Low-energy inference machine with multilevel HfO_2 RRAM arrays

V. Milo[1], C. Zambelli[2], P. Olivo[2], E. Pérez[3], O. G. Ossorio[4], Ch. Wenger[3,5] and D. Ielmini[1*]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano and IU.NET, 20133 Milano, Italy

[2] Dipartimento di Ingegneria, Università degli Studi di Ferrara, 44121 Ferrara, Italy

[3] IHP-Leibniz-Institut für innovative Mikroelektronik, Im Technologiepark 25, 15236 Frankfurt, Germany

[4] Dpto. Electricidad y Electrónica, Universidad de Valladolid, Paseo de Belén 15, 47011 Valladolid, Spain

[5] Brandenburg Medical School Theodor Fontane, Fehrbelliner Strasse 38, 16816 Neuruppin, Germany

*email: daniele.ielmini@polimi.it

***Abstract*—Recently, artificial intelligence reached impressive milestones in many machine learning tasks such as the recognition of faces, objects, and speech. These achievements have been mostly demonstrated in software running on high-performance computers, such as the graphics processing unit (GPU) or the tensor processing unit (TPU). Novel hardware with in-memory processing is however more promising in view of the reduced latency and the improved energy efficiency. In this scenario, emerging memory technologies such as phase change memory (PCM) and resistive switching memory (RRAM), have been proposed for hardware accelerators of both learning and inference tasks. In this work, a multilevel 4kbit RRAM array is used to implement a 2-layer feedforward neural network trained with the MNIST dataset. The performance of the network in the inference mode is compared with recently proposed implementations using the same image dataset demonstrating the higher energy efficiency of our hardware, thanks to low current operation and an innovative multilevel programming scheme. These results support RRAM technology for in-memory hardware accelerators of machine learning.**

Keywords: resistive switching memory (RRAM); artificial intelligence; machine learning; in-memory computing; neural network; backpropagation; energy efficiency.

I. Introduction

Machine learning has made extensive progress in the last ten years, thanks to the availability of large training datasets and the maturity of high-performance computers such as the graphics processing unit (GPU) [1] and the tensor processing unit (TPU) [2]. Deep learning networks [3] operated with this specialized hardware have been shown to outperform the human performance in image/face recognition [4]. Operating these networks, however, generally requires a large power consumption and computational time because of the von Neumann bottleneck affecting all conventional processors [5]. New computing architectures capable of combining high energy efficiency, high performance and high density of synaptic weights are thus receiving strong research efforts.

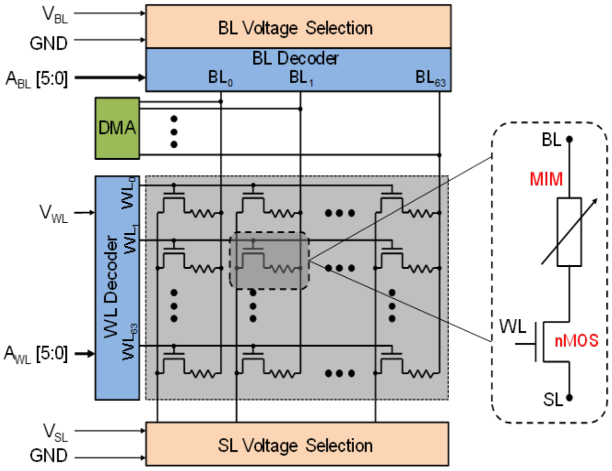

Fig. 1. Schematic of 4kbit RRAM array architecture where each cell consists of a 1T1R structure based on serial connection of a NMOS transistor and an Al:HfO_2 RRAM device.

In this scenario, novel memory technologies such as phase change memory (PCM) and resistive switching memory (RRAM) might enable improved energy efficiency and scaling, thanks to in-memory analogue matrix-vector multiplication (MVM) for both training and inference [6]. Hardware machine-learning accelerators using PCM devices as synaptic weights have been demonstrated for image classification of the MNIST dataset of handwritten digits [7], reaching an inference accuracy of 83% [8]. A higher inference accuracy of about 92% with MNIST classification was shown with Ta/HfO_2/Pt RRAM synaptic devices [9]. However, a hardware implementation combining high accuracy and low energy operation has not been demonstrated yet.

In this work, we demonstrate a hardware neural network implemented in a 4kbit array of Al:HfO_2 RRAM devices with one-transistor/one-resistor (1T1R) structure [10]. Synaptic weights are stored by using an efficient multilevel programming scheme with 5 conductance states. The 1T1R array is

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 648635) and from the Università degli Studi di Ferrara under the Bando per il finanziamento della ricerca scientifica "Fondo per l'Incentivazione alla Ricerca" (FIR) - 2018. This work was also supported by German Research Foundation (DFG) in the frame of the research group FOR2093.

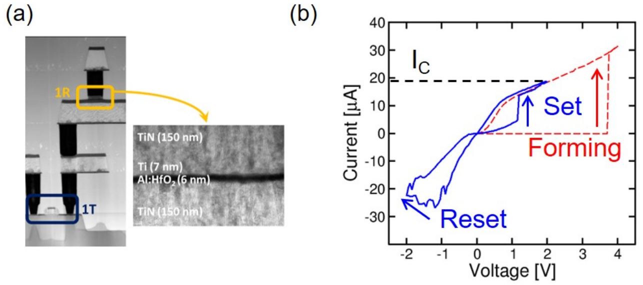

Fig. 2. (a) TEM cross-sectional image of 1T1R cell structure with a detailed description of Al:HfO$_2$ RRAM stack. (b) I-V characteristics of a single 1T1R cell displaying forming, set and reset processes.

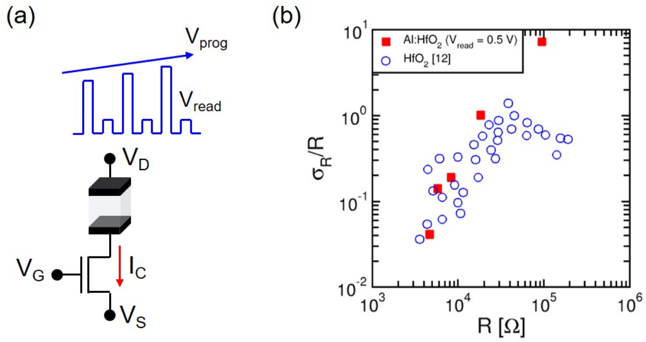

Fig. 3. (a) Schematic representation of multilevel program and verify algorithm (M-ISPVA) used to achieve 5 current levels to program synaptic weights into the array [11]. (b) Median relative variability of resistance as a function of median resistance of 5 target levels for V$_{read}$ = 0.5 V evidencing a good agreement with LRS measurements on HfO$_2$ RRAM cells presented in [12].

then programmed to store the weights of a 2-layer feedforward neural network, and the experimental accuracy of MNIST classification is characterized by simulations. We characterize the classification accuracy before and after annealing at elevated temperature. The array performance is finally compared with recently presented networks, supporting the good energy efficiency of multilevel RRAM in inference mode.

II. RRAM ARRAY STRUCTURE AND OPERATION

Fig. 1 shows the architecture of the 4kbit RRAM array used in this work, which includes 1T1R memory cells consisting of a NMOS transistor manufactured with the IHP's 0.24 μm CMOS technology that is contacted on its drain terminal by a metal-insulator-metal (MIM) element [10]. The 1T1R cells are arranged into a 64x64 matrix and they can be accessed through the wordline (WL) and bitline (BL) decoders. The voltages applied to the WL, BL, and the sourceline (SL) for forming, set, and reset operations are provided by an external test equipment and routed on the array through the voltage selectors and decoders. Also, a Direct Memory Access (DMA) interface connects with the BL decoder to provide the readout current of a selected cell. Fig. 2(a) provides a detailed view of 1T1R cell evidencing the TiN/Ti/Al:HfO$_2$/TiN RRAM stack and the thickness of each layer, while Fig. 2(b) shows the current-voltage ($I - V$) characteristics of a single 1T1R cell before the array integration. After forming process, the reset transition from low resistance state (LRS) to high resistance state (HRS) is triggered by application of a reset voltage of -1 V, whereas the set transition from HRS to LRS occurs at set voltage of about 1 V and it is limited by a compliance current I$_C$ = 20 μA. Tuning resistance in 1T1R cells relies on the multilevel capability of the RRAM technology tightly coupled with a good control of its intrinsic variability [11]. Here, we consider 5 target levels, namely one HRS level (L$_1$) and four different LRS levels (L$_2$-L$_5$). HRS is achieved through the application of the ISPVA approach [11] using the following device parameters: gate voltage V$_G$ = 2.7 V, source voltage V$_S$ from 0 to 3 V and a target readout current I$_{L1}$ = 5 μA. On the other hand, to achieve LRS levels we adopted the multilevel variation of ISPVA algorithm called M-ISPVA algorithm (Fig. 3(a)), where different I$_C$ values are obtained changing V$_G$ from 1 V to 1.6 V and V$_D$ from 0 to 3 V with readout current targets fixed to I$_{L2}$ = 15 μA, I$_{L3}$ = 30 μA, I$_{L4}$ = 45 μA, and I$_{L5}$ = 60 μA, respectively. Also, to maximize the level separation we formed all the cells in the

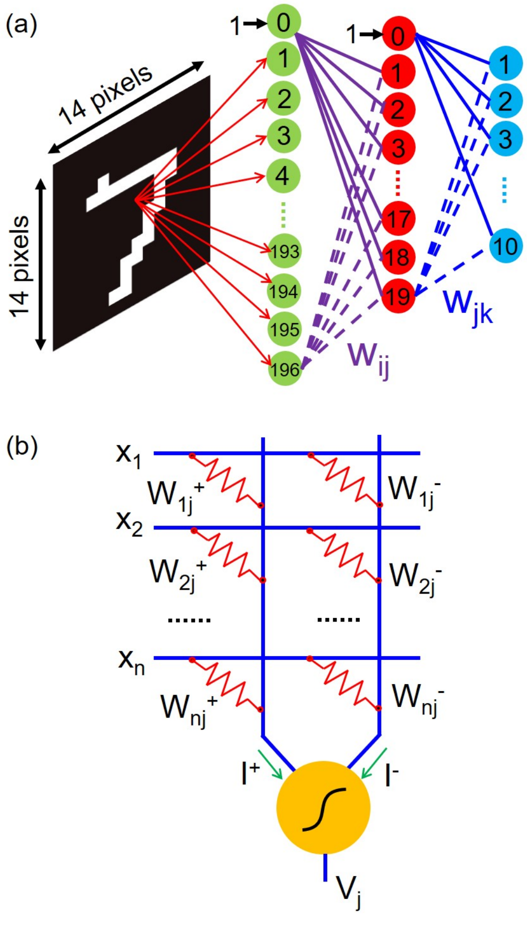

Fig. 4. (a) Sketch of 2-layer feedforward neural network implemented in the 4kbit array where any synaptic weight is achieved by the differential readout of currents activated within 2 1T1R cells (b).

array applying the M-ISPVA scheme tuned for a readout target I_{L5}. This enabled to program 5 resistance levels whose relative variability, which is shown in Fig. 3(b) as a function of median resistance R at V_{read} = 0.5 V, displays a behavior consistent with LRS variability of HfO_2 RRAM cells investigated in [12].

III. Neural network implementation

Using our 4kbit RRAM array, we designed a feedforward neural network as the one depicted in Fig. 4(a). It consists of an input layer with 197 neurons fully connected with a hidden layer comprising 20 neurons which feed in turn each of 10 output neurons. Note that both input layer and hidden layer include a bias neuron which is always on. This network was trained via software according to the backpropagation algorithm [7] submitting a binary version of the 60,000 handwritten digit images of MNIST training dataset downscaled from 28x28 pixels to 14x14 pixels to the input layer. Since weights calculated during training phase can have both negative and positive values, we mapped effective synaptic weights of the neural network into the difference between the conductance W^+ of a synaptic 1T1R device and the conductance W^- of the corresponding 1T1R cell within the reference line associated to input or hidden layer. As shown in Fig. 4(b), the sum of current signals given by dot product between inputs and synaptic weights is collected at the input of any j-th neuron of following layer (in this case the hidden layer) and it is converted through a sigmoidal activation function into a voltage V_j, which in turn becomes the input for weights between hidden and output layer.

After achieving a classification accuracy of about 92% on the 10,000 images of MNIST test dataset using the weight matrix calculated with 64-bit floating point precision, we applied a rounding scheme in software to lower weight precision to 5 levels and programmed the array cells based on new weight matrix with 5-level precision. Fig. 5(a) shows the PDF curves of 5 readout current levels (L_1-L_5) indicating synaptic weights of both network layers in the array, which exhibit some overlap, especially between L_1 and L_2. Note that PDF distributions of 5 levels are shown for V_{read} = 0.5 V because it is the read voltage for which we obtained the lowest experimental variability. In addition, Fig. 5(b) also shows the PDF distributions of the 5 current levels at the same V_{read} measured after 1h-long annealing experiment performed at temperature T = $125°C$, displaying a significant worsening in terms of level separation as a result of higher variability.

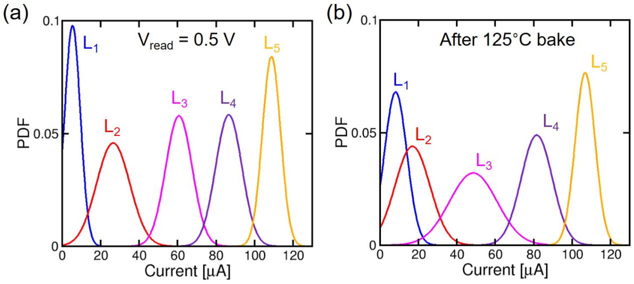

Fig. 5. PDF distributions of 5 experimental readout current levels at V_{read} = 0.5 V (a) before and (b) after an annealing experiment for 1 hour at $125°C$.

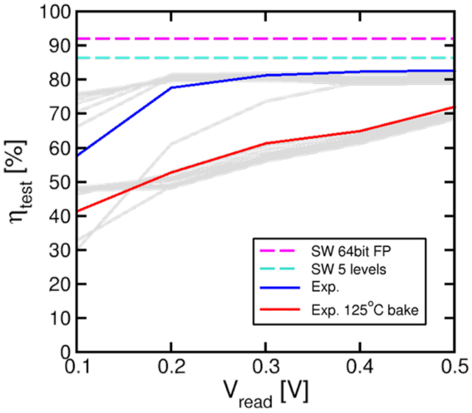

Fig. 6. Evolution of experimental classification accuracy η_{test} as a function of V_{read} for variable activation function slope (gray curves) before and after a high-temperature annealing experiment. These results are also compared with software accuracy values achieved with 64bit floating point precision and 5 ideal levels.

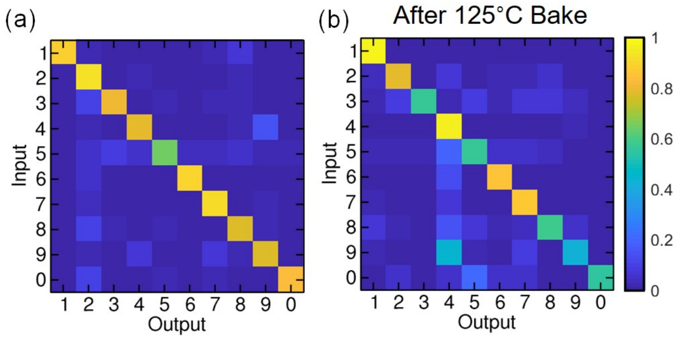

Fig. 7. Color maps evidencing network ability to correctly associate an input MNIST test image with the corresponding class (a) before and (b) after annealing experiment at high temperature.

IV. Classification performance of the network

After programming synaptic devices by 5 current levels, we tested network inference ability by presentation of all the MNIST test images in simulation. Fig. 6 shows the evolution of classification accuracy η_{test} as a function of V_{read} for increasing slope of sigmoidal function of the neurons from 10^4 V/A to 10^5 V/A. Compared to floating point accuracy (92%) and accuracy with 5 software levels with no variability (86.5%), we achieved a maximum experimental classification performance η_{test} = 82.82% at V_{read} = 0.5 V using an activation function slope of $2 \cdot 10^4$ V/A. This result can be attributed to the experimental variability of current levels preventing to capture uniform level separation considered in software simulations, the weight approximation due to the low number of levels, and the presentation of some corrupted digit images because of image downscaling.

TABLE I

SUMMARY AND COMPARISON OF ARRAY-LEVEL HARDWARE NEURAL NETWORKS WITH 1T1R SYNAPTIC DEVICES.

Work	**Device**	**# synapses**	**Input size**	**G$_{max}$**	η_{test}
[8]	PCM	165 k	22x24	22 μS	82.9 %
[9]	RRAM	8 k	8x8	400 μS	91.7 %
This work	RRAM	4 k	14x14	200 μS	82.82 %

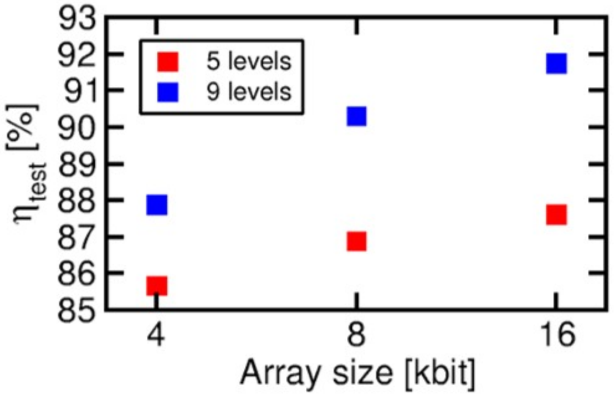

Fig. 8. Projection of network inference performance using 5 levels with lower variability (red dots) and 9 current levels (blue dots) for array sizes of 4 kbit, 8 kbit, and 16 kbit, respectively.

In parallel to this classification study, we investigated network inference capability after annealing at high temperature, achieving a maximum classification accuracy of about 72%, which supports the detrimental impact of increased device variability on classification accuracy. Fig. 7 provides a more detailed description of classification performance of our neural network by confusion matrices showing the probability that each submitted image is correctly classified during inference phase. Note that at network level it means that the output neuron corresponding to submitted input has to generate a voltage higher than all the other output neurons. In particular, Fig. 7(a) shows that the lowest number of correct classifications was obtained for class '5' whereas Fig. 7(b) confirms the decrease of classification capability for many digit classes.

This network implementation was also compared with other recent array-level hardware demonstrations on MNIST dataset. As reported in Table I, [8] achieves a test accuracy very close to performance of our network but using a very large neural network with 2-PCM synapses trained with 22x24 images. On the other hand, [9] reaches a classification accuracy higher than our implementation but using gray-scale 8x8 MNIST images and RRAM devices exhibiting a very linear conductance response and mainly a high maximum conductance value, which suggests a lower energy efficiency than our network.

Finally, Fig. 8 supports that inference capability of our neural network with 5 levels can be improved using larger arrays by reduction of variability of 5 levels or increasing the number of levels from 5 to 9, which leads to achieve a maximum classification accuracy of 91.76% by a 16kbit array. To further improve network performance, combination of high-density arrays and synaptic devices capable of a better multilevel operation is required.

V. CONCLUSIONS

In this work, the implementation of a 2-layer feedforward neural network capable of image classification on MNIST dataset by multilevel programming of a 4kbit 1T1R RRAM array is proposed. Based on results obtained by software simulations of neural network, we mapped a weight matrix with 5-level precision into the array and tested its inference ability achieving a classification accuracy of about 83%. Although other array-level implementations achieved better classification performance, our network enables to combine low energy operation and high inference accuracy. These results are thus seminal for new mixed hardware/software investigations aiming at building compact and low-power hardware accelerators for machine learning tasks.

REFERENCES

[1] A. Coates, B. Huval, T. Wang, D. J. Wu, A. Y. Ng, and B. Catanzaro, "Deep learning with COTS HPC systems," *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, no. 3, pp. 1337–1345, 2013.

[2] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a Tensor Processing Unit^{TM}," *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA)*, pp. 1–12, 2017.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," *IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015.

[5] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nat. Nanotechnol.*, vol. 10, no. 3, pp. 191–194, 2015.

[6] S. Yu, "Neuro-inspired computing with emerging nonvolatile memory," *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, 2018.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element," *IEEE IEDM Tech. Dig.*, pp. 697–700, 2014.

[9] C. Li *et al.*, "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks," *Nat. Commun.*, vol. 9, p. 2385, 2018.

[10] C. Zambelli *et al.*, "Statistical analysis of resistive switching characteristics in ReRAM test arrays," *International Conference on Microelectronic Test Structures (ICMTS)*, pp. 27–31, 2014.

[11] E. Pérez, C. Zambelli, M. K. Mahadevaiah, P. Olivo, and Ch. Wenger, "Towards reliable multi-level operation in RRAM arrays: improving post-algorithm instability and assessing high temperature retention," *Trans. on Electron Devices*, Submitted.

[12] D. Ielmini, "Resistive switching memories based on metal oxides: mechanisms, reliability and scaling," *Semicond. Sci. Technol.*, vol. 31, no. 6, p. 063002, 2016.