gates of the 1T NMOS selector transistors are connected to node "GA" (Gate Active) while the other gates are connected to node "GD" (Gate Default). The memristor top plate node of the "active" row is connected to node "RA" (Row Active), while the others are connected to node "RD" (Row Default). Finally, the active column is connected to node "CA" (Column Active), and the other columns are connected to node "CD" (Column Default).

The operation to be performed at the individually selected 1T1R synapse is set by the 3-bit digital control word "OP". It may be a "Forming", "Write", "Erase", "Stand-by", or "Read" operation. The stand-by mode is "Read-Mode", in which all terminals are connected to the "default" value. This is a safeguard measure to avoid undesirable glitches when switching active rows/columns or operation modes. Stand-by mode should therefore be inserted when switching between "Forming", "Write", "Erase" or "Read" operations. Similarly, stand-by mode should also be used when changing active columns/rows. It should also be noted that the "Forming", "Write", "Erase", and "Read" operations are to be performed during well-defined time durations, while "Stand-by" can have an arbitrary duration.

For each of the six modes in Fig. 7, GA (Gate Active), GD (Gate Default), RA (Row Active), RD (Row Default), CA (Column Active), and CD (Column Default) the corresponding active lines should be connected to five different bias voltages, depending on the selected operation. This results in a total of 30 different bias voltages, each of which can be adjusted individually on the custom PCB. These 30 bias voltages are available at the 30 nodes in Fig. 7 which are labeled with three capital letters XYZ , where 'X' is either 'G' (Gate), 'R' (Row), 'C' (Column), 'Y' is either 'A' (Active) or 'D' (Default), and 'Z' is either 'F' (Forming), 'E' (Erase), 'W' (Write), 'R' (Read), or 'S' (Stand-by).

Fig. 8 indicates the gate, column, and row voltages to be set for the active and default columns and rows for the five different operation modes and the inference mode.

These voltage bias settings are summarized in Table 2. During element-wise reading, the aim is to accurately read the resistance of the selected 1T1R synapse. To do this, the current sensing-circuit shown in inset (b) in Fig. 7 is connected to node RAR. By reading voltage V_R, it is possible to infer the resistance of the corresponding 1T1R synapse by using

$$R_{ij} = R_{sense}\frac{V_{HIGH} - V_{LOW}}{V_R - V_{RAR}} = R_{sense}\frac{0.1V}{V_R - 1.6V} \tag{4}$$

The sizing of the selector transistor is critical. On one hand, it is desirable it is not too large. This way, when scaling up the system, higher synapse densities can be achieved. However, on the other hand, its size ratio (W/L) should be large enough for allowing the maximum required currents for the different

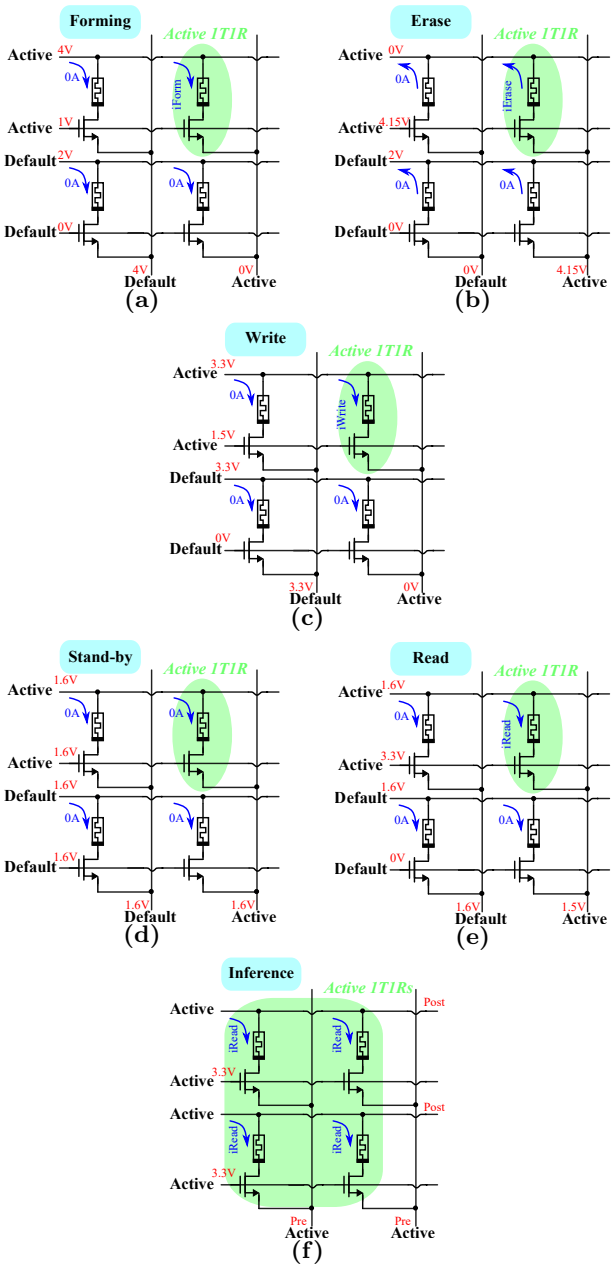


Figure 8: Illustration of the "active" and "default" voltage levels used for the different operating modes. a) Forming, b) Erase, c) Write, d) Stand-by, e) Read, and f) in the parallel Inference mode

Table 2: Active and Default Bias Voltages Used for the Different Operations

		Element-Wise Mode					Inference-Mode
		Forming	Erase	Write	Read	Stand-by	
Gate	Active	1 V	4.15 V	3.3 V	1.6 V	1.6 V	3.3 V
	Default	0 V	0 V	0 V	0 V		
Row	Active	4 V	0 V	3.3 V	1.6 V	1.6 V	1.6 V
	Defualt	2 V	2 V	3.3 V	1.6 V		
Column	Active	0 V	4.15 V	0 V	1.5 V	1.6 V	1.6 V / 1.5 V
	Default	4 V	0 V	3.3 V	1.6 V		

operations. The most critical operation is resetting the memristor from LRS to HRS, since the memristor has a low resistance while we need to apply to it a relatively large voltage of about 3V. We sized our selector transistor to have minimum length ($L = 350nm$) and a width of $W = 13.4\mu m$. Under these conditions, we could erase safely all fabricated and operational memristors while applying an erase voltage to the 1T1R compound of 4.15V, as shown in Table 2.

7 Experimental Results

Three separate chips were fabricated. One was a custom memristor chip containing 32 individual memristors and the other two were ASIC chips fabricated in TSMC 180nm CMOS technology, one of which contained the NMOS transistor selectors and the other the CMOS current attenuators and neurons for the post-synaptic circuits.

Fig. 9 shows micrographs of the three chips. Details of the custom memristor chip can be seen in Fig. 9a. This chip has a total of 32 independent, individual memristors, each connected to an Au (gold) top electrode and a Pt (platinum) bottom electrode. Each electrode is $5\mu m$ wide and the circular active area of each memristor at the cross points on both electrodes has a diameter of $1.5\mu m$. The chip has a total of 64 pins (32 TE pins and 32 BE pins). Not all of the 32 memristor devices were fully functional. From the ones that were, 16 were selected to be connected to the CMOS chips.

Fig. 9b shows a micrograph of the fabricated NMOS selector chip. Each NMOS transistor has a size of $W = 6.7\mu m$ and $L = 350nm$. The fabricated array has 8×8 selectors. Since we used only 16 memristors, a sub-array of 4×4 selectors was used in the setup. Fig. 9c shows a micrograph of the post-synaptic chip. It includes a total of 8 post-synaptic circuits, each including one current attenuator, one CMOS neuron, and one output buffer. Each post-synaptic circuit has an area of $180 \times 50\mu m^2$, of which $2520\mu m^2$ are occupied by the current-attenuator, $840\mu m^2$ by the neuron, and $1272\mu m^2$ by the buffer. In our setup, 4 post-synaptic circuits were used.

Fig. 10 shows the full experimental setup, including one PCB holding the memristor chip, another PCB holding the selector chip, another PCB holding the post-synaptic circuit chip, the custom PCB with all the switches, multiplexers and potentiometers providing all the biases, and the FPGA-based controller PCB.

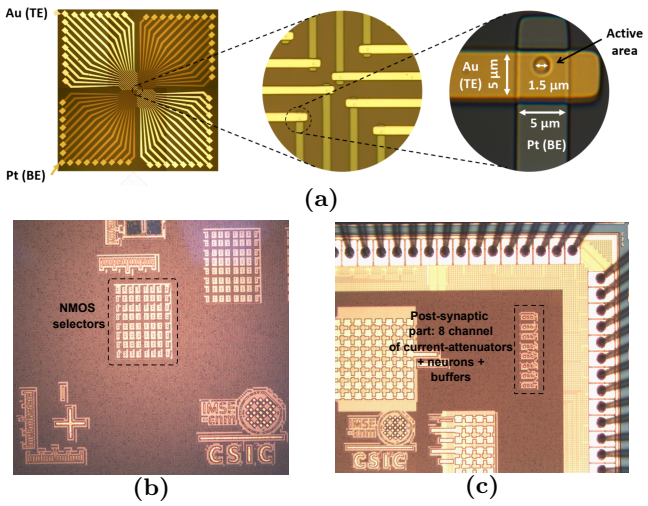


Figure 9: Micrographs of (a) the memristor chip (with progressive zoom-ins), (b) the selector transistor chip, and (c) the post-synaptic circuit chip.

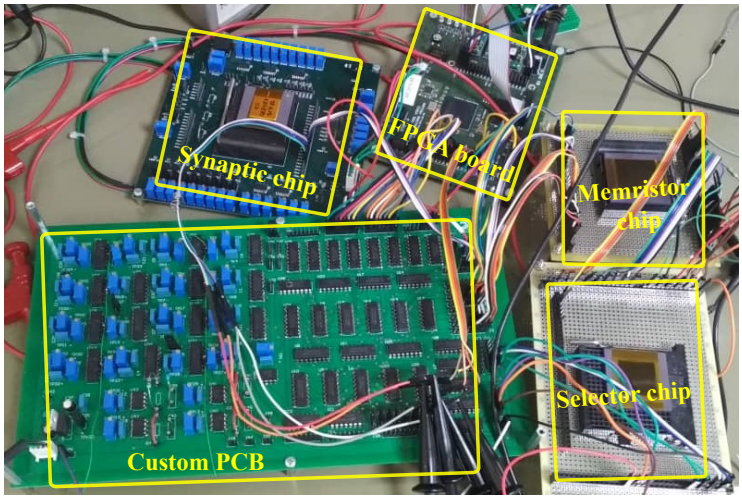


Figure 10: Experimental test setup showing all PCBs and chips

7.1 Current-Attenuator Test

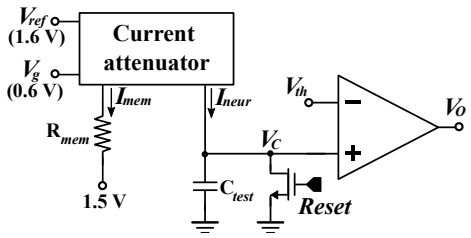


Figure 11: On-chip dedicated circuit used to characterize the current attenuator output current

Fig. 11 shows the on-chip dedicated circuit used to characterize the current attenuator. Since the attenuator output currents can be very small (less than 1 pico ampere), currents cannot be driven off-chip and measured by external instruments [17]. With the circuit in Fig. 11 it was possible to measure the charging time of the integrating capacitor and estimate the charging current coming from the attenuator output I_{neur}. The current attenuator input current I_{mem} was set by changing the off-chip resistance R_{mem}.

$$I_{mem} = \frac{100mV}{R_{mem}} \tag{5}$$

To measure the attenuated output current I_{neur}, capacitor C_{test} was initially discharged to 0V and then charged by I_{neur} until the capacitor voltage reached V_{th}. By observing the digital output V_o of the voltage comparator, it was possible to measure the time Δt between capacitor reset and the instant at which the capacitor voltage reached V_{th}

$$I_{neur} = \frac{C_{test}V_{th}}{\Delta t} \tag{6}$$

Capacitor C_{test} was designed with a capacitance of 1.5pF and voltage V_{th} was set at 500mV.

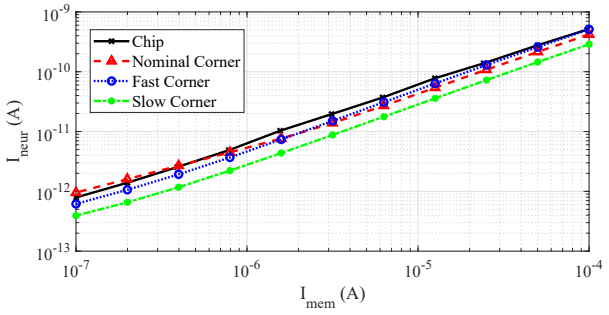


Figure 12: Current-attenuator output current I_{neur} vs its input current I_{mem}

Fig. 12 shows the characterization results for the measured values of current I_{neur} vs the attenuator input current I_{mem}. It can be seen that the attenuation factor for synaptic memristances R_{mem} in the range of 1 kΩ to 10 kΩ was

between 1.9×10^5 and 1.6×10^5. Table 3 summarizes the values used for R_{mem}, the values measured for Δt, and the inferred values for I_{mem}, I_{neur}, and the attenuation factor.

Table 3: Input (I_{mem}) and output (I_{neur}) currents of the attenuator along with the measured charging times (Δt) at capacitor C_{test} and the resulting current attenuation factors for a set of different equivalent input memristor resistances R_{mem}

#	R_{mem} (kΩ)	Δt (ms)	$\mathbf{I_{mem}}$	$\mathbf{I_{neur}}$	**Attenuation** ($\mathbf{10^5}$)
1	1.02	1.47	black97.7 μA	50.86 pA	1.92
2	2	2.70	black50.0 μA	27.73 pA	1.80
3	4.13	5.41	black24.2 μA	13.83 pA	1.74
4	7.97	9.65	black12.5 μA	7.77 pA	1.61
5	16.1	20.23	black6.21 μA	3.70 pA	1.67
6	32	38.34	black3.12 μA	1.95 pA	1.59
7	64	72.75	black1.56 μA	1.03 pA	1.51
8	127.7	153.99	black783.0 nA	487.03 fA	1.60
9	256.1	295	black390.4 nA	254.23 fA	1.53
10	500	537.59	black200 nA	139.51 fA	1.43
11	997	947.68	black100.3 nA	79.14 fA	1.26

7.2 Neuron Circuit Test

Fig. 13a shows the test configuration used to characterize the neuron cell circuit. In this arrangement, an external $1k\Omega$ resistor was connected between the input node of the current-attenuator and a $1.5V$ reference voltage. For this resistance, it can be seen from Table 3 that the attenuator output current to the neuron I_{neur} was slightly below $1pA$. Three buffers were used to isolate the three capacitor voltages in Fig. 13a ($Output$, V_{ref}, and V_{rec}) and thus allow efficient, undisturbed off-chip observation. Fig. 13b shows these three neuron voltages during a time period of $150ms$. In this experiment, the neuron threshold voltage was set to $2.2V$ and the refractory period was set to about $7ms$.

7.3 One Shot WTA Training

In a first system-level experimental setup, we performed a one-shot WTA-driven training demonstration. For this purpose, all 16 memristors in the crossbar were first initialized to their ON state (LRS). The four 4-bit patterns p1, p2, p3, p4 shown in Fig. 14 were then used as input patterns. For each pattern, simultaneous pulses were applied by the pre-synaptic neurons with an active bit. The patterns were applied for long enough to allow at least one of the output neurons to reach its threshold level. The neuron that first reached its threshold level was the winning neuron, and the weights of the synapses connecting to it were updated. For a synapse between the winning post-synaptic neuron and an active pre-synaptic neuron, no action was taken. For a synapse between the winning post-synaptic neuron and an inactive pre-synaptic neuron resistance was set to high (HRS) by performing an erase operation.

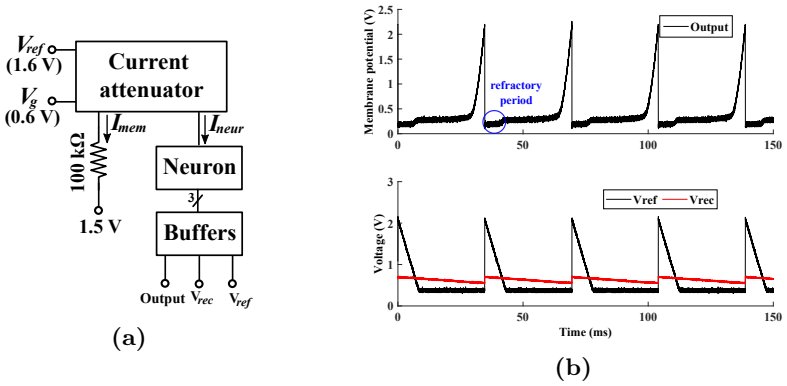


Figure 13: (a) Test configuration for the neuron circuit, (b) Membrane potential ($Output$), refractory variable voltage (V_{ref}), and recovery variable voltage (V_{rec}) of the neuron circuit

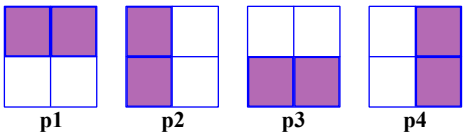


Figure 14: Input patterns to the SNN

Fig. 15 shows the experimentally measured membrane voltages of the post-synaptic neurons when the input patterns in Fig. 14 were applied sequentially and the weights updated in accordance with this WTA-driven training method. The left-most column corresponds to input pattern p1, the second left-most column to p2, and so on. The top row (a-d) shows the measured membrane voltages of the four post-synaptic neurons for each input pattern when all memristors were initially set to LRS. The second row (e-h) shows the membrane voltages after implementing weight changes when pattern p1 was applied and post-synaptic neuron 1 (blue) won the WTA competition. Therefore, only connections to neuron 1 get updated. Note that when applying input patterns 2 to 4 (see (f-h)), neuron 1 behaves differently than in (b-d), since its input weights have changed. The third row (i-l) shows the voltages after the next input pattern p2 was applied and post-synaptic neuron 3 (black) won. Now, neuron 3 changes its behavior in (i,k,l) with respect to (e,g,h). The fourth row (m-p) shows the voltages after the next pattern p3 was applied and post-synaptic neuron 4 (green) won. Now neuron 4 changed its behavior in (m,n,p) with respect to (i,j,l). Finally, the fifth row (q-t), shows the voltages after the next input pattern p4 was applied and post-synaptic neuron 2 (red) won, showing a different behavior in (q-s) with respect to (m-o). In this last row it can clearly be seen that each output neuron responded very strongly to only one of the input patterns.

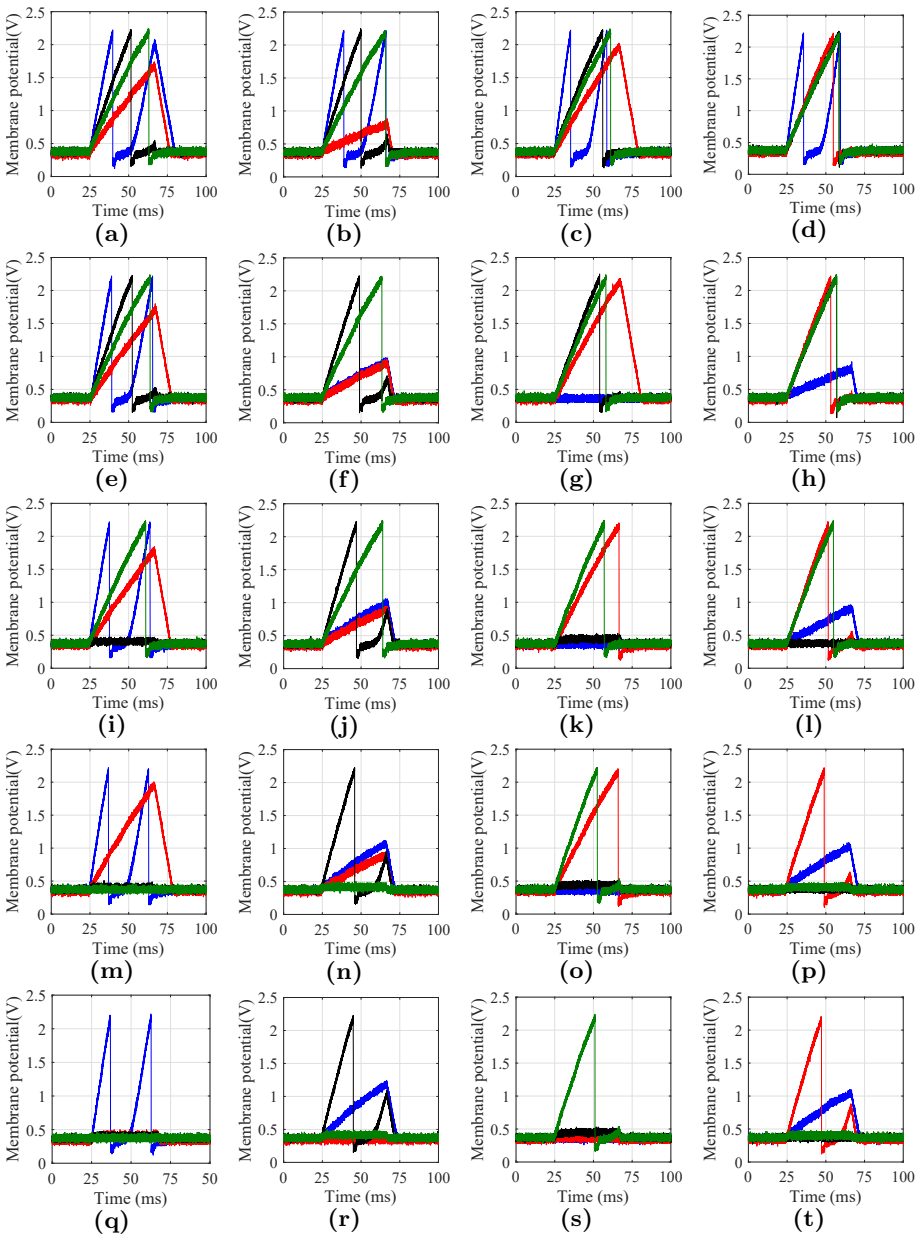


Figure 15: Neuron outputs for each input pattern. (a-d) before any weight update, (e-h) after presenting p1 and corresponding weight updates, (i-l) after presenting p1 and p4 and corresponding weight updates, (m-p) after presenting p1, p4, and p3 and corresponding weight updates, (q-t) after presenting all the input patterns and all corresponding weight updates. blue•: neuron1, red•: neuron2, black•: neuron3, green•: neuron4

7.4 Stochastic Binary STDP

Spike-Timing-Depedent Plasticity (STDP) [22] is a bio-inspired learning rule for SNNs and allows, in principle, for continuous on-line learning. Stochastic binary STDP (SB-STDP) is a variant of the original STDP learning rule in which synapses only present an ON and an OFF state and the weight updates follow a stochastic rule [11, 12]. Therefore, SB-STDP is quite appropriate for binary RRAM synapse SNNs. Originally, SB-STDP was proposed by simply substituting the original deterministic STDP gradual update [22] by a non-gradual stochastic one [11]. However, later on, it was shown that for correct operation on scaled-up systems some regularization techniques were required [12]. In the example case we are considering here, which is just a small 4×4 crossbar, we only needed to consider one regularization technique, namely homeostasis.

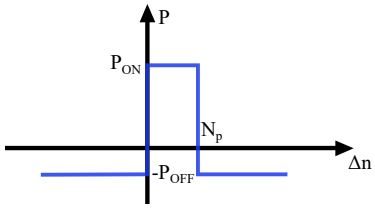


Figure 16: Weight change probability for Stochastic Binary STDP. If the sequence order between post-synaptic and pre-synaptic spikes is positive and less than N_p, the corresponding synapse is changed to ON with probability P_{ON}. Otherwise, all other synapses connecting to this post-synaptic neuron are set to OFF with probability P_{OFF}.

In SB-STDP, the synapses connected to a firing post-synaptic neuron are updated following the process explained below with reference to Fig. 16. The most recent N_P input spikes are kept on a list. Input and output spikes are all indexed sequentially by a counter n. Whenever an output neuron j spikes, all previous N_P input spikes are retrieved. For each input neuron i, only its most recent spike is considered, so a synapse ij between an active input neuron i on the list and active output neuron j is updated only once for each output spike. If synapse ij is already ON, it is left untouched. But if it is OFF, it is changed to fully ON with a probability P_{ON}. Once all active connections obtained from the list have been updated probabilistically, the total number of ON synapses is counted. In SB-STDP, to implement homeostasis, the sum of ON synapses connecting to an output neuron j is kept constant. Let us call this constant M. If the sum is greater than M, then one of the synapses which was not retrieved from the list and is ON is chosen randomly and its weight is set to OFF. This process is repeated until the sum of ON synapses is M.

To perform SB-STDP in our setup, we used the four input patterns shown in Fig. 17b bottom, where each input pattern is a horizontal 4-bit row. Each input pattern was applied by having its corresponding active input neurons present a sequence of randomly spaced spikes. Initially, all synaptic memristors were initialized to their ON state (LRS). As soon as an output neuron spiked, the memristors connecting to it were updated following the SB-STDP rule described above. All neurons were then reset, and a new input pattern was applied.

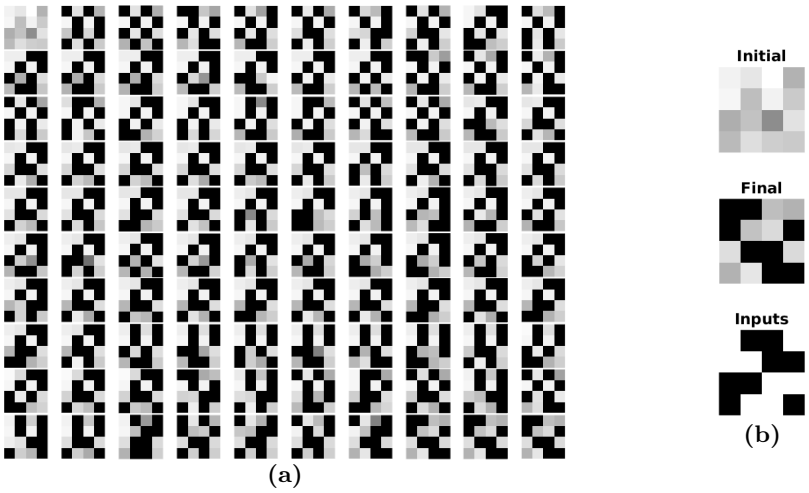


Figure 17: (a) Samples of the sequence of weight updates in SB-STDP, (b) Initial and final weights, and the four (row-wise) input patterns used for SB-STDP training

Fig. 17a shows samples of a sequence of weight updates, starting with the initial set of weights (all ON, top left) and finishing with the stabilized set of weights (bottom right). The number iterations until convergence varied from about 15 up to 200, with an average of about 50. The ON weights fall approximately in the 2-6$k\Omega$ range and are shown in gray scale, the minimum resistance in all 16 memristors over the full learning sequence being white. Fig. 17b shows the initial weights (top), the final weights (center), and the input patterns (bottom). Each row in Fig. 17b bottom corresponds to one 4-bit input pattern. It can be seen that the final weights correspond to the input patterns but are shuffled row-wise. Consequently, the system successfully learned the input patterns.

8 Benchmarking for Energy Efficiency

In order to compare with other reported state-of-the-art neural processing chip systems, let us first analyze which would be the optimum settings for a chip following our approach to minimize energy consumption per synaptic computation. The largest currents in our system are currents flowing through memristors which are at LRS. Therefore, the minimum energy per operation would be given by finding the fastest pulses that can be applied to LRS memristors while guaranteeing that the circuit providing current to them (transistor M_h in Fig. 2 and the differential amplifier) can propagate the corresponding memristor-domain charge packet down to the neuron-domain with sufficient integrity. By assuming an average LRS value of 10$k\Omega$, we found (by simulations) this limit to be $T_p \simeq 100ns$, in which case the memristor-domain charge packet was found to be $0.68pC$ (about 30% below the ideal value), and the circuit needed a total time of $360ns$ to settle. Under these conditions, we analyzed which would be the neuron-domain charge packets if selecting for I_{neur} the different ladder branches in Fig. 2, from I_0 to I_4. To find these charge packets we obtained the voltage

increments ΔV_{spk} induced at the neuron membrane capacitance C_{memb}. From these voltage increments, we can compute the neuron-domain charge packets as $\delta q_{neur} = C_{memb} \times \Delta V_{spk}$. The results are summarized in Table 4. For maximum

Table 4: Effective charge-packet attenuation for maximum speed and minimum energy

Attenuator branch	ΔV_{spk}	δq_{neur}	Effective attenuation
I_0	$183mV$	$27fC$	25.2
I_1	$19.5mV$	$2.9fC$	234
I_2	$1.88mV$	$0.28fC$	2.43×10^3
I_3	$268\mu V$	$0.04fC$	1.7×10^4
I_4	-	-	-

speed pulses of $100ns$ stimulating the memristors, one can observe voltage increments at the neuron membrane voltage when selecting ladder branches I_0 to I_3. When selecting branch I_4, no change is appreciated. Depending on the selected ladder branch, the effective charge packet attenuation ranges from 25.2 up to 1.7×10^4. Also, for practical charge packet sizes (those that induce membrane voltage increments in the range of $1mV - 50mV$, as discussed in Section 2), one would require attenuation factors in the range of hundreds to thousands for such fast $100ns$ stimulation pulses.

Regarding energy consumption in our system, there is energy due to stand-by power and energy due to memristor currents and circuit transients during input spikes. The power consumption of the neurons is negligible. A neuron firing at a high rate of $1KHz$ consumes about $140nW$. The circuit component consuming most of the stand-by power is the differential amplifier in Fig. 2, which consumes about $15\mu W$ each. During a synaptic event, the dominant currents are flowing through LRS memristors. The currents flowing through the ladder branches are little fractions of these currents, and may affect also the ratio E_{LRS}/E_{SOP}. In our 4×4 crossbar system, when setting all memristors to LRS and feeding all columns with $T_P = 100ns$ width stimulation pulses at a rate of one pulse every $360ns$, we obtain an average current consumption of $49.52\mu A$ at $3.3V$ power supply. This corresponds to 16 (LRS) synaptic operations (SOP) every $360ns$. Thus, the overall effective energy per (LRS) synaptic operation E_{SOP} is given by[1]

$$E_{SOP} = \frac{1}{16} \times 49.52\mu A \times 3.3V \times 360ns = 3.7pJ \tag{7}$$

The breakdown of the $49.52\mu A$ average current consumption is as follows: $29.49\mu A$ (59.6%) is consumed by the memristors ($1.84\mu A$ by each of the 16 LRS memristors), $17.6\mu A$ (35.58%) by the differential amplifiers driving them, $2.24\mu A$ (4.5%) by the attenuator circuit, and $170nA$ (0.3%) by the neurons. Note that the energy dissipated by an individual LRS memristor generating a $\delta q_{memr} = 0.68pC$ charge packet is $E_{LRS} = 3.3V \times 0.68pC = 2.24pJ$, about 60% of the energy in eq. (7). The other 40% are contributed by the corresponding share of the rest of the circuitry. For scaled-up systems, in which the common circuitry is shared by more memristors, the resulting E_{SOP} value should slowly

[1]The E_{SOP} figure of merit is the inverse of another popular figure of merit used many times in neural processing systems, which is the "*number of synaptic operations per second and per watt*".

approach the baseline of $E_{LRS} = 2.24pJ$ (or even less if LRS memristors are sparse). However, if scaling up aggressively, the combined differential amplifier in Fig. 2 and transistor M_h may need to be redesigned for properly handling larger currents.

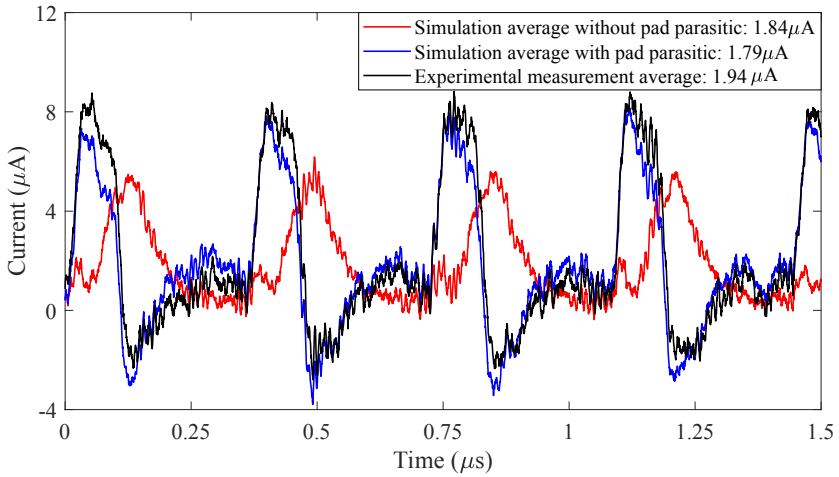


Figure 18: Currents flowing through one single LRS memristor, stimulated with $100ns$ pulses at a rate of one pulse every $360ns$.

The energy figures mentioned above were obtained by simulations, as we could not measure experimentally the detailed breakdown of the current consumption of all sub-circuits. However, the current consumption that we could measure experimentally was below 10% difference with respect to the simulated one. Additionally, our setup allowed us to measure precisely the average current consumed by one single LRS memristor because both of its terminals were accessible in our hybrid multi-chip architecture. Fig. 18 shows three current traces flowing through an LRS memristor. The black trace corresponds to an experimental measurement, yielding an average memristor current of $1.94\mu A$. The red trace corresponds to the simulation mentioned above, with an average of $1.84\mu A$. The black and red traces are quite different, although the average is very similar (5% difference). This is because in the experimental setup, the top plate of the memristors is connected to two chip pads, PCB traces, and the oscilloscope probe, thus adding a large parasitic capacitance. This makes this node to move much slower than the bottom plate making the polarity of the memristor change sign, as well as the current, as can be seen in the black trace in Fig. 18. By adding in the simulation an extra $8pF$ capacitor to the top plate node, the blue trace in Fig. 18 is obtained, which is almost identical to the experimental one, having an average of $1.79\mu A$. Therefore, a full monolithic realization would follow the red trace. However, the presence of the top plate very large parasitic capacitance in the experimental setup is not affecting dramatically the average current, nor the average power consumption.[2]

For a fully integrated on-chip system, additional communication circuitry

[2]This is so because the top plate node is a virtual ground node. In this case, a capacitor connected to it will always return whatever charge it absorbs temporarily, and therefore should not alter the total charge packets travelling through this node.

would be required to send spikes in and out, or to communicate spikes between on-chip computing crossbars. The energy consumption of such communication circuits is not considered in eq. (7). Table 5 shows a comparison with some neural processing chips, spiking and non-spiking, using RRAM or SRAM for weight storage, that have been reported recently.

Table 5: Comparison with some reported state-of-the-art neural processing chip systems

	TrueNorth [23]	Loihi [24]	LETI [9]	Yao [5]	This work
Technology	28nm	14nm	130nm	130nm	180nm
Coding	Spike	Spike	Spike	Formal	Spike
Weight storage	SRAM	SRAM	RRAM	RRAM	RRAM
E_{SOP}	27pJ	105pJ*	180pJ	91pJ**	4pJ*,***

* *Energy consumption of communication circuits not included.*
whiteo ** *Most of energy consumption is due to peripheral analog-to-digital converters.*
white.*** *All memristors are assumed to be at minimum resistance (LRS), thus consuming maximum power.*

9 Conclusions

In this paper we have shown the successful experimental operation of a small SNN that used a 4×4 1T1R memristor crossbar as synapses together with CMOS analog neurons. We also used one compact MOS ladder-based current-splitter circuit per neuron to aggressively downscale the memristor-domain micro-amp current levels to the required analog CMOS neuron-domain current levels. The SNN was assembled using three separate chips. The first chip provided individual novel Ti/C/Au top-plate memristors with low set/reset voltage while presenting high OFF resistance. The second chip was fabricated in a standard 180nm CMOS technology and provided the NMOS selector transistors required for all 1T1R synapses. The third chip, fabricated in the same 180nm CMOS technology, provided the post-synaptic circuitry, including the current attenuator circuits and the neuron circuits. These three chips interacted with a custom PCB and an FPGA-PCB. The custom PCB provided all the analog biases, which were independently adjustable, and the pre-synaptic stimulation pulses, while the FPGA-PCB digitally controlled all the switches and multiplexers on the custom PCB. The system was used to showcase two learning scenarios. One was based on one-shot Winner-Takes-All training, while the other implemented Stochastic-Binary STDP. Successful operation was demonstrated in both scenarios. The setup is clearly very useful in that it facilitates experimentation with new custom-made memristors. Energy measurements reveal this approach as highly promising for ultra-low power systems. Although the hardware example cases shown are small-size from the computational point-of-view, they are capable of performing computations, such as stochastic-binary STDP, which have been demonstrated previously capable of solving much larger scale computing systems [12].

JAF, BLB, and TSG conceived the CMOS chips and overall system architecture, JAF did the actual chip and PCB designs, and performed all tests. SR, SH, and DI designed, manufactured, and characterized the memristors. All authors contributed to writing the manuscript.

The authors declare that they have no competing interests.

This work was funded by EU H2020 grants 824164 HERMES and 871371 MEM-SCALES, and by Spanish grants from the Ministry of Economy and Competitivity TEC2015- 63884-C2-1-P (COGNET) and PID2019-105556GB-C31 (NANOMIND) (with support from the European Regional Development Fund).

References

[1] Z.-R. Wang, Y. Li, Y.-T. Su, Y.-X. Zhou, L. Cheng, T.-C. Chang, K.-H. Xue, S. M. Sze, and X.-S. Miao, "Efficient implementation of Boolean and full-adder functions with

1T1R RRAMs for beyond von Neumann in-memory computing," *IEEE Transactions on Electron Devices*, vol. 65, no. 10, pp. 4659–4666, 2018.

[2] B. Q. Le, A. Grossi, E. Vianello, T. Wu, G. Lama, E. Beigne, H.-S. P. Wong, and S. Mitra, "Resistive RAM With Multiple Bits Per Cell: Array-Level Demonstration of 3 Bits Per Cell," *IEEE Transactions on Electron Devices*, vol. 66, no. 1, pp. 641–646, 2019.

[3] E. Covi, S. Brivio, A. Serb, T. Prodromakis, M. Fanciulli, and S. Spiga, "Analog memristive synapse in spiking networks implementing unsupervised learning," *Frontiers in neuroscience*, vol. 10, p. 482, 2016.

[4] A. Serb, J. Bill, A. Khiat, R. Berdan, R. Legenstein, and T. Prodromakis, "Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses," *Nature communications*, vol. 7, no. 1, pp. 1–9, 2016.

[5] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, "Fully hardware-implemented memristor convolutional neural network," *Nature*, vol. 577, no. 7792, pp. 641–646, 2020.

[6] Z. Liu, J. Tang, B. Gao, P. Yao, X. Li, D. Liu, Y. Zhou, H. Qian, B. Hong, and H. Wu, "Neural signal analysis with memristor arrays towards high-efficiency brain–machine interfaces," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[7] C. Mohan, L. A. Camuñas-Mesa, M. José, E. Vianello, T. Serrano-Gotarredona, and B. Linares-Barranco, "Neuromorphic Low-Power Inference on Memristive Crossbars With On-Chip Offset Calibration," *IEEE Access*, vol. 9, pp. 38 043–38 061, 2021.

[8] M. Payvand, M. V. Nair, L. K. Müller, and G. Indiveri, "A neuromorphic systems approach to in-memory computing with non-ideal memristive devices: From mitigation to exploitation," *Faraday Discussions*, vol. 213, pp. 487–510, 2019.

[9] A. Valentian, F. Rummens, E. Vianello, T. Mesquida, C. L.-M. de Boissac, O. Bichler, and C. Reita, "Fully Integrated Spiking Neural Network with Analog Neurons and RRAM Synapses," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 14.3.1–14.3.4.

[10] K. Bult and G. Geelen, "An inherently linear and compact MOST-only current division technique," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 12, pp. 1730–1735, 1992.

[11] M. Suri, D. Querlioz, O. Bichler, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Bio-Inspired Stochastic Computing Using Binary CBRAM Synapses," *IEEE Transactions on Electron Devices*, vol. 60, no. 7, pp. 2402–2409, 2013.

[12] A. Yousefzadeh, E. Stromatias, M. Soto, T. Serrano-Gotarredona, and B. Linares-Barranco, "On practical issues for stochastic STDP hardware with 1-bit synaptic weights," *Frontiers in neuroscience*, vol. 12, p. 665, 2018.

[13] M. Payvand, Y. Demirag, T. Dalgaty, E. Vianello, and G. Indiveri, "Analog Weight Updates with Compliance Current Modulation of Binary ReRAMs for On-Chip Learning," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–5.

[14] B. Linares-Barranco, T. Serrano-Gotarredona, L. A. Camuñas-Mesa, J. A. Perez-Carrasco, C. Zamarreno-Ramos, and T. Masquelier, "On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex," *Frontiers in neuroscience*, vol. 5, p. 26, 2011.

[15] A. Sebastian, M. Le Gallo, G. W. Burr, S. Kim, M. BrightSky, and E. Eleftheriou, "Tutorial: Brain-inspired computing using phase-change memory devices," *Journal of Applied Physics*, vol. 124, no. 11, p. 111101, 2018.

[16] Q. Liu, B. Gao, P. Yao, D. Wu, J. Chen, Y. Pang, W. Zhang, Y. Liao, C.-X. Xue, W.-H. Chen, J. Tang, Y. Wang, M.-F. Chang, H. Qian, and H. Wu, "A Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 500–502.

[17] B. Linares-Barranco and T. Serrano-Gotarredona, "On the design and characterization of femtoampere current-mode circuits," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 8, pp. 1353–1363, 2003.

[18] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, and G. Indiveri, "A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128k synapses," *Frontiers in neuroscience*, vol. 9, p. 141, 2015.

[19] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. Van Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Häfliger, S. Renaud *et al.*, "Neuromorphic silicon neuron circuits," *Frontiers in neuroscience*, vol. 5, p. 73, 2011.

[20] H. Y. Lee, P. S. Chen, T. Y. Wu, Y. S. Chen, C. C. Wang, P. J. Tzeng, C. H. Lin, F. Chen, C. H. Lien, and M.-J. Tsai, "Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO_2 based RRAM," in *2008 IEEE International Electron Devices Meeting*, 2008, pp. 1–4.

[21] K. G. Young-Fisher, G. Bersuker, B. Butcher, A. Padovani, L. Larcher, D. Veksler, and D. C. Gilmer, "Leakage current-forming voltage relation and oxygen gettering in HfO_x RRAM devices," *IEEE electron device letters*, vol. 34, no. 6, pp. 750–752, 2013.

[22] G.-q. Bi and M.-m. Poo, "Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type," *Journal of neuroscience*, vol. 18, no. 24, pp. 10 464–10 472, 1998.

[23] 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, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.

[24] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C.-K. Lin, A. Lines, R. Liu, D. Mathaikutty, S. McCoy, A. Paul, J. Tse, G. Venkataramanan, Y.-H. Weng, A. Wild, Y. Yang, and H. Wang, "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.