1S1R sub-threshold operation in Crossbar Arrays for Neural Networks hardware implementation

Joel Minguet Lopez, Manon Dampfhoffer, Tifenn Hirtzlin, Lucas Reganaz, Laurent Grenouillet, Gabriele Navarro, Mathieu Bernard, Thomas Magis, Catherine Carabasse, Niccolo Castellani, et al.

1S1R Sub-threshold Operation in Crossbar Arrays for Neural Networks Hardware Implementation

Joel Minguet Lopez[1], Manon Dampfhoffer[1], Tifenn Hirtzlin[1], Lucas Reganaz[1], Laurent Grenouillet[1], Gabriele Navarro[1], Mathieu Bernard[1], Thomas Magis[1], Catherine Carabasse[1], Niccolo Castellani[1], Valentina Meli[1], Elisa Vianello[1], Damien Deleruyelle[4], Jean-Michel Portal[3], Gabriel Molas[*], François Andrieu[1]

[1]CEA-Leti, Univ. Grenoble Alpes, GRENOBLE, France [*]Now with Weebit Nano Ltd., France

[2]CEA, CNRS, Grenoble INP, INAC-Spintec, Univ. Grenoble Alpes, GRENOBLE, France

[3]Aix Marseille Univ., CNRS, IM2NP, MARSEILLE, France

[4]INL CNRS, INSA Lyon, VILLEURBANNE, France

joel.minguetlopez@cea.fr

***Abstract*—This paper presents an outlook of Crossbar memory array capabilities while operated in the sub-threshold regime. By means of experimental data obtained on a RRAM resistive device co-integrated in series with an OTS back-end selector, the pertinence of 1S1R sub-threshold read operation for both standard Binarized Neural Networks (BNNs) and Binarized Spiking Neural Networks (B`SNNs) inference implementation in hardware is elucidated.**

***Keywords*—chalcogenide, crossbar, crosspoint, OTS, RRAM, PCM, 1S1R, BNN, BSNN, inference**

I. Introduction

The integration of single memristor Crossbar arrays into computing core units holds great promise for a successful deployment of deep learning accelerators. By leveraging this approach, significant reductions in both latency and power consumption can be achieved [1-2]. In particular, the hardware implementation of Neural Networks (NNs) synaptic weights using non-volatile resistive memory devices, such as Resistive Random Access Memory (RRAM) and Phase Change Memory (PCM), within 1T1R arrays is one of the most compelling [3].

However, the intrinsic operating variability of these emerging memory devices can severely limit the overall network performance. To address this challenge, both memory device and NNs optimization are primordial. Various smart strategies have been proposed at the device level, such as developing bit-error correcting codes and adaptive programming schemes [4-5]. At the NNs level, it has been shown that employing wide topologies with redundant behavior can effectively improve resilience to memory imperfections [6-7]. Nevertheless, this approach imposes significant constraints on memory storage capacity. Firstly, ensuring high memory capacity is crucial to avoid unnecessary memory array partitioning to perform the computation. Secondly, increasing memory density is essential to enable hardware implementations of wide NN architectures while maintaining a reasonable silicon footprint. Lastly, achieving fast and highly parallel computation on the synaptic memory arrays is crucial to maintain reasonable computation latency in such topologies.

In this context, to replace the classical 1T1R arrays by denser 1 Selector – 1 Resistor (1S1R) based Crossbar structures has been proposed [8]. Remarkably, the utilization of Ovonic Threshold Switch (OTS) back-end selectors on this kind of structures is already a reality with the 3DXpoint product. However, due to the high current required for OTS selector OFF-to-ON opening transition, the ability to operate 1S1R Crossbar structures in the sub-threshold regime is currently gaining in importance in the context of Neural Network synapses hardware implementation [9-10]. In our approach, two different kind of Neural Networks are considered: Artificial Neural Networks (ANNs) and Spiking Neural Networks (SNNs). On one hand, standard ANNs computations mostly rely on Multiply-and-Accumulate (MAC) operations between input activations and synaptic weights, which can directly be implemented using the Kirchhoff law in Crossbar arrays [11-12]. However, unconstrained ANNs do not benefit from activation sparsity, the input layer activations being usually non-null values. Therefore, all the synapses are active during the NN inference, which requires massive and repeated reading of the memory devices. On the other hand, SNNs encode the information using sparse temporal events (called spikes) [13-15]. SNN neurons integrate input spikes and produce an output spike when reaching a threshold. Spikes being produced with high sparsity, SNNs can benefit from event-based implementations, where computations and memory accesses are triggered only in the presence of a spike event [16]. Therefore, the high spike sparsity allows to significantly reduce the amount of read operations in the memory device per inference. Altogether, both ANNs and SNNs computational models are summarized in Fig. 1A, and the 1S1R crossbar array for synaptic weight implementation is illustrated in Fig. 1B.

In this paper, we propose a review of 1S1R-based Crossbar arrays operated in the sub-threshold regime. Focusing on experimental results on HfO_2-based RRAM (OxRAM) memory device co-integrated with an OTS selector, the pertinence of this approach for Neural Network inference implementation in hardware is discussed.

First, the 1S1R device operating characteristics in the sub-threshold regime are illustrated. Second, two different strategies to achieve high memory capacity are discussed, relying on the OTS thickness engineering and 1S1R cell size downscaling. Third, the ability to achieve massively parallelized read operation in the Crossbar arrays using the 1S1R sub-threshold reading strategy is explored. Fourth, the pertinence of 1S1R-based Crossbar arrays operated in the sub-threshold regime for

NNs implementation is discussed, considering both the standard Binarized Neural Networks (BNNs) and bio-inspired Binarized Spiking Neural Networks (BSNNs). In this context, BNN and BSNN figures of merit are explored, by means of off-chip network training simulations to perform the standard MNIST handwritten digit recognition task. Based on this analysis, general guidelines for both BNN and BSNN architecture design are illustrated.

All in all, the pertinence of the 1S1R Crossbar arrays operation in the sub-threshold regime for Neural Network inference applications is elucidated, opening the way for a system-level investigation towards specific applications. Moreover, the need to carry out a parallel co-development between the memory device and the Neural Network characteristics to optimize the overall circuit performance is illustrated.

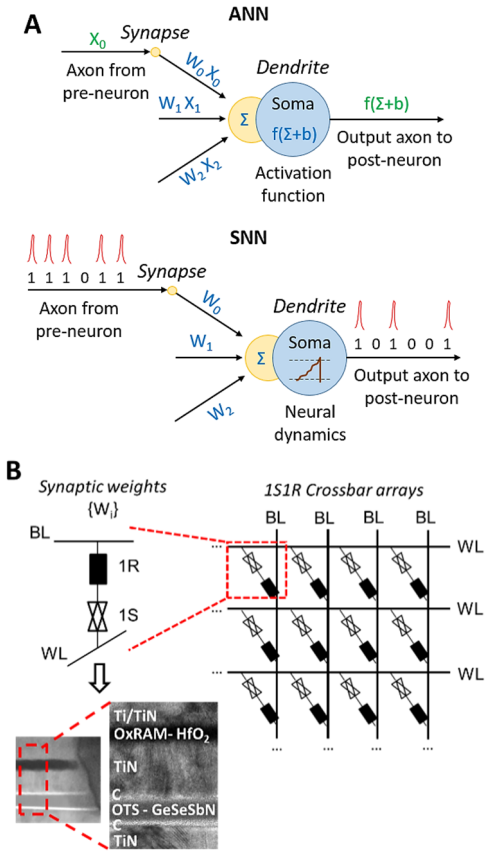


Fig. 1 A) Typical ANNs and SNNs computation models schematic summary. Adapted from [17]. B) Illustrative scheme of 1S1R Crossbar arrays, which are considered for Neural Network synaptic weights hardware implementation. Experimental data in this paper is extracted from OxRAM+OTS 1S1R devices, whose TEM characteristic is illustrated.

II. Results and Discussion

A. 1S1R sub-threshold read operation in Crossbar arrays

Fig. 2A presents a typical 1S1R current-voltage characteristics after the initial forming operation, required to initialize both OTS and OxRAM devices. The 1S1R switching voltages are strongly impacted by the OxRAM memory resistive state. On one hand, if the memory is at the Low Resistive State (LRS), the threshold couple (I_{th},$V_{th\text{-}LRS}$) is required for 1S1R switching process. On the other hand, if the memory is at the High Resistive State (HRS), the threshold couple (I_{th},$V_{th\text{-}HRS}$ > $V_{th\text{-}LRS}$) is required for 1S1R switching process due to additional voltage drop on the OxRAM device. Additionally, taking advantage of the OTS bipolar characteristics, a negative bias V_{RESET} is required to perform RESET operation (LRS-to-HRS transition) on the OxRAM device. In this context, the application of a reading bias V_{read} < $V_{th\text{-}LRS}$ allow to perform a 1S1R read operation in the sub-threshold regime (Fig. 2B), which prevents the OTS selector switching during the operation. Interestingly, the 1S1R sub-threshold read current is strongly dependent on the OxRAM resistive state. When the memory is at the LRS (resp. HRS), I_{LRS} (resp. I_{HRS}) currents are read. Hence, the I_{LRS}/I_{HRS} ratio represents the 1S1R sub-threshold read current margin. Remarkably, the more V_{read} becomes closer to $V_{th\text{-}LRS}$, the more I_{LRS} becomes closer to I_{th} and the larger the resulting 1S1R read current margin [9]. Therefore, ideal sub-threshold read configuration relies on the application of a reading voltage V_{read} equal to $V_{th\text{-}LRS}$. In this configuration, I_{LRS} becomes equal to I_{th}.

In this context, one of the most promising strategies to perform the read operation of an individual 1S1R device embedded in a Crossbar array structure relies on the $V_{read}/2$ biasing scheme (Fig. 2C). In this configuration, only the

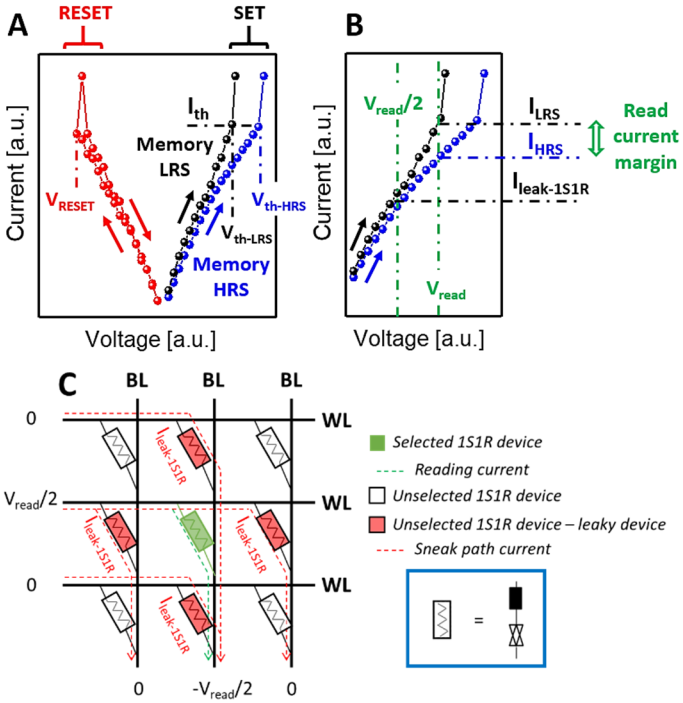


Fig. 2 A) Typical 1S1R quasi-static current-voltage characteristics. The resulting 1S1R switching voltages are strongly influenced by the memory resistive state. If the memory is at the High Resistive State, $V_{th\text{-}HRS}$ is required for 1S1R device switching operation. If the memory is at the Low Resistive State, $V_{th\text{-}LRS}$ is required for 1S1R device switching operation. B) 1S1R sub-threshold reading operation description, enabling to perform 1S1R read operation while preventing the OTS selector switching. A reading voltage V_{read}, lower than $V_{th\text{-}LRS}$, is required for reading operation. If the memory is at the Low Resistive State, I_{LRS} current is read. If the memory is at the High Resistive State, I_{HRS} current is read. Therefore, the I_{LRS}/I_{HRS} ratio corresponds to the 1S1R sub-threshold read current margin. C) $V_{read/2}$ 1S1R Crossbar array biasing scheme of interest, used to perform the reading operation on an individual 1S1R device in the array. In this biasing configuration, the neighboring 1S1R devices sharing the same Crossbar Word-Line and Bit-Line contribute to the overall leakage current in the array.

neighboring unselected 1S1R devices sharing the same Word-Line (WL) or Bit-Line (BL) with the device of interest contribute to the overall leakage currents on the Crossbar ($I_{leak-Crossbar}$). Accordingly, the overall Crossbar leakage current scales linearly with the amount of WLs and BLs in the structure. $V_{read}/2$ being the bias at the origin of the leakage current, $I_{leak-1S1R}$ represents the current contribution of a single unselected leaky 1S1R device in the overall leakage current in the structure (Fig. 2B & Fig. 2C).

B. 1S1R device optimization for high capacity Crossbar arrays

Focusing on the 1S1R reading configuration introduced in the previous section, the ability to perform a read operation of an individual 1S1R device in a Crossbar environment is linked to the ability to satisfactorily distinguish both I_{LRS} and I_{HRS} currents during the operation. Therefore, the 1S1R reading operation in the sub-threshold regime is considered as prohibitive when the cumulated leakage currents on the Crossbar reach I_{LRS}*0.9. The dependence among the 1S1R leakage current, the 1S1R I_{LRS} currents and the maximal Crossbar bank size is presented in Fig. 3. On one hand, the smaller $I_{leak-1S1R}$, the larger the maximal acceptable bank size at parity of I_{LRS}. On the other hand, the bigger I_{LRS}, the larger the maximal acceptable bank size at parity of $I_{leak-1S1R}$. Interestingly, the impact of both $I_{leak-1S1R}$ and I_{LRS} on the overall Crossbar bank size is inversely proportional. As an example, a 10x $I_{leak-1S1R}$ reduction (at parity of I_{LRS}) impact on the Crossbar bank size is equivalent to a 10x I_{LRS} increase (at parity of $I_{leak-1S1R}$) one.

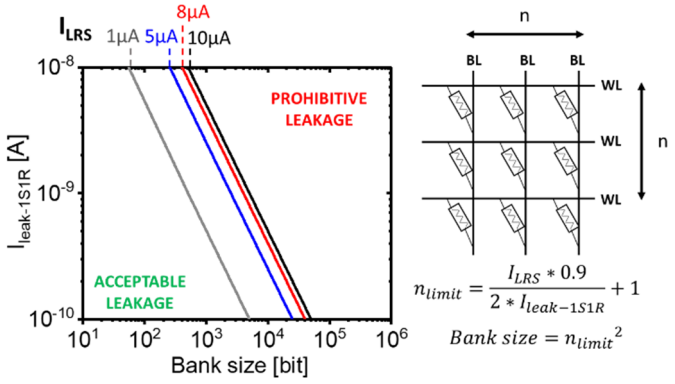


Fig. 3 1S1R acceptable leakage current for an increasing Crossbar bank size target, considering various 1S1R LRS currents. $V_{read/2}$ biasing strategy is considered for the analysis. Squared Crossbar arrays are considered. When the total leakage current in the Crossbar is equal to 0.9*I_{LRS}, the read operation in the 1S1R cell of interest is considered not to be possible. Both $I_{leak-1S1R}$ and I_{LRS} are shown to have opposite effects on the resulting 1S1R Crossbar bank size.

a. OTS thickness engineering for high-capacity Crossbar arrays

Therefore, to optimize both the 1S1R leakage currents and I_{LRS} currents remains essential to maximize the overall Crossbar capacity. Engineering the OTS selector thicknesses is one of the most common strategies to do so.

However, two main challenges limit the pertinence of this strategy for Crossbar capacity maximization. First, due to the existence of a tradeoff between the OTS selector leakage currents and programming voltages [18-19], $I_{leak-1S1R}$ reduction is obtained at the expense of higher 1S1R V_{th-LRS} switching voltages (Fig. 4A). Second, not only the OTS leakage currents but also the OTS threshold currents I_{th} decrease while increasing the OTS selector thicknesses [18]. Accordingly, considering ideal V_{read} choice, $I_{leak-1S1R}$ reduction is obtained at the expense of an I_{LRS} reduction (Fig. 4B). Nevertheless, $I_{leak-1S1R}$ currents are observed to scale faster than I_{LRS} with the OTS thickness characteristics, allowing the overall Crossbar capacity improvement. Overall, Fig. 4C summarizes the OTS thicknesses engineering impact on the 1S1R Crossbar design. Indeed, to increase the OTS selector thicknesses is demonstrated to satisfactorily improve the overall Crossbar capacity with the degradation of the 1S1R operating voltages as a counterpart, what goes against the 1S1R integration feasibility on CMOS advanced node technologies.

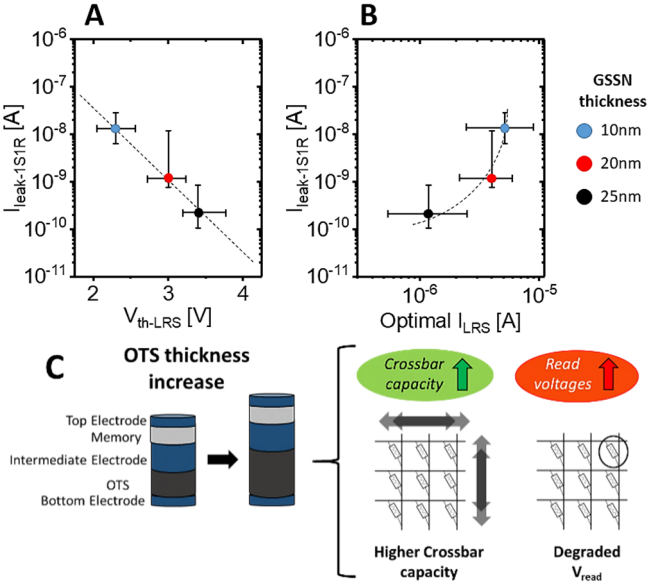


Fig. 4 A) (resp. B)) 1S1R leakage currents (resp. $I_{leak-1S1R}$) evolution with the device switching voltages (resp. device optimal LRS currents). The memory is considered to be read ideally, meaning V_{read} chosen identically as the 1S1R V_{th-LRS} switching values. The color code indicates various OTS selector thicknesses. On one hand, a tradeoff between the 1S1R switching voltages and leakage currents exist [18-19]. On the other hand, a correlation between the 1S1R optimal I_{LRS} and leakage currents exist. C) OTS thicknesses increase influence on 1S1R-based Crossbar design. First, the overall Crossbar capacity is increased by OTS thicknesses enlargment. Second, the 1S1R reading voltages degrade for thicker OTS selectors.

b. 1S1R scaling perspectives for high-capacity Crossbar arrays

Moreover, 1S1R cell size characteristics engineering is an additional lever to maximize the overall Crossbar capacity [18]. In this context, Fig. 5A presents both the $I_{leak-1S1R}$ and I_{LRS} evolution with the 1S1R cell section. Remarkably, $I_{leak-1S1R}$ is demonstrated to scale faster than I_{LRS} with 1S1R cell size, allowing to satisfactorily improve the overall Crossbar capacity. In addition, Fig. 5B presents the V_{th-LRS} switching voltages evolution with the 1S1R cell section. Interestingly, V_{th-LRS} does not scale with 1S1R cell size, which strongly simplifies the 1S1R operation in the array and prevents the electrical consumption associated to device read operation to degrade. Overall, Fig. 5C summarizes the 1S1R downscale impact on the 1S1R Crossbar design. Indeed, to shrink the 1S1R cell dimensions is demonstrated to satisfactorily improve the overall Crossbar capacity with no counterpart in terms of reading voltages, which enhances the 1S1R integration on CMOS advanced node technologies.

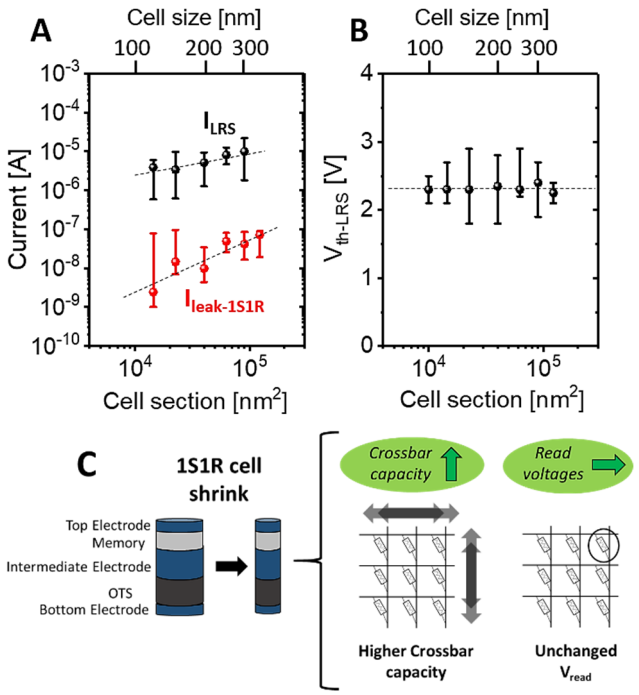


Fig. 5 A) (resp. B)) I_{LRS} and $I_{leak-1S1R}$ (resp. V_{th-LRS}) evolution with 1S1R cell section. Experimental tests performed on 15nm-thick OTS selectors are used for the analysis, considering the influence of the memory LRS resistance in the OTS sub-threshold regime negligible [18]. $I_{leak-1S1R}$ is observed to scale faster than I_{LRS} with respect to cell dimensions. 1S1R switching voltages remain constant with cell shrink. Therefore, 1S1R read voltages remains unchanged with cell dimensions reduction. C) 1S1R cell shrink influence on 1S1R-based Crossbar design. The overall Crossbar capacity is demonstrated to improve with no counterpart in terms of reading voltages.

C. 1S1R read parallelization in high capacity Crossbar arrays for low latency inference

The enlargement of the 1S1R Crossbar bank size increasing the amount of devices accommodated in the structure, it induces an overall computation time degradation on the system. To deal with it, achieving highly parallelized 1S1R read operation on the array remains key. In this context, the ability to perform parallelized 1S1R readings in Crossbar arrays is strongly dependent on the ability to satisfactorily supply sufficient current on the active devices during the operation. Therefore,

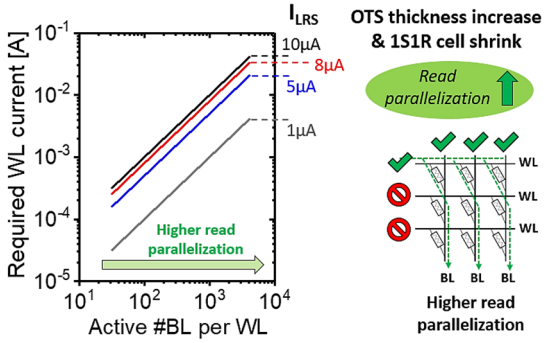


Fig. 6 Required current per Word Line evolution with the amount of active Bit Lines per Word Line on the array. 50% of the cells per WL are considered to be at the LRS. A read current margin of 10 is considered for the analysis [9]. Various I_{LRS} read currents are considered. Allowing to reduce the 1S1R device sub-threshold read currents I_{LRS} and I_{HRS}, OTS thickness increase and 1S1R cell shrink are demonstrated to be strategies to consider for an overall computation time optimization in high capacity Crossbar arrays.

the read parallelization feasibility in the array is mostly governed by the maximal current density rules in metal lines for the technological node of interest. Fig. 6 presents the dependence between the required current per WL and the amount of active BLs in the Crossbar, considering various I_{LRS} values. Reducing the I_{LRS} current on the 1S1R devices decreases the required WL current, and hence increases the maximum number of BL that can be activated per WL in parallel. Accordingly, OTS thickness increase and 1S1R cell shrink, allowing to reduce I_{LRS}, enhance highly parallelized read operation in the arrays.

D. 1S1R pertinence for state of art Neural Networks hardware implementated inference

Both device and Crossbar design optimization aspects have been introduced in the previous sections. In order to fully leverage 1S1R sub-threshold reading approach in NNs hardware implementation, it is essential to consider the device characteristics during the NN design. To do so, off-chip training simulations on a fully connected Neural Network with one hidden layer are performed, focusing on the standard MNIST handwritten digit recognition task (Fig. 7A). Different number of neurons on the hidden layer (X=[512 ; 1024 ; 4096]) are considered for the analysis.

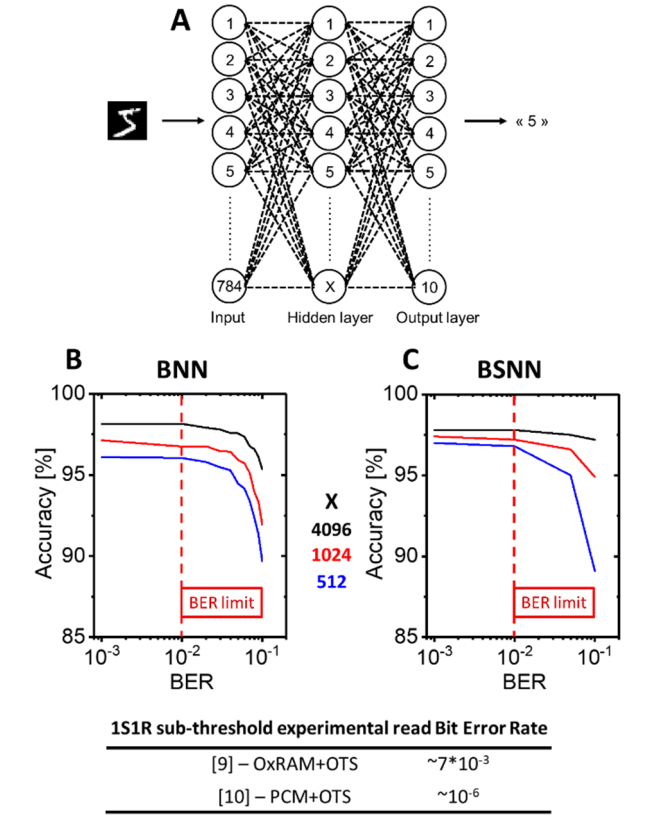
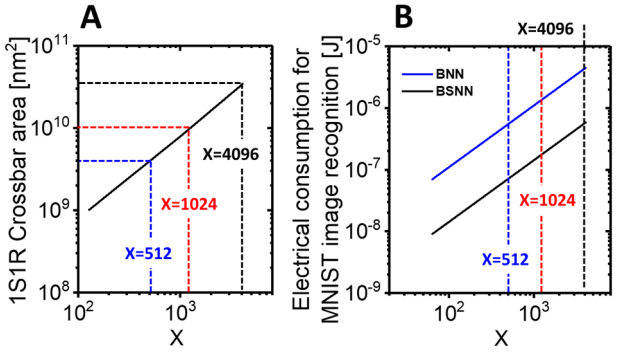


Fig. 7 A) Fully connected neural network topologies with one hidden layer of X neurons are considered in this paper, considering both Binarized Neural Network (BNN) and Binarized Spiking Neural Network (BSNN). The MNIST handwritten digit recognition task is considered. B) (resp. C)) BNN (resp. BSNN) accuracy evolution with the synaptic binarized Bit Error Rate (BER). No adapted training strategy to the BER is considered in this study. The experimental results provided in [9-10] are demonstrated to be perfectly tolerated by the BNN (resp. BSNN) networks.

Two different types of NNs are explored: Binarized Neural Network (BNN) and Binarized Spiking Neural Network (BSNN). In both cases, binary synaptic weights (+1 or -1) are considered, where 1S1R HRS (resp. LRS) states are used to encode +1 (resp. -1) weights. Concerning the BSNN, the pixels of input images are converted to spikes following the procedure described in [20].

Errors are simulated during inference in both BNN and BSNN synaptic weights, allowing to evaluate the impact of 1S1R non-idealities on the network accuracy [20-22]. The resulting accuracy evolution with the 1S1R Bit Error Rate (BER) for both BNN and BSNN is provided in Fig. 7B and Fig. 7C, respectively. Altogether, both BNN and BSNN follow similar trends. First, increasing the number of neurons in the hidden layer is demonstrated to improve both BNN and BSNN maximal attainable accuracy for the task of interest. Second, a similar maximal tolerable BER of ~10^{-2} is demonstrated for both BNN and BSNN, which guarantees an optimal tolerance to the 1S1R experimental sub-threshold reading Bit Error Rates published in the literature [9-10].

Moreover, Fig. 8A presents the evolution of the Crossbar footprint as a function of the number of neurons in the hidden layer. A CD_{min} metal width and space between metal lines is considered, focusing on the 28nm technological node. The Crossbar footprint increasing with the number of neurons in the hidden layer, a tradeoff between the network maximal attainable accuracy and its area capabilities is illustrated (Fig. 7B, Fig. 7C and Fig. 8A). In addition, Fig. 8B presents the electrical consumption for MNIST image classification evolution with the number of neurons in the hidden layer for both BNN and BSNN networks of interest. Remarkably, BSNNs promise higher energy efficiency than BNNs, due to a high spike sparsity in the network (87%, meaning 0.13 spikes per neuron per inference, measured on average on these experiments).



1S1R reading conditions [9]	
Reading LRS current	10µA
Reading HRS current	1µA
Reading voltage	2.5V
Reading time	100ns

Fig. 8 A) Crossbar footprint evolution with the number of neurons in the hidden layer of the network. The area required for peripherals accomodation is not considered. B) Electrical consumption, associated to 1S1R device read operation, per MNIST image classification evolution with the number of neurons in the hidden layer of the network. The considered reading conditions are provided, based on experimental results published in [9]. Both BNN and BSNN (with 0.13 spikes per neuron per inference) are considered.

III. Conclusion

The pertinence of 1S1R-based Crossbar arrays operated in the sub-threshold regime for Neural Network inference implementation in hardware is discussed, by coupling Neural Network off-chip training simulations with experimental data. At the device level, both OTS thickness engineering and 1S1R cell downscale have been demonstrated to be valuable parameters to optimize the overall Crossbar capacity and reading latency capabilities. At the Neural Network level, general guidelines for both BNN and BSNN architecture design have been provided, targeting enlarged inference accuracy, improved tolerance to 1S1R bit errors, optimized energy efficiency and reduced inference latency. Overall, this is a first step for Neural Network hardware implementations based on 1S1R-based Crossbar arrays.

Acknowledgment

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

References

[1] A. Pedram, S. Richardson, M. Horowitz, S. Kvatinsky and S. Galal, "Dark Memory and Accelerator-Rich System Optimization in the Dark Silicon Era", *IEEE Design Test* 2017, *34*, 39-50.

[2] V. Sze, "Efficient Processing of Deep Neural Networks: from Algorithms to Hardware Architectures", presented at *NEURIPS*, Vancouver, Canada, 2019.

[3] W. H. Chen, C. Dou, K. X. Li, W. Y. Lin, P. Y. Li, J. H. Huang, J. H. Wang, W. C. Wei, C. X. Xue, Y. C. Chiu, Y. C. King, C. J. Lin, R. S. Liu, C. C. Hsieh, K. T. Tang, J. J. Yang, M. S. Ho and M. F. Chang, "CMOS-integrated memristive non-volatile computing in-memory for AI edge processors", *Nat. Electron.* 2019, *2*, 420-428.

[4] P. Jain, U. Arslan, M. Sekhar, B. C. Lin, L. Wei, T. Sahu, J. Alzate-Vinasco, A. Vangapaty, M. Meterelliyoz, N. Strutt, A. B. Chen, P. Hentges, P. A. Quintero, C. Connor, O. Golonzka, K. Fischer and F. Hamzaoglu, "2 Embedded Non-Volatile ReRAM Macro in 22nm FinFET Technology with Adative Forming/Set/Reset Schemes Yielding Down to 0.5V with Sensing Time of 5ns at 0.7V, presented at *ISSCC*, San Francisco, CA, USA, 2019, pp 212-214.

[5] C. Chou, Z. Lin, C. Lai, C. Su, P. Tseng, W. Chen, W. Tsai, W. Chu, T. Ong, H. Chuang, Y. Chih and T. J. Chang, "A 22nm 96KX144 RRAM Macro with a Aelf-Tracking Reference and a Low Ripple Charge Pump to Achieve a Configurable Read Window and a Wide Operating Voltage Range", presented at *VLSI Circuits*, Honolulu, HI, USA, 2020, pp. 1-2.

[6] 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", presented at *IEDM*, San Francisco, 2019, pp 14.3.1–14.3.4.

[7] T. Hirtzlin, M. Bocquet, B. Penkovsky, J. O. Klein, E. Nowak, E. Vianello, J. M. Portal and D. Querlioz, "Digital biologically plausible implementation of binarized neural networks with differential hafnium oxide resistive memory arrays", *Front. Neurosci.* 2020, *13*, 1383.

[8] D. Kau, S. Tang, I. V. Karpov, R. Dodge, B. Klehn, J. A. Kalb, J. Strand, A. Diaz, N. Leung, J. Wu, S. Lee, T. Langtry, K. Chang, C. Papagianni, J. Lee, J. Hirst, S. Erra, E. Flores, N. Righos, H. Castro and G. Spadini, "A stackable cross point phase change memory", presented at *IEDM*, Baltimore, MD, USA, 2009, pp. 1-4.

[9] 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, M. Bocquet, J. M. Portal, G. Molas and F. Andrieu, "1S1R sub-threshold operation in Crossbar arrays for low power BNN inference computing", presented at *IMW*, Dresden, Germany, May 2022, pp. 1-4.

[10] N. Lepri, P. Gibertini, P. Manocci, A. Pirovano, I. Tortorelli, P. Fantini and D. Ielmini, “In-memory neural network accelerator based on phase change memory (PCM) with one-selector/one-resistor (1S1R) structure operated in the subthreshold regime”, presented at *IMW*, Monterey, CA, USA, May 2023, pp. 1-4.

[11] D. Strukov, G. Indiveri, J. Grollier and S. Fusi, “Building brain-inspired computing”, *Nat. Commun.* 2019, 10, 4838.

[12] D. Garbin, E. Vianello, E. Bichler, Q. Rafhay, G. Ghibaudo, B. De Salvo and L. Perniola, “HfO2-based OxARM devices as sunapses for convolutional neural networks”, *IEEE Trans. Electron Devices* 2015, *62*, 2494-501.

[13] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, F. Chen, N. Deng, S. Wu, Y. Wang, Y. Wu, Z. Yang, C. Ma, G. Li, W. Han, H. Li, H. Wu, R. Zhao, Y. Xie and L. Shi, “Towards artificial general intelligence with hybrid Tianjic chip architecture”, *Nature* 2019, *572*, pp. 106-111.

[14] M. Dampfhoffer, T. Mesquida, A. Valentian and L. Anghel, "Are SNNs Really More Energy-Efficient Than ANNs? an In-Depth Hardware-Aware Study," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 3, pp. 731-741, 2023.

[15] M. Dampfhoffer, T. Mesquida, A. Valentian and L. Anghel, “Backpropagation-based Learning Techniques for Deep Spiking Neural Networks: A Survey”, *IEEE Transactions on Neural Networks and Learning Systems,* 2023.

[16] M. Davies, N. Srinivasa, T. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C. Lin, A. Lines, R. Liu, D. Mathaikutty, S. McCoy, A. Paul, J. Tse, G. Venkataramanan, Y. Weng, A. Wild, Y. Yang ang H. Wang, “Loihi: A Neuromorphic Manycore Processor with On-Chip Learning”, *IEEE* 2018, *38*, pp. 82-99.

[17] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, F. Chen, N. Deng, S. Wu, Y. Wang, Y. Wu, Z. Yang, C. Ma, G. Li, W. Han, H. Li, H. Wu, R. Zhao, Y. Xie and L. Shi, “Towards artificial general intelligence with hybrid Tianjic chip architecture”, *Nature* 2019, *572*, pp. 106-111.

[18] J. Minguet Lopez, N. Castellani, L. Grenouillet, L. Reganaz, G. Navarro, M. Bernard, C. Carabasse, T. Magis, D. Deleruyelle, M. Bocquet, J. M. Portal, E. Nowak and G. Molas, Ge-Se-Sb-N-based OTS scaling perspectives for high-density 1S1R crossbar arrays, presented at *IMW*, Dresden, Germany, May 2021, pp. 1-4.

[19] A. Verdy, M. Bernard, N. Castellani, P. Noe, J. Garrione, G. Bourgeois, M. C. Cyrille, G. Navarro and E. Nowak, “Tunable Performances in OTS Selectors Thanks to Ge3Se7-As2Te3”, presented at *IMW*, Monterey, CA, USA, 2019, pp. 1-4.

[20] J. Minguet Lopez, T. Hirtzlin, M. Dampfhoffer, L. Grenouillet, L. Reganaz, G. Navarro, C. Carabasse, E. Vianello, T. Magis, D. Deleruyelle, M. Bocquet, J. M. Portal, F. Andrieu and G. Molas, “OxRAM+OTS optimization for binarized neural network hardware implementation”, *Semicond. Science Tech*. 2022, 37, 014001.

[21] J. Minguet Lopez, Q. Rafhay, M. Dampfhoffer, L. Reganaz, N. Castellani, V. Meli, S. Martin, L. Grenouillet, G. Navarro, T. Magis, C. Carabasse, T. Hirtzlin, E. Vianello, D. Deleruyelle, J. M. Portal, G. Molas and F. Andrieu, “1S1R optimization for high-frequency inference on Binary Spiking Neural Networks”, *Advanced Electronic Materials* 2022, 2200323.

[22] M. Dampfhoffer, J. Minguet Lopez, T. Mesquida, A. Valentian and L. Anghel, “Improving the Robustness of Neural Networks to Noisy Multi-Level Non-Volatile Memory-based Synapses”, *IEEE International Joint Conference on Neural Networks (IJCNN),* 2023.