Fully-Integrated Spiking Neural Network using SiO_X-based RRAM as synaptic device

Amir REGEV[1], Alessandro BRICALLI[1], Giuseppe PICCOLBONI[1], Alexandre VALENTIAN[2], Thomas MESQUIDA[2], Gabriel MOLAS[2], Jean-François NODIN[2]

[1]Weebit Nano, Hod Hasharon, Israel, [2]CEA-Leti, Grenoble, France

***Abstract*— This paper presents, to the best of the authors' knowledge, the first complete integration of a Spiking Neural Network combining analog neurons and SiO_x-based resistive memory (RRAM) synapses. The implemented topology is a perceptron, and the circuit is aimed at performing MNIST digits classification. An existing framework was adapted for off-line learning and weight quantization, and the network was later converted into its spiking equivalent. The test chip, fabricated in 130 nm CMOS, shows a classification accuracy of 82%, with a 180 pJ energy dissipation per spike.**

Keywords*— *silicon oxide, resistive memories, spiking neural network, neuromorphic computing

I. Introduction

In recent years, Artificial Neural Networks (ANNs) ensured consistent improvements in the field of Artificial Intelligence (AI), allowing to reach unprecedent accuracy in machine learning tasks such as speech recognition and image classification. The high degree of precision of such networks came at the cost of enormous energy consumption, which is mainly related to data movement between the computing cores and the memory modules. This caused a strong push in the development of circuit architectures where data movement is reduced to a minimum, by physically integrating dense, low-power and non-volatile memories close to the computing elements. Filamentary resistive memories (RRAMs) represent a great candidate for this kind of application.

Although ANN implementations using classical formal coding and RRAM devices have recently made significant progress [1, 2], new solutions more inspired by the human brain have attracted a lot of attention, thanks to the possibilities they offer in terms of highly reduced power consumption. Spiking Neural Networks (SNNs, sometimes referred to as the 3rd generation neural networks) and Neuromorphic Computing represent a promising approach for the implementation of highly-efficient systems [3]. Differently from classical ANNs, inside a SNN neurons communicate sending spikes to each other and the memory elements (synapses) are close to the computing elements (neurons), thus requiring a shift of architectural paradigm with respect to the common von Neumann machine. This work presents the first complete integration of a fully-connected SNN where the synaptic weights are implemented with SiO_x-based resistive memories.

II. Device Technology

Silicon oxide has proved to be particularly interesting as a dielectric material for resistive memory technology given its good resistive switching properties, as well as the advantages it ensures in terms of manufacturability [4, 5]. Fig. 1.a shows a SEM cross-section of the SiO_x-based memory cell integrated on top of a 130 nm CMOS technology. The device stack, shown in the inset of Fig. 1.a, Is composed of a TiN bottom-electrode (BE), an SiO_x active layer (AL) and a Ti top-electrode (TE). The Ti TE is used as an oxygen-scavenging layer. The single resistive memory cell is of 1 transistor-1 resistor (1T1R) type, i.e. the selection element in the array is a transistor. In order to be able to properly write and erase the device, an electroforming procedure is generally needed, where a conductive filament is first created between the top and bottom metallic electrodes. Fig. 1.b shows the distributions of resistances in a SiO_x memory array before and after the forming procedure. In this work, the pristine resistance was used as high-resistive state (HRS) for the cells inside the neural network.

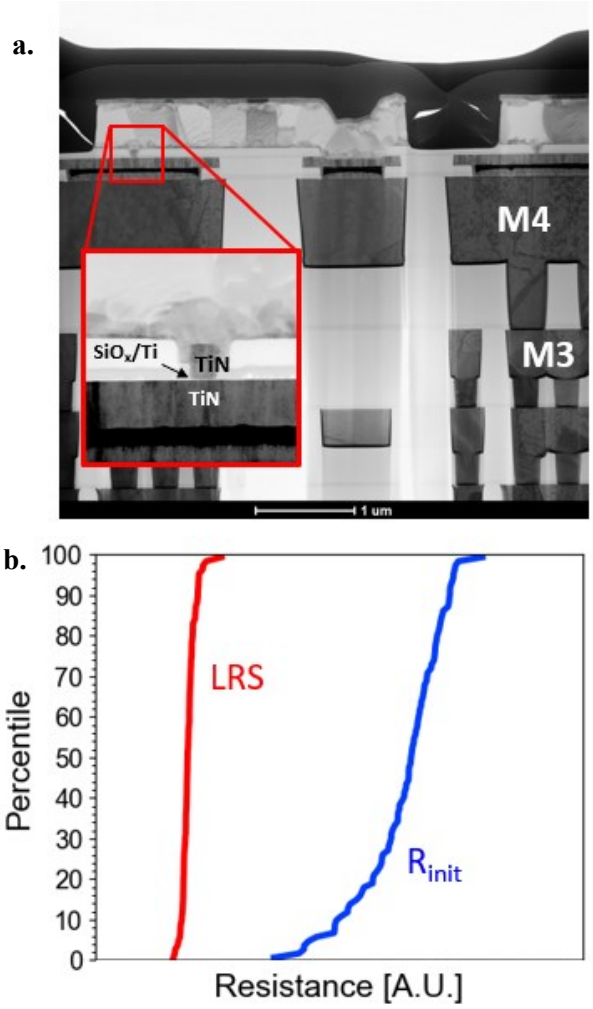


Figure 1. (a) SEM cross-section of the RRAM cell integrated on top of the 130 nm CMOS; (b) Low resistance state and pristine resistance distributions of a SiO_x-based array.

III. Network Training

SNNs are notoriously harder to train using supervised methods compared to their non-spiking counterparts. Although unsupervised biological learning methods exist, such as *Spike Timing Dependent Plasticity* (STDP), those

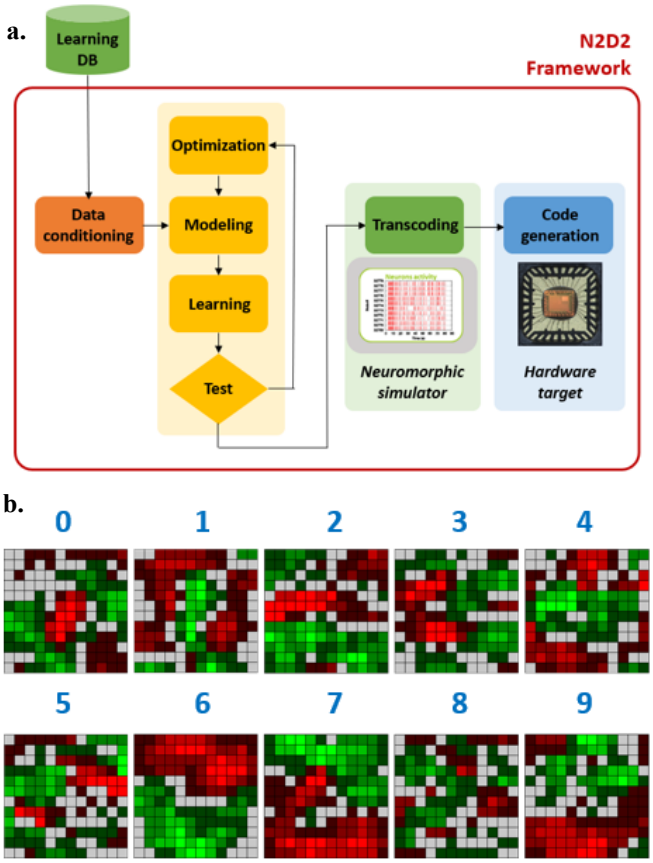


Figure 2. (a) Schematic workflow of the training and conversion phase in the N2D2 framework; (b) Synaptic weights obtained during training for the 10 digits.

typically provide worse performance compared to supervised training models. In our work, the learning task is executed off-line in the classical domain, using the common backpropagation algorithm based on gradient descent. The obtained network is later converted into a quantized, spike-based network (Fig. 2.a). For both learning and conversion, we used N2D2 [6], which is a public deep learning framework integrating transcoding of spiking DNNs, simulation and dedicated hardware code generation. The rate-based transcoding principle that we use was first partially formalized for Leaky-Integrate & Fire (LIF) neurons in [7]. Here we improve and extend it for simple Integrate & Fire (IF) neurons.

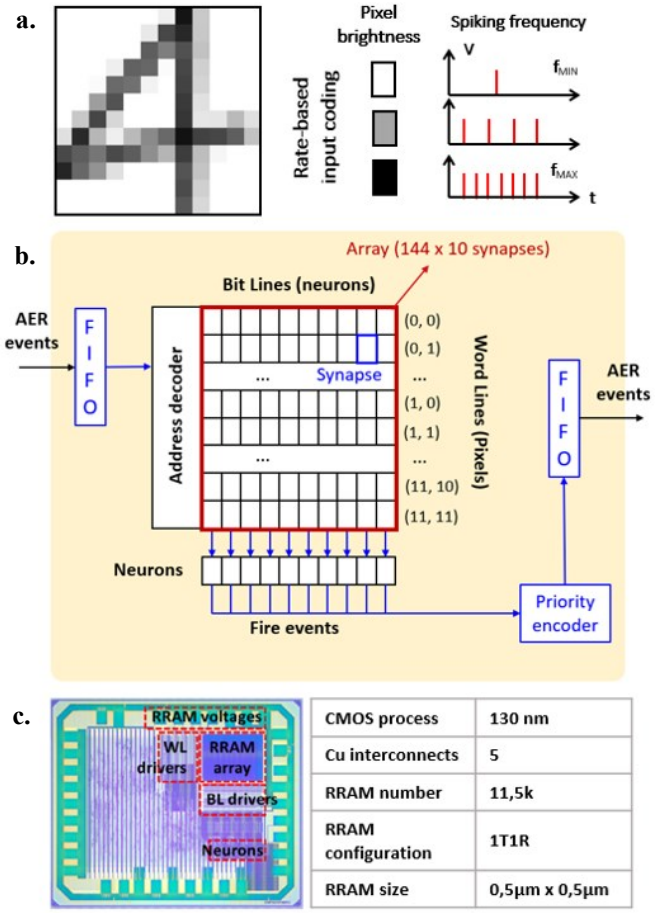
The basic operation in classical coding is the Multiply-and-Accumulate (MAC) between the input x_i and the synaptic weight w_{ij}:

$$y_i = h\left(\sum_i x_i w_{ij}\right)$$

Omitting the non-linear function in the previous equation, we can convert it in a rate-based one using an IF neuron model with a threshold x_{th}:

$$\frac{n_j}{T_{acc}} = \left|\frac{\sum_i n_i w_{ij}}{x_{th}}\right| \frac{1}{T_{acc}}$$

Where n_i and n_j are the number of spikes at input x_i and output y_j respectively, over an accumulation period of T_{acc}. In this equation, spikes can carry a sign, and therefore n_i and n_j can be either positive or negative. In this case, the neuron has



CMOS process	130 nm
Cu interconnects	5
RRAM number	11,5k
RRAM configuration	1T1R
RRAM size	0,5µm x 0,5µm

Figure 3. (a) Schematic representation of the conversion scheme of each pixel of the input image into its spiking counterpart: (b) Schematic representation of the circuit architecture; (c) Chip micrograph.

both a positive threshold ($x_{th+}>0$) and a negative one ($x_{th-}<0$). When one of the two thresholds is reached, we assume the integration is reset to its value minus $x_{th}\, sign(n_j)$ and not to 0. The chosen activation function is the hyperbolic tangent (*tanh*) function, which, for inference only, can be approximated with a simple saturation function h_{sat}:

$$h_{sat}(x) = \begin{cases} -1 & x < -1 \\ x & -1 \le x \le 1 \\ 1 & x > 1 \end{cases}$$

In this case, in order to ensure mathematical equivalence with the *tanh* function, the IF neuron model can be written as:

$$\frac{n_j}{T_{acc}} \approx \frac{1}{T_R} h_{sat}\left(\frac{\sum_i n_i w_{ij}}{\left|x_{th}\, sign(n_j)\right| T_{acc}}\right)$$

Where T_R is the refractory period of the neuron. Finally, given that in spike coding there is a temporal dimension, it is necessary to define when the neural network has received enough information to make a decision. For this, the ΔS termination parameter was introduced, which is the difference between the most spiking output neuron and the second most spiking one and determines when the recognition task is terminated. The synaptic weights obtained during the training phase are shown in Fig. 2.b for each digit, where the green pixels represent the potentiating synapses, while the red ones represent the inhibitory synapses.

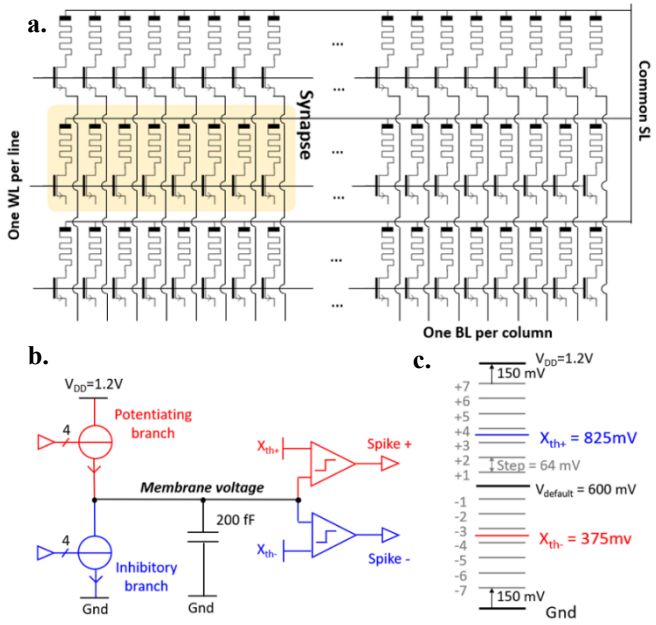


Figure 4: (a) Schematic representation of the synaptic array; (b) Schematic representation of the analog neuron; (c) Possible values of the voltage on the MOM capacitor.

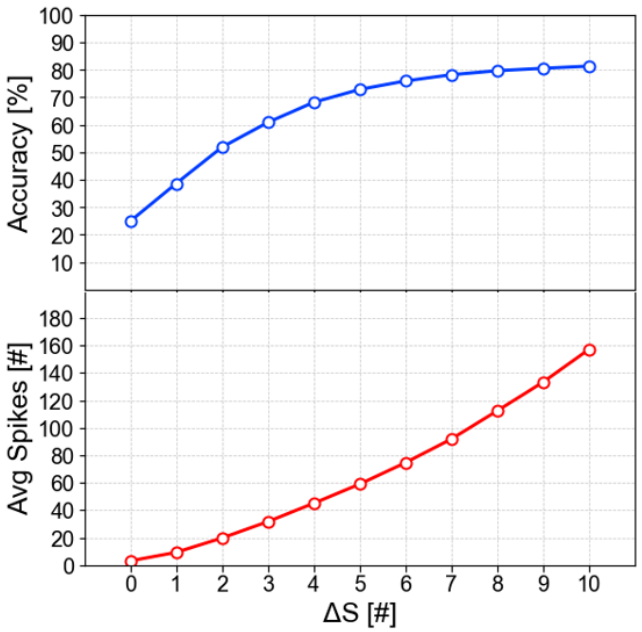


Figure 5: Accuracy and average number of incoming spikes as a function of the termination parameter ΔS.

IV. Circuit Architecture

The topology of the neural network is fully-connected, with a single-layer of analog output neurons. Since the goal is to perform classification of the 10 digits of the MNIST dataset, 10 analog neurons were integrated. When performing the recognition task, the image corresponding to the digit must be converted into a spike-based input (Fig. 3.a): to this purpose, each pixel is encoded as a value in a 256-levels greyscale, and the greyscale value is then converted into a corresponding input-spike frequency. The complete architecture of the circuit is shown in Fig. 3.b. The addresses of the spiking pixels are communicated through an SPI interface in the form of an Address Event Representation (AER) message and stored into a FIFO. The input AER message is the (*x, y*) position of the spiking pixel. The pixel addresses are decoded for reading the corresponding Word Line (WL) of the RRAM array. The array is organized in such a way that a single WL read feeds the ten neurons with inputs weighted by the resistive synapses: in other words, a single pixel events is seen by all the neurons at the same time, thus making the network fully connected. The currents are integrated by the output neurons, which eventually generate spike events as outputs when a threshold in their membrane potential is reached. Output spikes are read through the same SPI interface. A chip micrograph is also illustrated in Fig. 3.c.

A. Synapse Design

Synapses are implemented with Single Level Cell (SLC) RRAM [8], *i.e.* only considering the low and high resistance levels. As discussed in Sec. II, the RRAM cell has a 1T1R structure, with an access transistor per cell (Fig. 4.a). The representation of multiple level weights is achieved by placing several RRAM cells in parallel. Synaptic quantization experiments done on the learning framework N2D2 showed that values ranging from -4 to +4 allow to achieve a good compromise between classification accuracy and number of RRAM devices. For the representation of negative synaptic weights, the Sign bit could have been encoded using RRAMs as well: however, since a fault-tolerant triple redundancy would have been needed, it was preferred to use 4 additional RRAMs for implementing the negative weights. A single synapse is therefore composed of 8 RRAM cells, which represents an equivalent precision of 3 bits. Synapses are arranged in a matrix for sharing WL, Source Line (SL) and Bit Line (BL) drivers. In the future, multilevel RRAM operation (for example by tuning LRS using current compliance) may be be used for analog representation of weights, allowing for a significant optimization of achievable synaptic density.

B. Neuron Design

The design of the *Integrate and Fire* (IF) analog neuron was guided by the need for mathematical equivalence with the *tanh* activation function employed in the offline network training phase, with the following specifications:

- A stimulation with a synaptic weight equal to ±4 must generate a spike
- Neurons must generate positive and negative spikes
- Neurons must have a refractory period, during which they cannot emit spikes, but must continue to integrate

The structure of the analog neuron was built to closely simulate a biological neuron, with a membrane potential and a firing threshold. This is illustrated in Fig. 4.b: neurons are architected around a MOM 200 fF capacitor, and two comparators are used to compare the voltage level on the capacitor to a positive and a negative threshold. Since RRAMs must be read with a voltage drop limited to 100 mV between their terminals, in order to prevent unwanted transition of the devices to LRS the currents cannot be directly integrated by the neurons: instead, they are copied by current injectors. In the figure, the whole branch relative to excitation and positive output spikes is represented in red,

	Science 2014 [3]	Micro 2018 [9]	VLSI 2018 [2]		This work
Technology	28nm	14nm	40nm	180nm	130nm
Coding	Spike	Spike	Formal	Formal	Spike
Weight storage	SRAM	SRAM	RRAM	RRAM	RRAM
Synapses	256M	130M	4M	2M	11.5K
Synapses/mm²	195K	2000K	1480K	160K	16K
Power	63mW	-	9.9mW	15.8mW	1.5mW
Energy/syn. event	27pJ	105pJ	-	-	180pJ
Accuracy	96%	-	90%		82%

Table 1: comparison of the main figures of merit of our SNN with other state-of-the-art accelerators.

while the branch relative to inhibition and negative output spikes is represented in blue. Finally, the voltage levels attainable in the capacitor are illustrated in Fig. 4.c. The default voltage value across the capacitor is 600 mV, while positive and negative thresholds are placed at 825 and 375 mV respectively.

V. Results Discussion

Classification accuracy on the 10k test images of the MNIST database is measured at 82%, which must be compared with the accuracy obtained from ideal simulations of 88%, limited by the simple network topology (1 layer, 10 output neurons). Fig. 5 plots the accuracy as a function of the difference between the most active and the second most active neuron of the network, ΔS. By changing the ΔS parameter, it is possible to trade off accuracy against spiking activity (i.e., power consumption). The energy dissipation per synaptic event is equal to 180 pJ: however, most of the energy is dissipated by the SPI interface, and it could be reduced by optimization of the communication protocol. Measurements show that an image classification requires around 155 input spikes on average (for ΔS = 10): when compared to an equivalent formal coding MAC operation in the same technology node, we estimate this leads to a 5x gain in terms of energy consumption at synapse and neuron level, and can become as high as 30x in case weight movement between memory and the processing unit is considered. In fact, it is well known that the energy cost of moving data can be orders of magnitude higher than the cost of computation [10]: for example, it was estimated that, in many typical workloads for consumer applications, around 60% of the total system energy is spent on data movement [11].

Finally, Table 1 summarizes the main characteristics of our SNN, comparing it to other implementations from the state of the art. In particular, it is worth noting how, compared to the results in [2], which uses classical coding and MAC operations, we can perform a similar task with strongly reduced power consumption and far less synapses. Although there is a difference in accuracy, it should also be considered that we can achieve an accuracy of 82% in a single-layer network, while the results in [2] were obtained on a network with 3 middle layers. Therefore, our technology demonstrates great potential for the implementation of low-power AI tasks.

VI. Conclusions

This work demonstrates the first fully-integrated Spiking Neural Network combining analog neurons and SiO_x-based RRAM synapses, aimed at demonstrating MNIST digits recognition. The network is trained off-line with backpropagation and then converted into a spiking network. The 130nm test chip shows an energy consumption of 180 pJ per spike, limited by the communication protocol.

References

[1] S. Ambrogio, P. Narayanan, H. Tsai, R.M. Shelby, I. Boybat, C. di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. P. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, G. W. Burr, "Equivalent-accuracy accelerated neural-network training using analogue memory", *Nature*, 2018, vol.558, no.778, pp.60.

[2] M. Reiji, K. Kazuyuki, H. Yuriko, N. Masayoshi, O. Takashi, S. Hitoshi, "A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture", VLSI Technology, 2018, pp. 175-176.

[3] 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, "A million spiking-neuron integrated circuit with a scalable communication network and interface", 2014, *Science*, pp. 668-673.

[4] A. Bricalli ; E. Ambrosi ; M. Laudato ; M. Maestro ; R. Rodriguez ; D. Ielmini, "SiO_x-based resistive switching memory (RRAM) for crossbar storage/select elements with high on/off ratio", IEDM, 2016

[5] 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, "Silicon Oxide (SiO_x): A Promising Material for Resistance Switching?", Advanced Materials, 2018, Volume30, Issue43

[6] https://github.com/CEA-LIST/N2D2

[7] J. A. Pérez-Carrasco, B. Zhao, C. Serrano, B. Acha, T. Serrano-Gotarredona, S. Chen, B. Linares-Barranco, "Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing--application to feedforward ConvNets", IEEE transactions on pattern analysis and machine intelligence 35, no. 11 (2013): 2706-2719

[8] A. Grossi, E. Vianello, C. Zambelli, P. Royer, J.-P. Noel, B. Giraud, L. Perniola, P. Olivo, E. Nowak, "Experimental Investigation of 4-kb RRAM Arrays Programming Conditions Suitable for TCAM", VLSI Technology, 2018, Volume: 26 , Issue: 12

[9] M. Davies, N. Srinivasa, T.H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, and Y. Liao, "Loihi: A neuromorphic manycore processor with on-chip learning". 2018, IEEE Micro, 38(1), pp.82-99.

[10] S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, and D. Glasco, "GPUs and the Future of Parallel Computing," IEEE Micro, 2011.

[11] A. Boroumand, S. Ghose, Y. Kim, R. Ausavarungnirun, E. Shiu, R. Thakur, D. Kim, A. Kuusela, A. Knies, P. Ranganathan, O. Mutlu, "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks", Proceedings of the 23rd International Conference on Architectural Support for Programming languages and Operational Systems (ASPLOS), Wiliamsburg, VA, USA, March 2018.