OxRAM+OTS optimization for Binarized Neural Network hardware implementation

Joel Minguet Lopez, Tifenn Hirtzlin, Manon Dampfhoffer, Laurent Grenouillet, Lucas Reganaz, Gabriele Navarro, Catherine Carabasse, Elisa Vianello, Thomas Magis, Damien Deleruyelle, et al.

OxRAM+OTS optimization for Binarized Neural Network hardware implementation

J. Minguet Lopez[1], T. Hirtzlin[1], M. Dampfhoffer[2], L. Grenouillet[1], L. Reganaz[1], G. Navarro[1], C. Carabasse[1], E. Vianello[1], T. Magis[1], D.Deleruyelle[4], M. Bocquet[3], J. M. Portal[3], F. Andrieu[1], G. Molas[1]

[1] CEA, LETI, Univ. Grenoble Alpes, 38000 GRENOBLE, France
[2] Univ. Grenoble Alpes, CEA, CNRS, Grenoble INP, INAC-Spintec, 38000 GRENOBLE, France
[3] Aix Marseille Univ, Université de Toulon, CNRS, IM2NP, 13009 MARSEILLE, France
[4] INL CNRS, INSA Lyon, 69621 VILLEURBANNE, France

E-mail: joel.minguetlopez@cea.fr , gabriel.molas@cea.fr

Abstract

Low-power memristive devices embedded on GPUs or CPUs logic core is a very promising non-von-Neumann approach to improve significantly the speed and power consumption of Deep Learning accelerators, enhancing its deployment on embedded systems. Among various non-ideal emerging neuromorphic memory devices, synaptic weight hardware implementation using RRAM memories within 1T1R architectures promises high performance on low precision Binarized Neural Networks (BNN). Taking advantage of the RRAM capabilities and allowing to substantially improve the density thanks to the OTS selector, this work proposes to replace the standard 1T1R architecture by a denser 1S1R crossbar system, where a HfO_2-based OxRAM is co-integrated with a Ge-Se-Sb-N-based OTS. In this context, an extensive experimental study is performed to optimize the 1S1R stack and programming conditions for extended Read Window Margin and endurance characteristics. Focusing on standard machine learning MNIST image recognition task, we perform offline training simulations in order to define the constraints on the devices during the training process. A very promising Bit Error Rate of ~10^{-4} is demonstrated together with 1S1R 10^4 error-free programming endurance characteristics, fulfilling the requirements for the application of interest. Based on this simulation and experimental study, BNN figures of merit (system footprint, amount of weight updates, accuracy and tolerance to errors) are optimized by engineering the amount of learnable parameters of the system. Altogether, an inherent BNN resilience to 1S1R parasitic bit errors is demonstrated.

Keywords: BNN, Resistive RAM, OTS, chalcogenide, crossbar

Outstanding achievements have been done in artificial intelligence over the last years, supported by the raise in maturity of deep learning (DL) accelerators. However, its deployment on embedded systems remains very limited due to excessive energy consumption associated to data transfer between Graphics or Central Processing Units (GPUs or CPUs) logic and memory blocks, commonly referred as von-Neumann bottleneck [1]. Therefore, rethinking the system architecture by logic and memory merging towards in or near-memory computing implementation makes perfect sense. While embedded on crossbar matrices, novel emerging memory technologies utilisation for deep neural network hardware implementation is a very promising path towards beyond von-Neumann architectures [2].

Basic operation in all neural network is the Multiply and Accumulate (MAC) that can be naturally implemented using the Kirchoff law in crossbar arrays [3, 4], where the memory array can store the weight values in an analog fashion [5].

However, these innovative architectures impose severe constraints and requirements on the memory technologies like very high capacity, multi level capabilities with high accuracy on device conductance and low variability, high endurance and low power consumption. Unfortunately, this universal memory does not exist at this stage [6]. In this context, several techniques have been proposed in the literature to mitigate imperfection of current technologies [7-8]. In particular, Resistive Random-Access Memory (RRAM) devices have been widely proposed for synaptic weight implementation [7,9-10]. Nevertheless, RRAM utilisation remains challenging due to device intrinsic variability, which induces parasitic bit errors. Moreover, big efforts have been done on bit error correcting codes and adaptive programming schemes deployment [11-12]. Also, aggressive RRAM programming strategies have been proposed, inducing device variability reduction at the expense of its endurance capabilities [13-14].

In our approach, we focus on low precision Binarized Neural Networks (BNN), which are more to cope with non idealities of the RRAM [15-16]. In this type of neural networks, both synaptic weights and neuron activations are implemented by binary values (+1 or -1) after network training process. Allowing a drastical reduction of the required memory for inference computation while preserving state-ofthe-art performance on image recognition tasks [17-18], BNNs attractivity for hardware implementation is increasing.

Memory-based synaptic weight hardware implementation is commonly performed using 1T1R architectures [19-22], where memory devices are accessed by using individual selecting CMOS transistors. In this context, we propose to replace standard 1T1R architectures by a denser 1S1R system, where the memory is co-integrated in series with a back-end selector. This allows to scale the bitcell size from $40F^2$ to $4F^2$ and leads to an increase of one order of magnitude in the memory density, enhancing the path towards on-chip implementation of complex state of art neural networks' architectures (like YOLOv3 or ResNet-152).

However, RRAM co-integration with a back-end selector in dense crossbar arrays leads to several challenges, including process integration complexity and a tradeoff between programming voltage and leakage current [23], which leads to a critical development of both memory and selector devices [24].

In this work, we propose to co-integrate an RRAM with an Ovonic Threshold Selector (OTS) for synaptic weight analog implementation using single memristor crossbar Arrays [25]. It combines the high performance of RRAM device with a higher crossbar density procured by the OTS backend selector. OTS (1S) devices having been satisfactorily co-integrated with HfO_2-based OxRAM (1R) devices recently, very promising performance have been achieved [26] thanks to both stack and programming conditions optimization [27]. OxRAM+OTS (1S1R) device working principle and switching characteristics are presented, underlining the Read Voltage Margin (RVM) criticality on device reliability [28]. Starting with voltage repartition between OTS and OxRAM devices, the 1S1R programming conditions for the stack of interest are optimized. High Read Voltage Margin with optimized variability is presented, demonstrating an ~10^{-4} very low Bit Error Rate (BER). Then, to explore the benefit of deploying Binarized Neural Networks (BNNs) with 1S1R crossbar arrays and to study the constraints on the devices we performed training simulations on fully connected binarized neural network with one hidden layer of 1024 neurons for an image classification task on the MNIST dataset [29], described in **Fig. 1**. Based on experimental programming and reading endurance characteristics, 1S1R pertinence for BNN neuromorphic hardware deployment is demonstrated and its inference lifetime is predicted. Moreover, by generalizing our BNN architecture, the BNN bit error tolerance is evaluated and guidelines for optimized footprint, minimal amount of synaptic weight updates and BNN performance for MNIST dataset are provided.

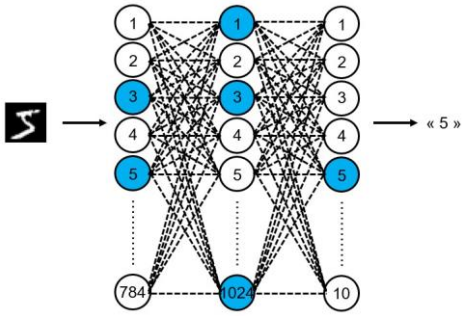

Fig.1. Binarized Neural Network considered in this section. Fully connected neural network with one hidden layer of 1024 neurons for MNIST handwritten recognition.

2. Technological details

OxRAM is co-integrated with OTS back-end selector in memory arrays. A transistor is used as a current limiter. Integrated within TiN and Ti top and bottom electrodes, a 10nm-thick HfO_2-based OxRAM is used as memory device. 10nm-thick Ge-Se-Sb-N-magnetron sputtered alloy, sandwiched between two Carbon electrodes for improved reliability [29], is used as OTS device. **Fig. 2** illustrates TEM cross section for the stack of interest, demonstrating the consistent co-integration of all deposited layers.

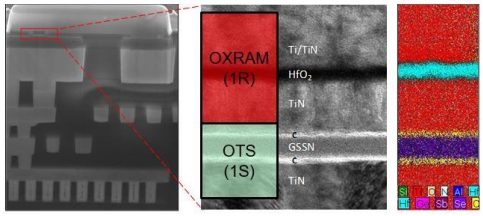

Fig.2. SEM (left), together with TEM (middle) and EDX (right) images for the 1S1R devices of interest, integrated on Metal 4.

3. Results and discussion

3.1 1S1R programming and reading capabilities

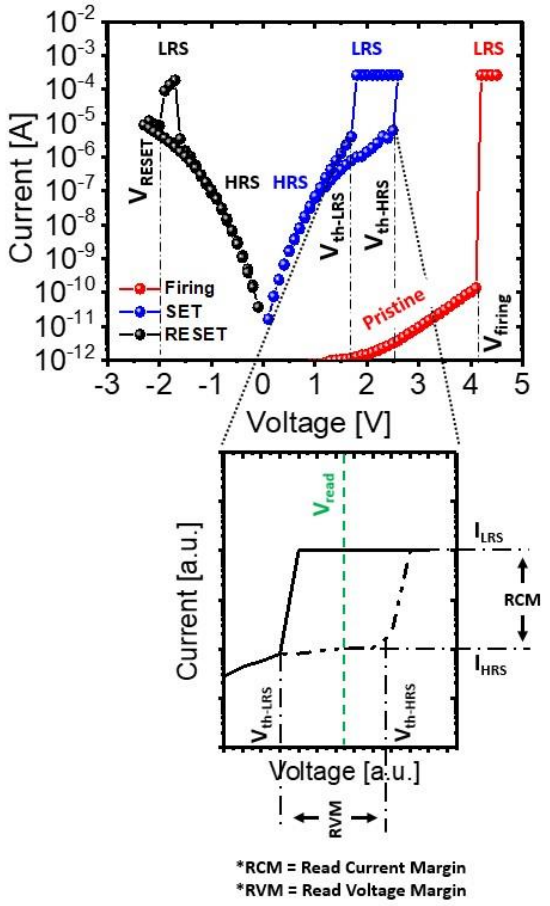

Fig.3. 1S1R current-voltage characteristics. A Read Voltage Margin can be defined within $V_{th\text{-}HRS}$ and $V_{th\text{-}LRS}$, which enables to perform reading operation by applying V_{read} potential difference at the device terminals.

Fig.3 shows 1S1R classical current-voltage characteristics together with the main electrical parameters used in this work. An initial firing process is required for 1S1R initialization from pristine to Low-Resistive State (LRS). Once the firing process is done, the OxRAM resistive state impacts the overall 1S1R switching voltages. On one hand, if the OxRAM is at High-Resistive State (HRS), $V_{th\text{-}HRS}$ is required for 1S1R switching. On the other hand, if the OxRAM is at Low-Resistive State (LRS), $V_{th\text{-}LRS} < V_{th\text{-}HRS}$ is required for 1S1R switching. $V_{th\text{-}HRS}$-$V_{th\text{-}LRS}$ defines the 1S1R Read Voltage Margin (RVM). The higher is the OxRAM HRS resistance, the higher will be $V_{th\text{-}HRS}$ due to additional voltage drop at the device and so the larger will be the 1S1R RVM [29]. Therefore, by applying a certain V_{read} comprised between V_{thLRS} and $V_{th\text{-}HRS}$ at the device terminals, 1S1R reading operation can be performed. If I_{LRS} (resp. I_{HRS}) is read, OxRAM is at LRS (resp. HRS). Additionally, taking advantage of the OTS bipolar characteristics, a negative bias V_{reset} is required for OxRAM RESET process.

The Read Voltage Margin variability being the most critical parameter for reliable 1S1R operation [26], both the 1S1R stack and the applied programming conditions must be engineered to guarantee satisfactory 1S1R Read Voltage Margin and reduce its variability while avoiding device early degradation [27]. **Fig.4** plots schematically the voltage repartition between OxRAM and OTS for both SET and RESET operations. V_{hold} drops at the OTS device while it is at the conductive state. Accordingly, being V_{app} the applied voltage to the overall 1S1R stack, V_{app}-V_{hold} drops at the OxRAM. Therefore, engineering both the 1S1R stack features and the applied programming voltages becomes key to efficiently open the OTS device ($V_{app}>V_{th\text{-}HRS}$ and $V_{app}>V_{thLRS}$) while remaining compatible with a functional OxRAM operation: enough voltage drop for a convenient programming operation but not excessive to prevent device early degradation.

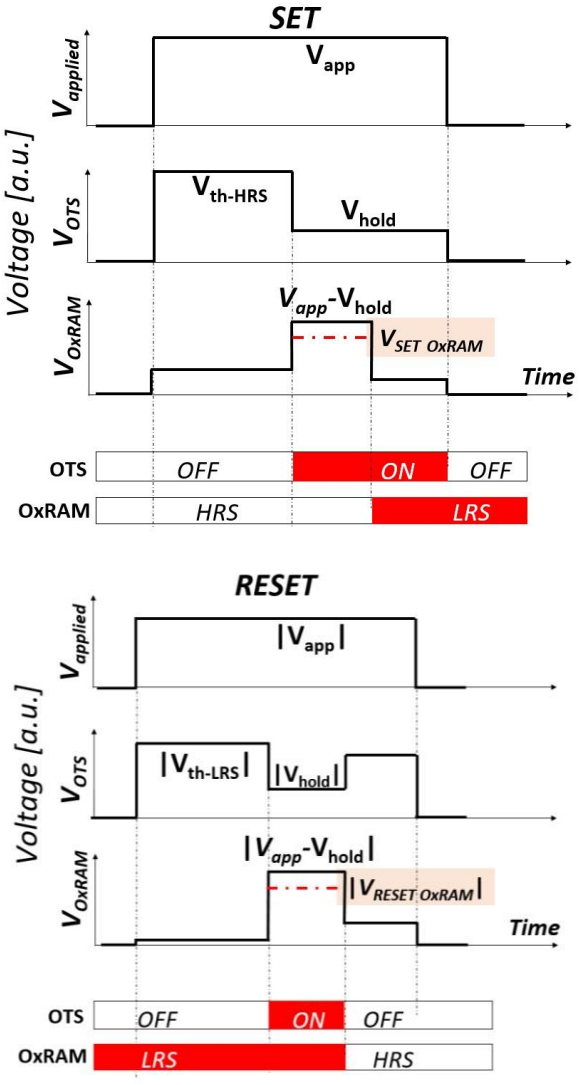

Fig.4. Voltage repartition in 1S1R structures for OTS and OxRAM devices during SET and RESET programming operation.

All in all, guidelines for 1S1R stack and programming conditions optimization are presented in **Fig.5**. 1S1R functional SET and RESET programming voltages are identified as a function of $V_{th\text{-}LRS}$ switching voltages and HfO_2 thicknesses, based on design-of-experiments endurance characteristics [27].

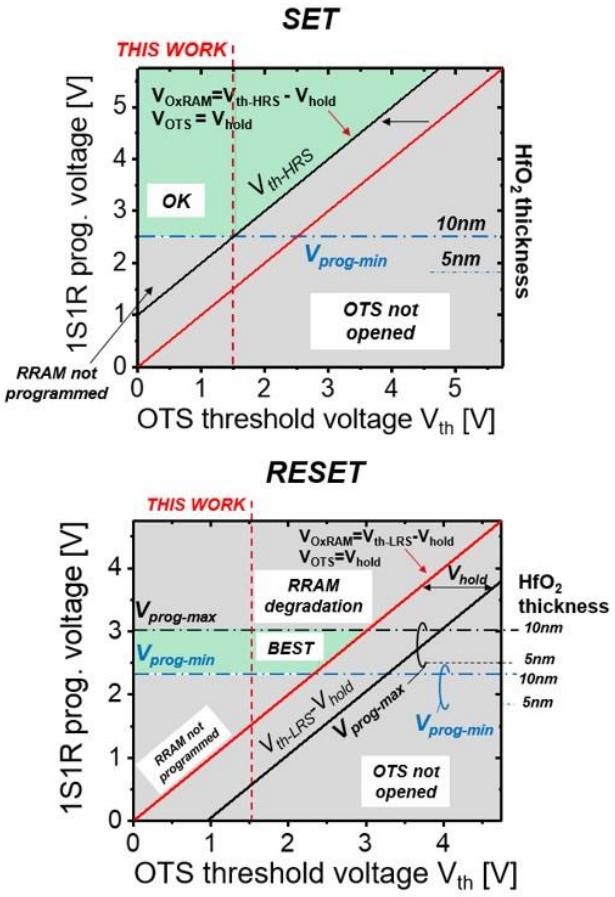

Fig.5. Guidelines for 1S1R programming operation, ensuring satisfactory RVM and preventing device early degradation. Functional 1S1R SET and RESET voltages operating regions are identified as function of 1S1R $V_{th\text{-}LRS}$ switching voltages.

Fig.6 a) shows RVM experimental distributions for adapted applied programming conditions for the stack of interest. Distinct $V_{th\text{-}LRS}$ and $V_{th\text{-}HRS}$ distributions are obtained up to 2.5σ. 1S1R reading voltage V_{read} is presented. 1S1R very low ~10^{-4} Bit Error Rate (BER) is extrapolated. The 1S1R read operation is performed and the experimental LRS and HRS current distributions are presented in **Fig.6 b)**, demonstrating statistical bit error free behavior for 1S1R devices up to 2.5σ.

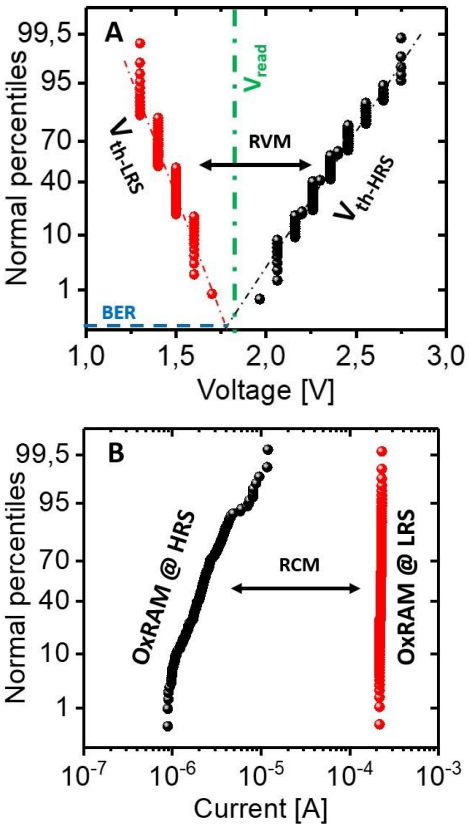

Fig.6. A) Experimental Read Voltage Margin distributions. Distinct $V_{th\text{-}LRS}$ and $V_{th\text{-}HRS}$ distributions are obtained up to 2.5σ for adapted programming conditions for the device of interest. B) Experimental Read Current Margin distributions. Distinct LRS and HRS currents are demonstrated.

3.2 1S1R programming and reading reliability

To evaluate the 1S1R pertinence for synaptic weight storage for BNN neuromorphic hardware application, the neural network of interest (**Fig.1**) is trained offline to perform MNIST image recognition task over 30 epochs using Binary Optimizer (Bop) approach [30]. During training process, weight updates are quantified. Depending on the switch of the binary weight (i.e. going from 1 to -1 or the opposite), the equivalent energy consumption associated to 1S1R device SET or RESET operation can be quantified. 97,5% of accuracy is obtained after the training process. Additionally, the amount of binary weight updates for the overall training process is presented in **Fig.7 a)**. We chose to represent only the weights between the hidden and output layers, as they are updated more often. 1547 maximum updates per weight are observed. Moreover, **Fig.7 b)** shows 1S1R experimental programming endurance characteristics for optimized applied programming conditions, where no bit fail is observed. Therefore, the 1S1R error-free programming endurance characteristics are demonstrated to be 10x higher than the maximal required weight updates for a single synaptic weight during the overall BNN training.

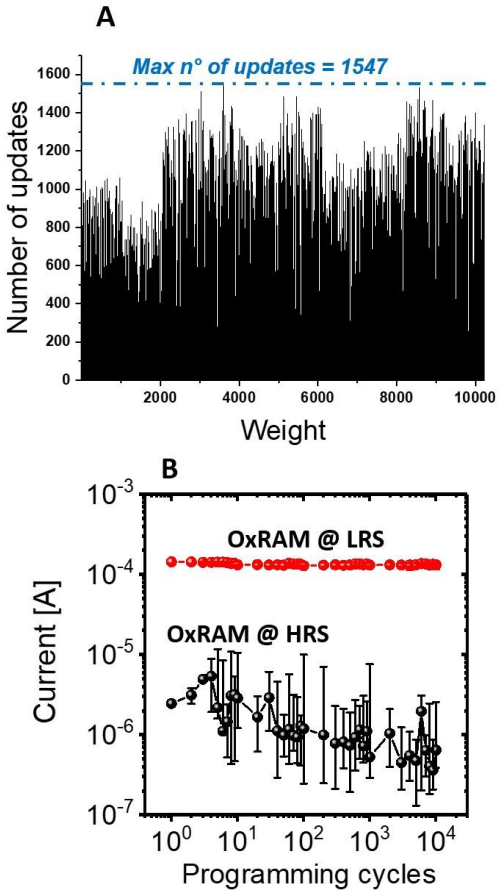

Fig.7. A) Weight updates for the output layer of the BNN architecture of interest. The maximum amount of weight updates is demonstrated to be equal to 1547, achieving 97,5% accuracy on MNIST image recognition task. B) 1S1R experimental programming endurance characteristics. No fail is observed for 10^4 SET+RESET programming cycles.

Once trained, BNN inference lifetime begins and is going to be limited by 1S1R device resilience to reading operations. 1S1R experimental read disturb characteristics are presented in **Fig.8**, demonstrating 10^9 reading cycles without any device degradation.

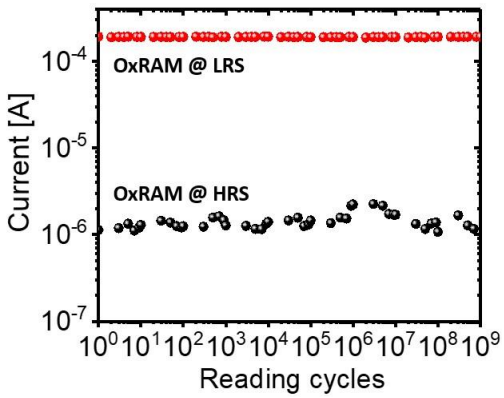

Fig.8. 1S1R experimental read disturb characteristics for the device of interest. 10^9 reading operations without any device degradation are demonstrated. Tests were stopped after 10^9 reading operations.

3.3 BNN figures of merit

At the device level, 1S1R relevance for synaptic weight storage for BNN neuromorphic hardware application is demonstrated on the previous section. At the system level, the crossbar footprint and the amount of synaptic weight updates (meaning SET and RESET operations on 1S1R devices, which can lead to 1S1R degradation and excessive energy consumption for weight updating), must be optimized while preserving satisfactory network accuracy. In particular, these parameters depend on the BNN architecture, as the amount of synaptic weights exponentially grows with the number of neurons. The bigger the amount of neurons, the bigger the overall 1S1R memory footprint. The number of neurons in the input (resp. output) layer being determined by the number of pixels on the image (resp. the number of classes), only the number of neurons in the hidden layer can be tuned. Therefore, we consider a variable number of neurons for the hidden layer of the BNN architecture of interest, belonging to X ϵ [64; 128; 256; 512; 1024; 4096].

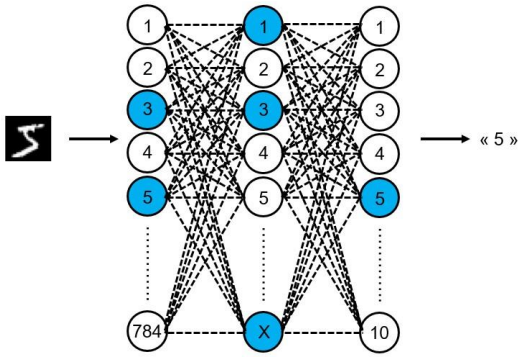

Fig.9. Binarized Neural Networks considered in this section. Fully connected neural network with one hidden layer of X neurons for MNIST handwritten recognition. X values belong to [64; 128; 256; 512; 1024; 4096].

In this context, the neural networks of interest (**Fig.9**) were trained offline to perform MNIST handwritten digit recognition task over 30 epochs using Binary Optimizer (Bop) approach [30]. Again, weight updates are quantified during training process. To quantify the energy consumption associated to binary weight update (going from 1 to -1 or the opposite), 1S1R SET and RESET programming and 1S1R reading conditions are summarized in **Table 1**. Because the programing time for SET and RESET 1S1R operations is reduced, we must apply higher programming voltages as the applied pulse sweep rate increases [27]. The electrical consumption required to program a single device one single time for both operations is next calculated. Less than 70pJ electrical consumption per operation can be achieved for both programming and reading operations.

TABLE I
1S1R programming and reading conditions

SET programming current	200µA
SET voltage	3.2V
SET programming time	100ns
SET electrical consumption	64pJ
RESET programming current	200µA
RESET voltage	2.9V
RESET programming time	100ns
RESET electrical consumption	58pJ
Reading current	200µA
Reading voltage	2.6V
Reading time	100ns
Read electrical consumption	52pJ

Fig.10 presents the BNN accuracy evolution with the amount of training epochs. First, the amount of neurons on the hidden layer has a strong impact on the network performance. The bigger the amount of neurons, the bigger the amount of learnable parameters (binary synaptic weights) and so the higher the accuracy, up to some extent. Indeed, by increasing too much the amount of neurons, the network may be subject to the overfitting phenomenon. Second, the amount of neurons on the hidden layer has a strong impact on the amount of iterations required for learning. While a very low amount of training epochs is required for smaller networks to reach their maximal accuracy, it strongly increases for bigger networks.

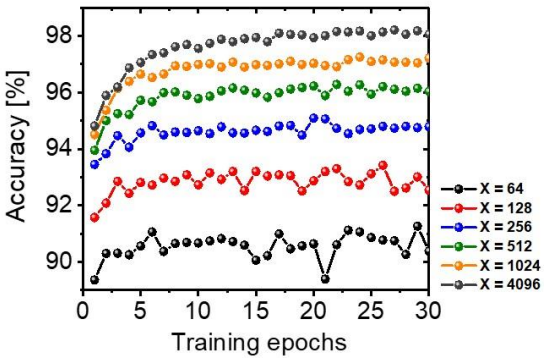

Fig.10. BNN accuracy as a fonction of the training epochs. Fully connected neural network with one hidden layer of X neurons for MNIST handwritten recognition is considered. X values belong to [64; 128; 256; 512; 1024; 4096].

Given the ability to accommodate the periphery under the crossbar memory array, the overall crossbar footprint is defined by the maximal between the footprint of the crossbar memory array and the footprint of the periphery. Assuming one driver transistor height per bitline and per wordline, the associated periphery footprint evolution with the amount of neurons on the hidden layer is calculated for 28nm highvoltage CMOS. The crossbar memory footprint is also calculated, assuming a CD_{min} metal width and space between metal lines. Both memory and periphery area or footprint (FP) evolution with the amount of neurons on the hidden layer is presented in **Fig.11**.

Targeting $FP_{periphery}<FP_{memory}$, a minimal amount of neurons on the hidden layer of 71 is demonstrated. All in all, the overall 1S1R crossbar footprint is compared with the equivalent 1T1R matrix footprint, based on experimental data from [31]. The overall footprint is reduced of about one order of magnitude by replacing the classical 1T1R architecture by a denser 1S1R crossbar architecture.

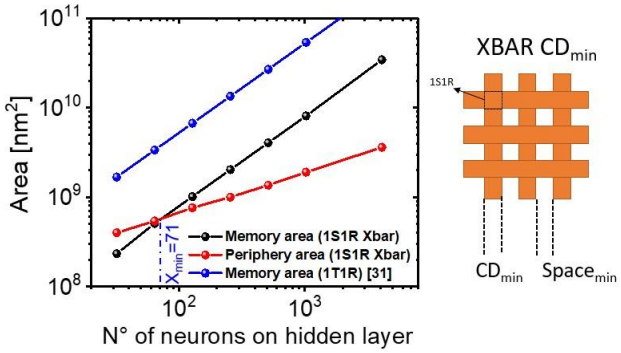

Fig.11. Crossbar footprint evolution with the amount of neurons in the hidden layer of the network. The presented area is calculated by addding the respective areas from both Input-Hidden crossbar matrix and the Hidden-Output crossbar matrix. By replicing the classical 1T1R architecture by a denser 1S1R crossbar architecture, the overall memory footprint is reduced about one order of magnitude.

Fig.12 presents the overall crossbar footprint and the amount of synaptic binary weight updates during the training as a function of the maximal attainable BNN accuracy. Accuracy degradation due to memory bit errors is not considered. We observe that the crossbar footprint and the amount of weight updates increase with accuracy. In particular, the crossbar footprint grows faster than the weight updates. Therefore, accuracy can be improved (up to some extent) by increasing the size of the network with reasonable weight updating at the expense of higher crossbar footprint.

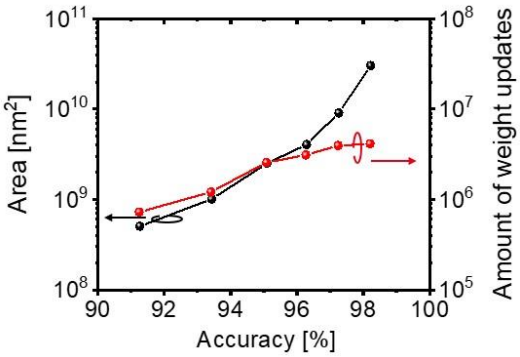

Fig.12. Crossbar overall footprint and amount of synaptic binary weight updates during the training as a function of the maximal attainable BNN accuracy.

After the training process, the BNNs of interest are tested on the dataset validation sets. At this point, fails are artificially introduced into the BNN synaptic weights in order to simulate the impact of 1S1R bit errors on the network performance. In particular, some +1 weights are replaced by -1 weights, and vice versa. **Fig. 13 A)** presents the obtained validation accuracy evolution with the 1S1R memory parasitic bit error rates. First, the 1S1R experimental BER characteristics in this work are demonstrated not to induce any BNN performance degradation. Then, the BNN bit error tolerance is demonstrated to be strongly influenced by the amount of neurons on the hidden layer. The higher the network width, the higher the network tolerance to unexpected bit errors.

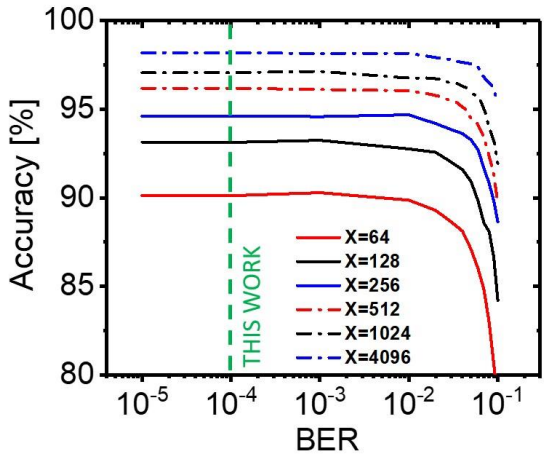

Fig.13. A) Maximal attainable BNN accuracy evolution with memory BER for the architectures of interest. No accuracy degradation is observed for our 1S1R devices.

4. Conclusion

OxRAM+OTS 1S1R pertinence for synaptic weight storage for BNN neuromorphic hardware application is explored by crossing offline training simulations with experimental data, focusing on MNIST handwritten digit recognition task. At the device level, general guidelines allowing to design the 1S1R stack and programming conditions for optimized Read Window Margin and extended endurance are illustrated. A very promising 1S1R bit error rate of ~10^{-4} is demonstrated experimentally, while preserving satisfactory endurance capabilities. Furthermore, at the system level, we provided general guidelines allowing to design the BNN architecture for optimized footprint, reduced synaptic weight updating operations, extended network accuracy and improved tolerance to unexpected bit errors. It appears that 1S1R parasitic fails are perfectly tolerated by the BNN, which performance is not degraded at all. All in all, this is the first step for the integration of an 1S1R-based binarized neural network system able to learn. This opens the path through inference computing on chip using 1S1R crossbar matrices, provided the limitation of the electrical consumption for programming and reading operations.

Acknowledgements

This work was partially funded by European commission, MIAI @ Grenoble Alpes (ANR-19-P3IA-0003), French State and Auvergne-Rhône Alpes region through ECSEL project ANDANTE and French Nano2022 program.

References

[1] A. Pedram et al., “Dark memory and accelerator-rich system optimization in the dark silicon era”, IEEE Design Test 34, 39-50 (2017)

[2] V. Sze et al., “Efficient processing of deep neural networks: from algorithms to hardware architecture”, NEURIPS2019

[3] D. Strukov et al., “Building brain-inspired computing”, Nat. Commun. 10, 4838 (2019)

[4] D. Garbin et al., "HfO2-Based OxRAM Devices as Synapses for Convolutional Neural Networks," in IEEE Transactions on Electron Devices, vol. 62, no. 8, pp. 2494-2501, Aug. 2015, doi: 10.1109/TED.2015.2440102.

[5] H. Tsai et al., “Recent progress in analog memory-based accelerators for deep learning”, J. Phys. D: Appl. Phys. 51 283001 (27pp)

[6] D. Ielmini et al., “Emerging Neuromorphic Devices”, Nanotech. 21;31(9):092001 (2020)

[7] S. Ambrogio et al., “Equivalent-Accuracy Accelerated NeuralNetwork Training Using Analogue Memory”, Nat. 558, 60-67 (2018)

[8] F. Cai et al., “Power-efficient combinatorial optimization using intrinsic noise in memristor Hopfield neural networks”, Nat. Electron. 3, 409-418 (2020)

[9] S. Yu et al, “Scaling-Up Resistive Synaptic Arrays For NeuroInspired Architecture: Challenges And Prospect”, IEDM2015 Tech. Dig., pp. 17.3.1-17.3.4

[10] M. Prezioso et al., “Training And Operation Of An Integrated Neuromorphic Network Based On Metal-Oxide Memristors”, Nat. 521, 61-64 (2015)

[11] P. Jain et al., “13.2 A 3.6Mb 10.1 Mb/mm2 Embedded Non-Volatile ReRAM Macro in 22nm FinFET Technology with Adaptive Forming/Set/Reset Schemes Yielding Down to 0.5V with Sensing Time of 5ns at 0.7V”, 2019 ISSCC, pp 212-214

[12] C. Chou et al., “A 22nm 96KX144 RRAM Macro with Self-Tracking Reference and a Low Ripple Charge Pump to Achieve a Configurable Read Window and a Wide Operating Voltage Range” 2020 IEEE Symposium on VLSI Circuits, pp.1-2

[13] C. Nail et al., “Understanding RRAM endurance, retention and window margin trade-off using experimental results and simulations”, IEDM2016 Tech. Dig., pp. 4.5.1-4.5.4

[14] C. Nail et al., “Hybrid-RRAM towards Next Generation of Non-volatile Memory: Coupling of Oxigen Vacancies and Metal Ions”, Advanced Electronic Materials, 1800658, 2018.

[15] T. Hirtzlin et al., “Digital Biologically Plausible Implementation of Binarized Neural Networks With Differential Halfnium Oxide Resistive Memory Arrays”, Front. Neurosci. 13:1383.

[16] M. Bocquet et al., “In-Memory and Error-Immune Differential RRAM Implementation of Binarized Deep Neural Networks”, IEDM2018 Tech. Dig., 20.6.1.

[17] I. Hubara et al. “Binarized neural networks”, *in Proc. of the 30th International Conference on Neural Information Processing Systems.* 2016

[18] M. Rastegari et al., “Xnor-net: Imagenet classification using binary convolutional neural networks”, European Conference on Computer Vision, Springer, Cham, pp. 525-542 (2016)

[19] V. Joshi et al., “Accurate deep neural network inference using computational phase-change memory”, Nat. Commun. 11, 2473 (2020)

[20] A. Valentian et al., “Fully Integrated Spiking Neural Network with Analog Neurons and RRAM Synapses”, IEDM2019, pp. 14.3.1-14.3.4

[21] A. Regev et al., “Fully-Integrated Spiking Neural Network Using SiOx-Based RRAM as Synaptic Device”, 2020 2nd AICAS, pp. 145-148

[22] X. Sheng et al., “Low-Conductance and Multilevel CMOS-Integrated Nanoscale Oxide Memristors”, Adv. Electron. Mater. 2019, 5, 180076

[23] G. Molas *et al.*, “Crosspoint Memory Arrays: Principle, Strengths and Challenges”, *in Proc. of IMW2020, pp.1-4*

[24] J. Minguet Lopez et al., “Optimization of RRAM and OTS selector for advanced Low voltage CMOS Compatibility”, *in Proc. of IMW 2020, pp.1-4*

[25] S. N. Truong, “Single Crossbar Array of Memristors With Bipolar Inputs for Neuromorphic Image Recognition”, in IEEE Access, vol.8, pp. 69327-69332, 2020

[26] D. Alfaro Robayo et al., “Reliability and Variability of 1S1R OxRAM-OTS for High Density Crossbar Integration”, IEDM2019 Tech. Dig., pp.35.3.1-35.3.4

[27] J. Minguet Lopez et al., “Elucidating 1S1R operation to reduce the read voltage margin variability by stack and programming conditions optimization”, IRPS2021, pp.1-6

[28] Le Cun et al., “Gradient-based learning applied to document recognition”, *in Proc. IEEE 86*, 2278-2324

[29] Verdy et al. “Optimized Reading Window for Crossbar Arrays Thanks to Ge-Se-Sb-N-based OTS Selectors”, in IEDM2018 Tech. Dig., pp. 37.4.1-37.4.4

[30] K. Helwegen et al. “Latent Weights Do Not Exist : Rethinking Binarized Neural Network Optimization”, Neurips2019

[31] L. Grenouillet et al., “16kb 1T1R OxRAM arrays embedded in 28nm FDSOI technology demonstrating low BER, high endurance, and compatibility with core logic transistors”, *in Proc. of IMW 2021*