1S1R sub-threshold operation in Crossbar arrays for low power BNN inference computing

J. Minguet Lopez, F. Rummens, L. Reganaz, A. Heraud, T. Hirtzlin, L. Grenouillet, G. Navarro, M. Bernard, C. Carabasse, N. Castellani, et al.

1S1R sub-threshold operation in Crossbar arrays for low power BNN inference computing

J. Minguet Lopez, F. Rummens, L. Reganaz, A. Heraud, T. Hirtzlin, L. Grenouillet, G. Navarro, M. Bernard, C. Carabasse, N. Castellani, V. Meli, S. Martin, T. Magis, E. Vianello, C. Sabbione, D. Deleruyelle[3], M. Bocquet[2], J. M. Portal[2], G. Molas*, F. Andrieu
[1]CEA, LETI, MINATEC Campus, GRENOBLE, France, joel.minguetlopez@cea.fr, tifenn.hirtzlin@cea.fr, [2]Aix Marseille Univ, Université de Toulon, CNRS, IM2NP, Marseille, France, [3]INL CNRS, INSA Lyon, France, *Now with Weebit Nano Ltd.

***Abstract*—We experimentally validated the sub-threshold reading strategy in OxRAM+OTS crossbar arrays for low precision inference in Binarized Neural Networks. In order to optimize the 1S1R sub-threshold current margin, an experimental and theoretical statistical study on HfO_2-based 1S1R stacks with various OTS technologies has been performed. Impact of device features (OxRAM R_{HRS}, OTS non-linearity and OTS threshold current) on 1S1R sub-threshold reading is elucidated. Accuracy and power consumption of a Binarized Neural Network designed in 28nm CMOS have been estimated with Monte Carlo simulations. A gain of 3 orders of magnitude in power consumption is demonstrated in comparison with conventional threshold reading strategy, while preserving the same network accuracy.**

Keywords— chalcogenide, crossbar, OTS, RRAM, BNN

I. Introduction

Deep learning hardware accelerators based on weight stationary approaches have shown promising speed and power consumption improvements [1]. This trend is particularly exacerbated using 1T1R RRAM cell coupled with near memory computing architecture for low precision Binarized Neural Networks (BNN) [2,3]. Nevertheless, the overall memory density improvement remains essential to large Neural Network (NN) with high accuracy. In this context, we propose to replace the 1T1R architecture by a denser 1S1R Crossbar system, where an Ovonic Threshold Switching (OTS) back-end selector [4,5] is used for memory selection. However, opening the selector (OFF-to-ON transition) during 1S1R reading operation implies several challenges at the system level, due to the high current flowing through the stack [6]. First, the overall electrical consumption becomes limiting. Second, following the maximal current density rules in metal lines for the node of interest, the maximum readable array size per timestep is limited to preserve satisfactory overall density. Within this framework, we present a theoretical and experimental study to optimize the 1S1R read current margin when OTS operates in sub-threshold regime (before switching), by optimizing the OTS technology (composition and thickness) and the 1S1R programming conditions. The main goal here is to achieve the largest sub-threshold current margin while limiting the I_{LRS} current. Based on experimental programming and reading endurance characteristics, we evaluate the benefit of 1S1R sub-threshold reading for BNN inference operation simulating a novel 28nm node BNN circuit. Furthermore, we perform off-chip training simulations on a fully connected BNN with one hidden layer of 1024 neurons for image classification tasks. All in all, we demonstrate overall satisfactory accuracy while minimizing the electrical consumption and maximizing the system density for inference computing on-chip.

II. Technological Details

OxRAM (1R) is co-integrated with an OTS (1S) back-end selector (1S1R configuration). 5nm-thick ALD HfO_2-based OxRAM is used as a memory device. Three various OTS technologies are compared, belonging to two different selector families: As-based [7] and Sb-based OTS [5]. Illustrative TEM cross section is provided in **Fig.1**, demonstrating a consistent co-integration of all layers.

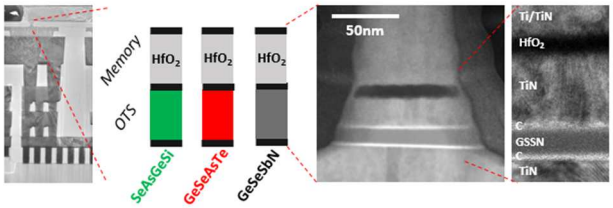

Fig. 1. Illustrative SEM and TEM cross sections for the stacks of interest, integrated in the BEOL of a 130nm CMOS.

III. Results and Discussion

A. Key parameters for large 1S1R sub-threshold margin

Fig.2 presents typical 1S1R current-voltage characteristics after the initial firing operation, which is required for OTS and OxRAM initialization. By applying a certain V_{read} lower than the OTS switching voltages (V_{th}), the 1S1R sub-threshold reading operation can be performed. In this context, the 1S1R read current $I_{read} < I_{th}$ at V_{read} is very dependent on the OxRAM resistive state. If the OxRAM is at Low Resistive State (LRS), I_{LRS} is read. If the OxRAM is at High Resistive State (HRS), $I_{HRS} < I_{LRS}$ is read. Thus, the I_{LRS}/I_{HRS} ratio represents the 1S1R sub-threshold current margin.

Accordingly, three main parameters play a key role on 1S1R sub-threshold current margin reliability: OxRAM R_{HRS}, OTS non-linearity (NL) and OTS threshold current (I_{th}). The influence of each individual parameter on 1S1R sub-threshold current margin is evaluated in **Fig.3**.

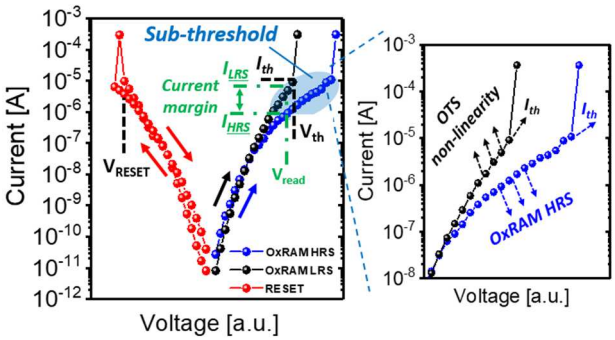

Fig. 2. 1S1R typical current-voltage characteristics. The key parameters for reliable 1S1R sub-threshold reading operation (OxRAM R_{HRS}, OTS non-linearity and OTS threshold current) are presented. Given OxRAM R_{LRS} << OTS R_{OFF}, OxRAM R_{LRS} impact on I_{LRS} is considered negligible.

Firstly, the higher R_{HRS}, the more the 1S1R I_{HRS} current at V_{read} decreases. Accordingly, the 1S1R sub-threshold current margin increases with the OxRAM R_{HRS} (**Fig.3A**). Indeed, R_{HRS} is very dependent on the applied 1S1R V_{RESET} (V_{RESET}<0), which is required for OxRAM RESET operation (LRS-to-HRS transition). In this context, applying $|V_{RESET}| > |V_{th}|$ ensures OTS satisfactory opening during the process. Secondly, the higher the OTS non-linearity, the more the 1S1R I_{LRS} current at V_{read} increases. Therefore, the 1S1R sub-threshold current margin increases with the OTS non-linearity (**Fig.3B**). Thirdly, the higher I_{th}, the higher V_{th} and so the more V_{read} can be increased. The 1S1R LRS currents increasing faster than the 1S1R HRS currents with respect to the applied voltages, increasing V_{read} thus implies 1S1R sub-threshold current margin increase (**Fig.3C**).

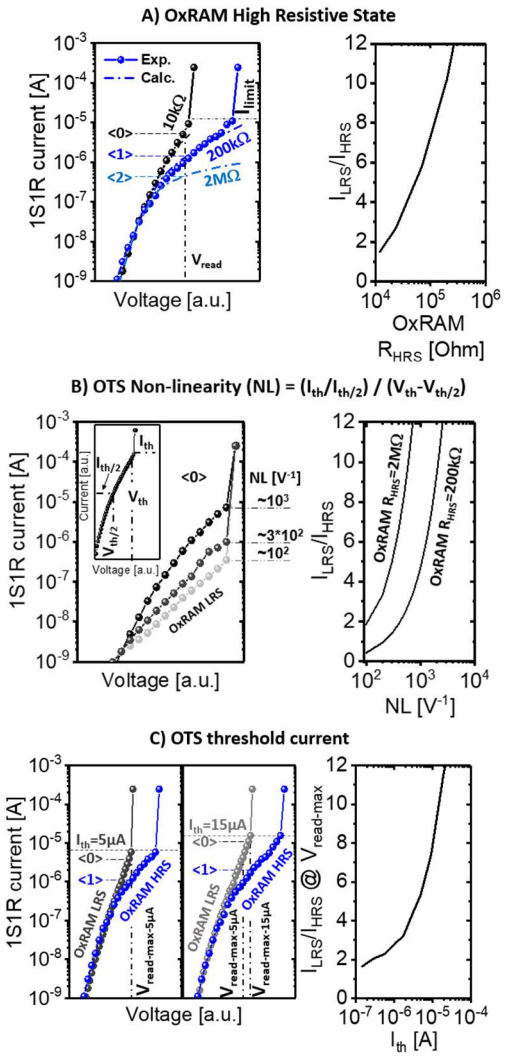

Fig. 3. A) 1S1R experimental IV characteristics for various programmed OxRAM resistive states. The OTS non-linearity (resp. OTS threshold currents) are fixed to 10^3 V^{-1} (resp. 10^{-5} A). The more the OxRAM is resistive, the more 1S1R I_{HRS} current at V_{read} decreases and so the larger is the current margin. B) 1S1R non-linearity (NL) definition. Three experimental OTS IV curves are presented. The 1S1R sub-threshold margin evolution with the OTS NL is presented, calculated for two OxRAM R_{HRS}. The higher the OTS non-linearity, the more 1S1R I_{LRS} current at V_{read} increases and so the larger is the current margin. C) 1S1R experimental IV characteristics for various OTS threshold current values (I_{th}). The OTS non-linearity (resp. OxRAM R_{HRS}) are fixed to 10^3 V^{-1} (resp. 200kΩ). The higher I_{th}, the more V_{read} can be increased. The higher V_{read}, the higher the 1S1R sub-threshold current margin.

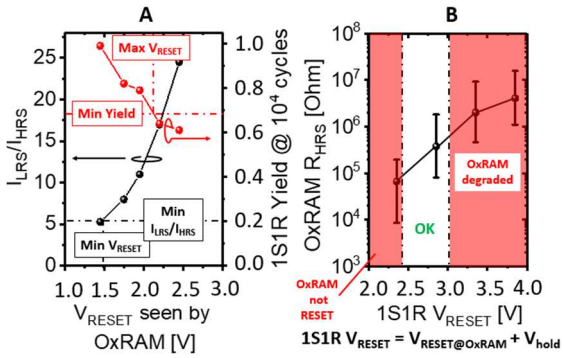

Fig. 4. A) Calculated 1S1R sub-threshold current margin and reading yield at 10^4 operating cycles evolution with the voltage seen by OxRAM during RESET operation. Reading yield is defined as the probability for I_{LRS} and I_{HRS} to be distinct. 1S1R I_{LRS}/I_{HRS} is calculated for fixed OTS characteristics (NL~10^3 V^{-1} and I_{th}~10^{-5} A). Minimum and maximum RESET voltages are identified, ensuring satisfactory 1S1R sub-threshold margin while limiting OxRAM early degradation. B) OxRAM R_{HRS} evolution with 1S1R applied RESET voltages. V_{hold} ~ 1V is assumed [5,8]. Overall functional 1S1R RESET voltage conditions are provided.

B. Optimization of 1S1R sub-threshold margin

1) Impact of 1S1R programming on OxRAM R_{HRS}

The influence of the applied RESET voltage seen by the OxRAM on the 1S1R sub-threshold current margin and reading yield at 10^4 operating cycles is presented in **Fig.4A**. One should note that we focus on low precision neuromorphic application, with relaxed constraints on read current margin. On one hand, a minimum voltage seen by the OxRAM of ~1.5V ensures a minimal 1S1R sub-threshold current margin of ~6 (**Fig.4A**), with maximal 1S1R yield. On the other hand, a maximal ~2.1V sustainable voltage seen by the OxRAM limits device early degradation (**Fig.4A**) but offers the best window margin. The RESET voltage dropping on the OxRAM being a function of 1S1R V_{RESET}, the 1S1R V_{RESET} influence on the programmed OxRAM R_{HRS} is experimentally illustrated in **Fig.4B**. Therefore, a maximal sustainable OxRAM R_{HRS}~$5*10^5\Omega$ is identified for the stack of interest.

2) Impact of OTS stack on OTS non-linearity and threshold current

OTS NL evolution with OTS thickness is presented in **Fig.5A**. The maximal attainable NL is demonstrated to be related with OTS composition. Moreover, while OTS NL remains constant with stack thicknesses for As-based OTS selectors, Sb-based OTS devices show increasing NL for thicker stacks.

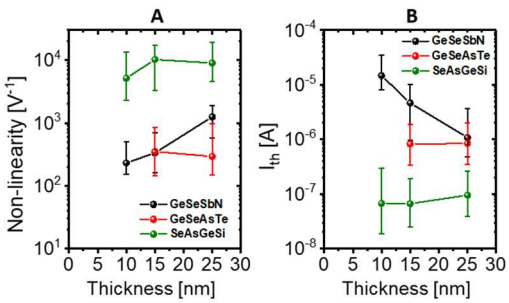

Fig. 5. A) OTS non-linearity and threshold currents (I_{th}) evolution with OTS thicknesses. Three OTS compositions are compared, belonging to As-based and Sb-based selector families. On one hand, no dependence between OTS parameters and thickness is oberved for As-based selectors. On the other hand, Sb-based selectors show increasing NL (resp. decreasing I_{th}) for thicker stacks. Moreover, the maximal (resp. minimal) attainable OTS non-linearity (resp. I_{th}) is linked to the OTS composition. Altogether, a tradeoff between OTS non-linearity and I_{th} is demonstrated.

Furthermore, OTS I_{th} evolution with stack thickness is presented in **Fig.5B**. While OTS I_{th} remains constant with stack thickness for As-based OTS selectors, Sb-based OTS devices show decreasing I_{th} for thicker stacks. Therefore, a tradeoff exists between OTS NL and OTS I_{th}.

3) 1S1R figures of merit on 1S1R sub-threshold margin

Based on these analyses, figures of merit on 1S1R sub-threshold margin are proposed. **Fig.6** shows the OTS I_{th} evolution with the OTS threshold switching voltages (V_{th}). Each circle color corresponds to a given OTS composition. For each OTS composition, the symbol shape refers to a given stack thickness. In this context, the functional voltage region is illustrated, preventing OxRAM degradation during 1S1R RESET operation. Four various OTS stacks appear here to allow satisfactory OxRAM programming operation.

Then, taking into account the OxRAM-compatible OTS stacks identified in **Fig.6**, the tradeoff between both OTS NL and I_{th} is illustrated in **Fig.7**. The symbol size represents the 1S1R sub-threshold current margin value. Optimal read window margin is extracted for 10nm GSSN OTS. This 1S1R stack is thus selected for next section of this paper.

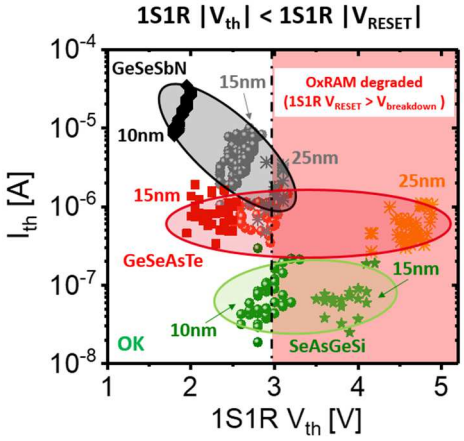

Fig. 6. Functional 1S1R stack characteristics (composition, thickness) as function of 1S1R switching voltages. Each point corresponds to a different 1S1R device. The various symbol shapes (square, sphere, diamond, star) correspond to various OTS thicknesses. Given the OTS bipolar characteristics, V_{RESET} must be higher than V_{th} for satisfactory OTS opening during RESET operation. Accordingly, the tolerated 1S1R switching voltages are limited in order to prevent OxRAM early degradation. All in all, four 1S1R stacks allow durable 1S1R programming operation.

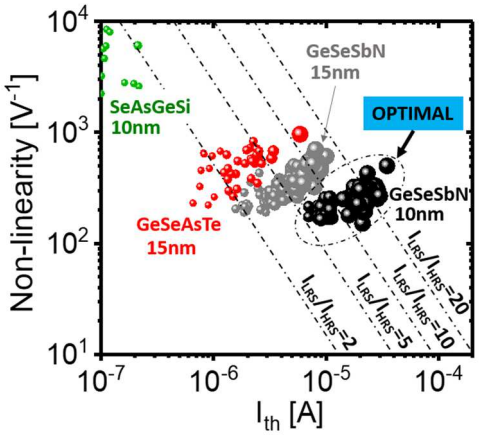

Fig. 7. Among the various OTS technologies promising durable 1S1R programming operation (**Fig.6**), the optimal 1S1R stack is extracted for 1S1R sub-threshold margin maximization. The tradeoff between OTS non-linearity and threshold currents is used for stack discretization.

4) 1S1R sub-threshold margin reliability

The optimized 1S1R sub-threshold experimental cycle-to-cycle programming current distributions are presented in **Fig.8A**. A BER of ~$7*10^{-3}$ with sub-threshold read scheme is demonstrated, opening the path to low precision neuromorphic applications. Moreover, **Fig.8B** presents experimental read disturb characteristics for V_{read}=2.5V. No margin degradation is observed up to 10^9 reading cycles. Therefore, a threshold current for OxRAM state identification (I_{REF}) is extracted.

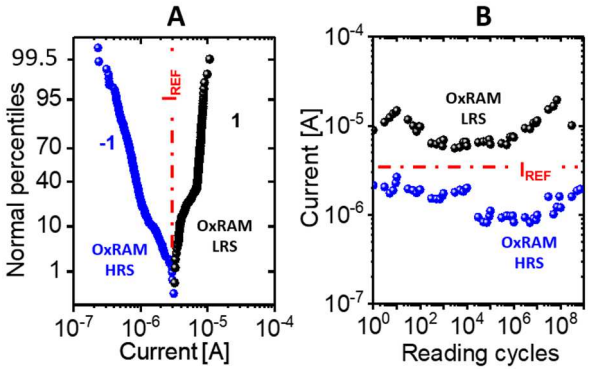

Fig. 8. A) 1S1R sub-threshold experimental cycle-to-cycle programming LRS and HRS read current distributions. Currents are extracted from five different devices. A BER ~ $7*10^{-3}$ is demonstrated. B) 1S1R sub-threshold reading endurance.

C. 1S1R sub-threshold operation for BNN inference computing

To evaluate the potential of 1S1R sub-threshold reading approach for synaptic weight storage for BNN inference hardware application, we performed off-chip training simulations on a fully connected NN with one hidden layer (**Fig.9**). We focused on the standard machine learning MNIST image recognition task.

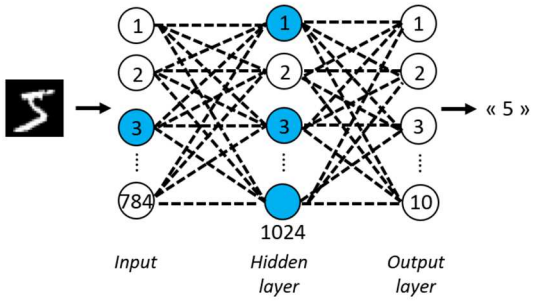

Fig. 9. Binarized Neural Network considered in this section. Fully connected neural network with one hidden layer of 1024 neurons for MNIST handwritten recognition.

In this context, the binarized weights are physically stored layer-per-layer in crossbar arrays, where the amount of Word-Lines (WL) (resp. Bit-Lines (BL)) corresponds to the amount of input neurons (resp. output neurons) of the layer. One input activation response is evaluated per timestep, by collecting the resulting currents on bottom of the crossbar BLs. At the end of every timestep, each BL current is separately compared with I_{REF}, which allows identifying the 1S1R resistive state (**Fig.10**). To this aim, we designed a 28nm node 1S1R sub-threshold based BNN circuit (**Fig.11**), where one sense amplifier is deployed per BL.

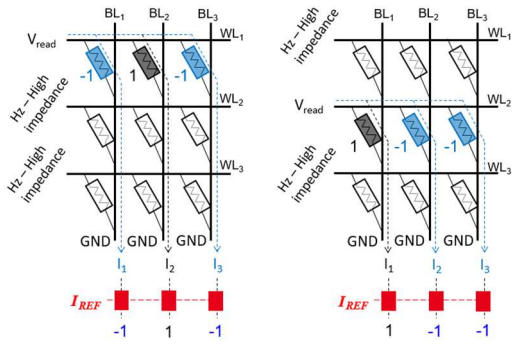

Fig. 10. Schematic biasing scheme on the Crossbar for BNN inference operation. One unique Word Line is activated per timestep. The resulting sub-threshold current on the bottom of the Bit Line is compared with the reference current (I_{REF}), enabling 1S1R device resistive state identification.

Fig.12 summarizes the overall BNN inference electrical consumption. Multiple Monte Carlo crossbar reading simulations are used for the calculation, where 50% of the devices are considered to be at LRS on the array. 40MHz circuit operation frequency is considered, where ~50ns are dedicated to sense the WL response. Then, the BNN circuit area is calculated, taking into account both crossbar and peripherals contribution (calculation details can be found in [6]). In this context, 1S1R reading in sub-threshold regime (this work) is benchmarked with 1T1R, 2T2R, and 1S1R operated in threshold regime, for the application of interest. All in all, 1S1R sub-threshold read operation appears very efficient in terms of energy consumption, achieving ultra-low 76 fJ per read bit (3 decades lower than the read energy for 1S1R in threshold regime). 32x32 crossbar array is considered for the calculation. Moreover, ~8x footprint improvement is demonstrated at crossbar level for back-end selector architectures, with respect to 1T1R standard ones. Finally, **Fig.13** presents the maximum attainable BNN accuracy evolution with memory BER, for the architecture of interest. The 1S1R sub-threshold experimental BER characteristics are perfectly tolerated by the network.

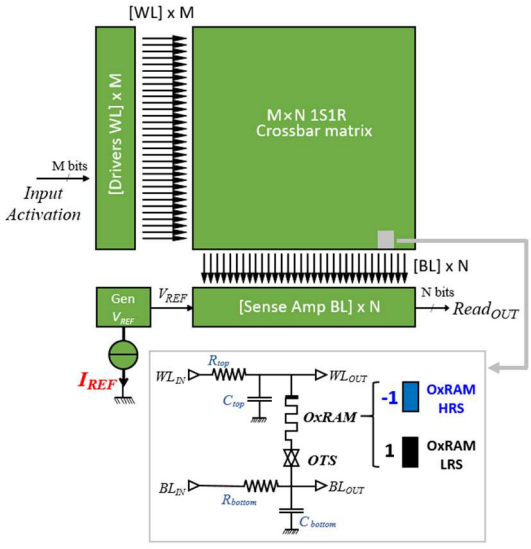

Fig. 11. Schematical description of the 1S1R sub-threshold based BNN circuit for inference computing, designed on 28nm node technology. The circuit is scalable to larger NN layer size (M and N values). One sense amplifier is deployed per BL. The previously extracted I_{REF} is used as a current comparator during sensing operation on bottom of each BL, allowing to identify both OxRAM HRS and LRS states at the end of every timestep.

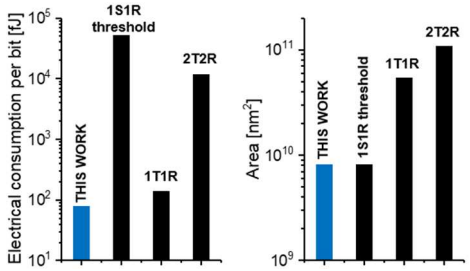

Fig. 12. BNN electrical consumption and area estimation for the application of interest. 1S1R in sub-threshold regime, 1S1R threshold regime, 1T1R and 2T2R configurations are presented. 50% of the devices on the crossbar are considered in LRS for the calculation. 1S1R in threshold regime data is extracted from [6]. 1T1R data is extracted from [9].

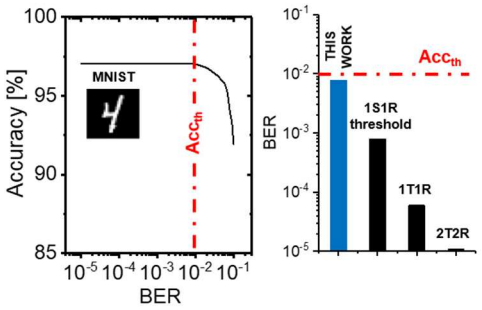

Fig. 13. BNN accuracy evolution with binarized weights BER. No accuracy degradation is observed for 1S1R devices exploited at sub-threshold regime.

IV. CONCLUSIONS

OxRAM+OTS crossbar with sub-threshold reading operation is studied for dense synaptic weight storage on BNN inference hardware application. Comparing three OTS technologies, general guidelines allowing to design 1S1R stack for optimized sub-threshold current margin and extended endurance are proposed. A promising ~7*10^{-3} BER on sub-threshold current margin is demonstrated, while preserving satisfactory read endurance capabilities. Ultra-low ~76 fJ/bit reading consumption (~1000 times lower than 1S1R arrays in threshold regime) is obtained with Monte Carlo simulations on 28nm node design. High density capability with ~8x area reduction is reached with respect to standard 1T1R configuration. Finally, 1S1R sub-threshold reliability is demonstrated to be compatible with one hidden layer fully connected Binarized NN, with no counterpart in terms of network accuracy. This opens the path to on-chip inference computing using 1S1R crossbar matrices operated in sub-threshold regime.

ACKNOWLEDGMENT

This work was partially funded by the European project ANDANTE and StorAIge, as well as the French IPCEI program.

REFERENCES

[1] V. Sze et al., *in Proc.* 2019 NeurIPS.
[2] M. Bocquet et al., *in Proc.* 2018 IEEE IEDM.
[3] G. Molas et al., *Appl. Sci.* 2021, *11*, 11254.
[4] S.R. Ovshinsky, Phys. Rev. Lett., 21 (1968), 1450-1453.
[5] J. Minguet Lopez et al., *in Proc.* 2021 IEEE IMW.
[6] J. Minguet Lopez et al., 2022 *Semicond. Sci. Technol.* 37 0140.
[7] A. Verdy at al, *in Proc.* 2019 IEEE IMW.
[8] J. Minguet Lopez et al., *in Proc.* 2021 IEEE IRPS.
[9] L. Grenouillet et al., *in Proc.* 2021 IEEE IMW