



1S1R optimization for high-frequency inference on binarized spiking neural networks

Joel Minguet Lopez, Quentin Rafhay, Manon Dampfhooffer, Lucas Reganaz, Niccolo Castellani, Valentina Meli, Simon Martin, Laurent Grenouillet, Gabriele Navarro, Thomas Magis, et al.

► To cite this version:

Joel Minguet Lopez, Quentin Rafhay, Manon Dampfhooffer, Lucas Reganaz, Niccolo Castellani, et al.. 1S1R optimization for high-frequency inference on binarized spiking neural networks. Advanced Electronic Materials, 2022, 2022 (8), pp.2200323. 10.1002/aelm.202200323 . cea-03707409

HAL Id: cea-03707409

<https://cea.hal.science/cea-03707409v1>

Submitted on 28 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1S1R Optimization for High-Frequency Inference on Binarized Spiking Neural Networks

Joel Minguet Lopez,* Quentin Rafhay, Manon Dampfhofer, Lucas Reganaz, Niccolo Castellani, Valentina Meli, Simon Martin, Laurent Grenouillet, Gabriele Navarro, Thomas Magis, Catherine Carabasse, Tifenn Hirtzlin, Elisa Vianello, Damien Deleruyelle, Jean-Michel Portal, Gabriel Molas, and François Andrieu

Single memristor crossbar arrays are a very promising approach to reduce the power consumption of deep learning accelerators. In parallel, the emerging bio-inspired spiking neural networks (SNNs) offer very low power consumption with satisfactory performance on complex artificial intelligence tasks. In such neural networks, synaptic weights can be stored in nonvolatile memories. The latter are massively read during inference, which can lead to device failure. In this context, a 1S1R (1 Selector 1 Resistor) device composed of a HfO_2 -based OxRAM memory stacked on a Ge-Se-Sb-N-based ovonic threshold switch (OTS) back-end selector is proposed for high-density binarized SNNs (BSNNs) synaptic weight hardware implementation. An extensive experimental statistical study combined with a novel Monte Carlo model allows to deeply analyze the OTS switching dynamics based on field-driven stochastic nucleation of conductive dots in the layer. This allows quantifying the occurrence frequency of OTS erratic switching as a function of the applied voltages and 1S1R reading frequency. The associated 1S1R reading error rate is calculated. Focusing on the standard machine learning MNIST image recognition task, BSNN figures of merit (footprint, electrical consumption during inference, frequency of inference, accuracy, and tolerance to errors) are optimized by engineering the network topology, training procedure, and activations sparsity.

in graphics processing units (GPUs) or central processing units (CPU) remains extremely challenging. On the contrary, the brain promises very high cognitive capacity while preserving exceptional energy efficiency. One of the main differences between GPUs or CPUs and the brain is memory management. On the one hand, a physical separation exists in GPUs and CPUs between arithmetic and storage units, which is at the origin of enormous energy consumption associated with data transfer between both units.^[1] This trend is particularly exacerbated for Artificial Neural Networks (ANNs), which require a very large amount of memory access. On the other hand, the biological neurons and synapses are close to each other in the brain. Accordingly, developing non-von-Neumann architectures to perform in or near-memory computing with non-volatile memory (NVM) technologies is one of the most promising strategies to improve the energetic efficiency of artificial intelligence.^[2] Another relevant dif-

ference between artificial processors and the brain is the way information is coded. On the one hand, GPUs and CPUs rely on high-precision floating-point outputs. On the other hand, binarized sparse and asynchronous spikes are used to communicate in the brain. In particular, a neuron receives through

1. Introduction

Deep learning accelerators have attained impressive performance on various cognitive tasks, such as image or audio recognition. However, its enormous electrical consumption

J. Minguet Lopez, L. Reganaz, N. Castellani, V. Meli, S. Martin, L. Grenouillet, G. Navarro, T. Magis, C. Carabasse, T. Hirtzlin, E. Vianello, G. Molas,^[†] F. Andrieu
CEA
LETI

MINATEC Campus, 17 rue des Martyrs, Grenoble 38054, France
E-mail: joel.minguetlopez@cea.fr

Q. Rafhay

Univ. Grenoble Alpes
Univ. Savoie Mont Blanc
CNRS

Grenoble INO, IMEP-LAHC, 3 Parvis Louis Néel, Grenoble 38000, France



The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/aelm.202200323>.

^[†]Present address: Weebit Nano Ltd., MINATEC Campus, 17 rue des Martyrs, 38054 Grenoble, France

M. Dampfhofer
Univ. Grenoble Alpes
CEA
CNRS

Grenoble INP, INAC-Spintec, 17 rue des Martyrs, Grenoble 38054, France
D. Deleruyelle

INL CNRS
INSA Lyon

7 Avenue Jean Capelle, Villeurbanne 69621, France

J.-M. Portal
Aix Marseille Univ
CNRS

IM2NP, 5 rue Enrico Fermi, Marseille 13009, France

DOI: 10.1002/aelm.202200323

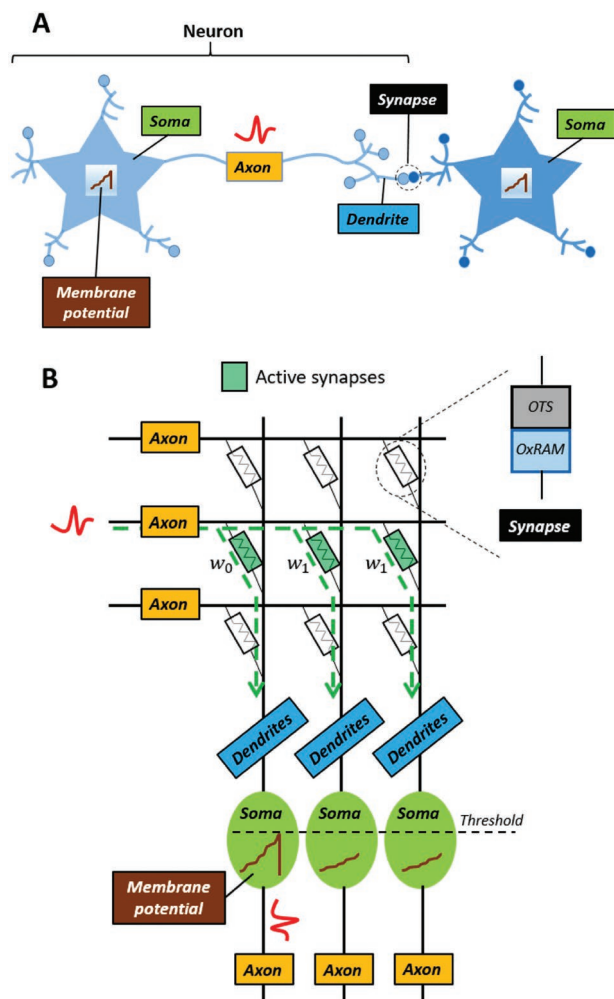


Figure 1. A) Biological neuron dynamics. An input spike flows through the neuronal dendrite until it is integrated into the soma and contributes to increase the membrane potential. When the neuron membrane potential reaches a certain threshold, an output spike is generated and transmitted to other neurons. Then, the neuron returns to its resting state. B) A 1S1R-based single memristor crossbar array is proposed in this work to implement the binarized spiking neural network synaptic weights.

a synapse an input spike, which flows through the dendrite until reaching the soma. This input spike is integrated into the soma of the neuron and contributes to increasing its membrane potential. When the neuron membrane potential reaches a certain threshold, it generates a spike that is transmitted to other neurons (post-neurons) through the axon. As a consequence, the neuron returns to its resting state (Figure 1A). The emerging bio-inspired spiking neural networks (SNNs) provide the opportunity to emulate these characteristics in artificial systems. Indeed, while classical ANNs rely on layer-by-layer multiply-and-accumulate operations between input activations and weights, SNNs computation is mostly based on the accumulation of spikes weighted by synaptic connections.

Therefore, only Accumulate operations are required in SNNs. In addition, SNNs can be processed in an event-based manner to benefit from their natural sparsity. Indeed, synaptic weight readings are triggered by the spikes received by a neuron. One

should notice that the synaptic weights are not read in the absence of input spikes. Therefore, a low spiking activity on the SNN results in very low energy consumption compared to an ANN where all synaptic weights are read at each inference.

Various NVM devices have been considered for NN synaptic weight hardware implementation in crossbar arrays.^[3–8] Nevertheless, at the moment, there is no ideal candidate due to NVM technologies imperfection, which can strongly limit the overall accelerator inference accuracy.^[9] Accordingly, considerable efforts have been invested in the precedent years to develop innovative techniques to deal with memory non-idealities.^[10] Among others, resistive random-access memory (RRAM) devices are one of the most promising technologies proposed for NN hardware implementation.^[10–12] However, the RRAM intrinsic operating variability remains challenging. In this context, innovative bit-error correcting codes and adaptive programming schemes have been developed.^[13,14] In addition, innovative and aggressive programming patterns have been considered to reduce device variability. Nevertheless, an important degradation of the device operating endurance capabilities appears as a counter-part.^[15,16]

In this work, we focus on low-precision Binarized Spiking Neural Networks (BSNNs), which benefit from very high energy efficiency while promising satisfactory tolerance to RRAM non-idealities.^[17] In this type of network, the synaptic weights are implemented by binarized values (–1 and +1) after the training process. Moreover, the standard 1T1R (1 Transistor 1 Resistor) architectures used to implement synaptic weights with memory devices are replaced by a denser 1S1R (1 Selector 1 Resistor) stack, where the memory (1R) is co-integrated in series with a back-end selector (1S). About one order of magnitude improvement in terms of memory density can be achieved with 1S1R with respect to conventional 1T1R architectures.^[18] However, RRAM co-integration with a back-end selector implies several challenges. First, the process integration complexity increases. Second, achieving high memory capacity while preserving reduced electrical consumption becomes challenging due to a trade-off between programming voltages and leakage currents. Third, to achieve high precision on device resistive states implies a degradation of its programming endurance capabilities due to a trade-off between memory window and endurance capabilities. To deal with those challenges, a specific design of both memory stack and applied programming conditions for the application of interest is required.^[19–22]

In this context, the co-integration of an HfO_2 -based OxRAM device with an ovonic threshold switch (OTS) selector on single memristor crossbar arrays to implement the BSNN synaptic weights (Figure 1B)) is proposed. OTS co-integration with HfO_2 -based OxRAM has been satisfactorily demonstrated in the previous years.^[18,20–23] Through stack design and applied programming conditions adaptation, the 1S1R dynamic switching capabilities have been elucidated, the 1S1R binarized window margin optimized, and its programming endurance capabilities enlarged.^[22] In particular, the 1S1R pertinence for standard low-precision synaptic weight encoding during network training on-chip has been demonstrated.^[18] However, beyond 1S1R-based crossbar arrays pertinence for low-precision neural network training, it remains essential to elucidate the 1S1R ability to perform reliable high-frequency inference on-chip. Inference in

SNNs requires a huge amount of repeated and frequent reading operation of 1S1R devices to sense the network synaptic weights. While reading is expected to be nondestructive, it can induce erratic switching of the selector and memory devices and affect network performance. This reliability issue has never been addressed in the literature, and the link between device physics and circuit accuracy has not been studied and clarified so far.

In this work, we first present an experimental statistical study on OTS switching probability when repeated sub-threshold voltages are applied and propose to reduce its variability through device lifetime by optimizing the applied voltage on the devices. To gain insight into the OTS switching operation microscopic mechanisms, a Monte Carlo statistical model based on Bernoulli's conduction point is implemented. The overall 1S1R reading reliability being directly linked to the OTS switching variability, and general guidelines on 1S1R reading conditions (applied voltages, reading frequency) optimization for low reading bit error rate (BER) are provided based on this experimental and theoretical statistical study. To evaluate the 1S1R pertinence for BSNN inference hardware implementation, training simulations are performed on one hidden layer fully connected BSNN for an image classification task on the MNIST dataset. By introducing errors in the synaptic weights during training, an optimized BSNN tolerance to synaptic errors for the devices of interest is demonstrated. Based on this analysis, guidelines for an optimized system footprint, a reduced electrical consumption, maximized inference frequency, and maximized BSNN accuracy for the MNIST task are provided.

2. Technological Details

OxRAM is co-integrated with an OTS back-end selector in a 4kb 1S1R array configuration, where a transistor is used to limit the current on the devices. The OTS is composed of a 10 nm-thick Ge-Se-Sb-N (GSSN) alloy, sandwiched between two Carbon electrodes, and is used as a selector device. Then, a 10 nm-thick HfO₂-based OxRAM deposited by Atomic Layer Deposition (ALD) is used as the memory device. This HfO₂ is deposited on a TiN inter-layer that separates the selector and the memory elements. A 10 nm Physical Vapor Deposition (PVD) Ti top layer acts as an oxygen scavenging layer for the memory. A TiN additional layer completes the top electrode. After the etching process, the memory dot is capped with a SiN layer for passivation. Afterward, the top contact and top metal line end the overall integration process. The 1S1R memory array is thus integrated into the Back-End-Of-Line (BEOL) of a 130 nm CMOS process between the fourth and fifth metal layers. **Figure 2A** provides an SEM image of the 1S1R devices integrated into the Back-End-Of-Line (BEOL), as well as a TEM cross-section for the stack of interest, which demonstrates a consistent co-integration of all the deposited layers.

3. Results and Discussion

3.1. 1S1R Programming and Reading Operation

Figure 2B provides typical 1S1R current-voltage characteristics after the device firing operation, which is required for both

OTS and OxRAM initialization. The 1S1R switching voltages are strongly impacted by the resistive state of the OxRAM. If the OxRAM is in the Low Resistive State (LRS) (resp. High Resistive State (HRS)), V_{th-LRS} (resp. $V_{th-HRS} > V_{th-LRS}$) is required for the 1S1R switching. The 1S1R read window margin is defined by $(V_{th-HRS} - V_{th-LRS})$. **Figure 2C** provides experimental programming endurance characteristics for the stack of interest for eight 1S1R devices. No error is observed until 10^4 programming cycles, demonstrating the ability to encode binarized weights on the memory. Since only inference is targeted on chip, weights have only to be programmed once per application, and thus 10^4 cycles are sufficient. Nevertheless, 1S1R stack adjustments and optimized programming patterns as well as OTS material engineering are paths of improvement to enhance endurance for more cycling demanding applications.^[22,23] In this context, the OxRAM resistive state can be read by applying a certain reading voltage (V_{read}) caught between V_{th-HRS} and V_{th-LRS} . If the OxRAM is in LRS, the OTS switching occurs, and I_{LRS} is read. If the OxRAM is in HRS, the OTS device remains at the OFF state, and I_{HRS} is read. For a 200 ns long standard trapezoidal reading pulse, **Figure 2D** shows experimental 1S1R reading disturb characteristics for a given relaxation time $t_{relax} \approx 100 \mu s$ between the applied pulses. No error is observed up to 10^9 cycles. However, provided the same reading pattern and V_{read} , **Figure 2E** presents experimental 1S1R HRS reading disturb characteristics when $t_{relax} \approx 1 \mu s$. Through the cycles, the read HRS current progressively increases until an erratic OTS switching occurs after 10^7 cycles, which corresponds to the first failing cycle noted $Cycle_{1st-fail}$. I_{HRS} is read on the subsequent cycles, which suggests that the OTS can nonetheless relax back to the OFF state and that this phenomenon is reversible. The apparition of reading fails on 1S1R devices being linked to the OTS switching reliability at V_{read} , and the main goal of the following sections is to explore the OTS dynamic switching behavior. The reading pulses characteristics provided in **Figure 2D,E** are used for the following sections.

3.2. 2D Monte Carlo Model for OTS Switching Operation

To study the OTS reading reliability from a phenomenological point of view, a novel 2D OTS reading endurance model is presented in this section. The main goal here is to statistically quantify the OTS switching statistics leading to the apparition of reading errors during device lifetime, which hence suggests a Monte Carlo simulation approach.

Conceptually, we focus on an OTS filamentary field-driven switching theory, where the OTS switching operation is described by the nucleation of metastable domains within the chalcogenide layer.^[24] In particular, after the initialization of the device, the application of a certain electrical field at the OTS terminals leads to the random nucleation of conductive GSSN dots in the chalcogenide layer. At a microscopic scale, it can be foreseen that these local conductive dots correspond to metastable metavalent bonding formations, resulting from a change in the local order and GSSN bond alignment.^[25] The alignment of these conductive dots implies the appearance

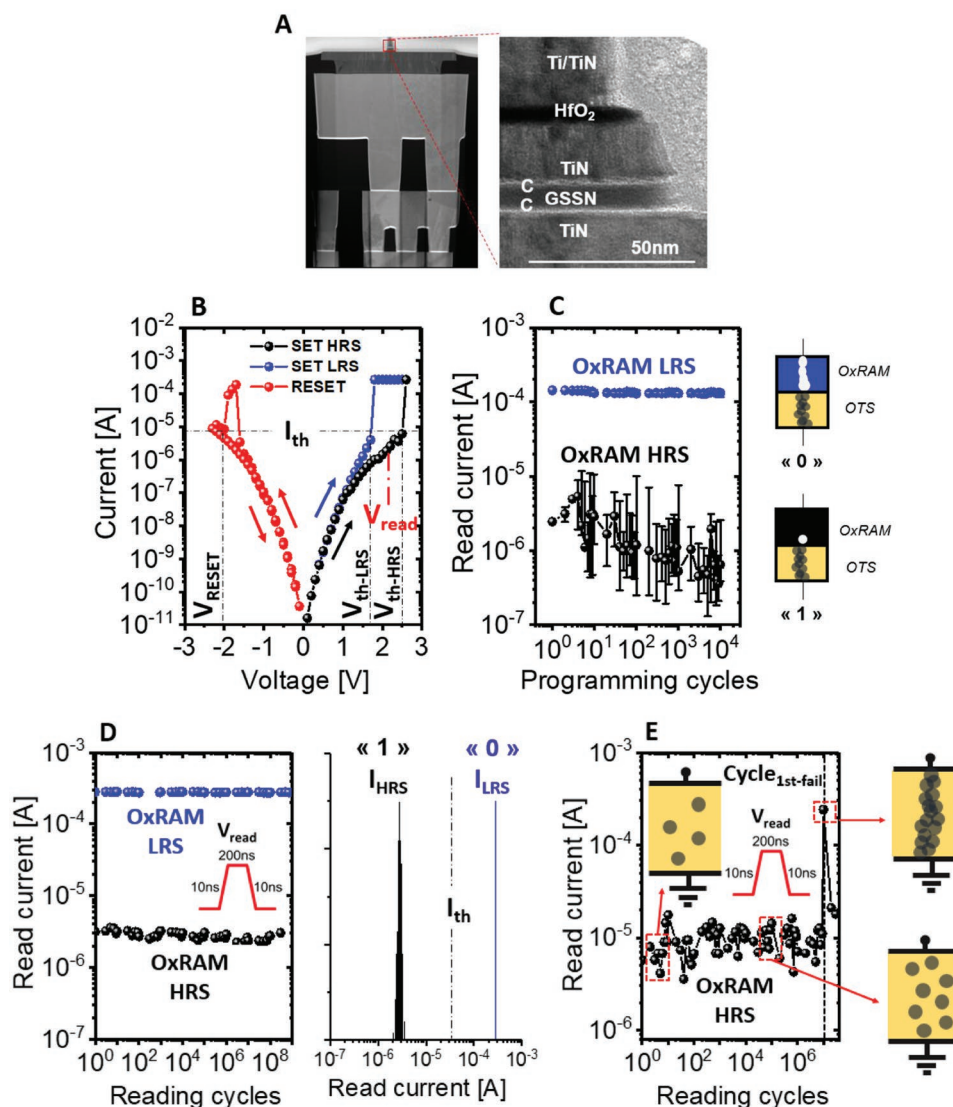


Figure 2. A) Technological details. SEM top view of the 1S1R devices integrated into the BEOL, together with a TEM cross-section of the 1S1R devices studied in this work. B) 1S1R typical current-voltage characteristics after the firing process. The 1S1R switching voltages are function of the OxRAM resistive state. If OxRAM is in low resistive state (LRS) (rep. high resistive state (HRS)), V_{th-LRS} (resp. $V_{th-HRS} > V_{th-LRS}$) is required for 1S1R switching operation. In this context, the application of a reading voltage V_{read} between V_{th-HRS} and V_{th-LRS} allows to identify the OxRAM resistive state. C) 1S1R programming endurance capabilities. Up to 10⁴ programming cycles are demonstrated without errors. D) Experimental 1S1R reading disturb characteristics at $V_{read} = 2.25$ V. A 100 μ s relaxation time between pulses is allowed to the device. No error is observed up to 10⁹ cycles. E) Experimental 1S1R high resistive state current reading disturb characteristics at $V_{read} = 2.25$ V. A short 1 μ s relaxation time between pulses is allowed to the device, which increases the probability of OTS erratic switching with respect to longer relaxation times. In this case, the OTS switches after 10⁷ reading cycles, leading to the appearance of a reading fail.

of a conductive filament and leads to OTS switching to the ON state. This approach is schematically summarized in Figure 3A.

Figure 3B schematically illustrates the OTS 2D Monte Carlo reading endurance model deployment, which remains compatible with the conceptual switching approach introduced previously. A homogeneous OTS bulk is considered at the beginning of each reading disturb simulation. Through the application of a virtual pulse, conductive dots can randomly appear in the GSSN matrix for every reading cycle. Between two successive cycles, a certain relaxation time is allowed for the device. On the

one hand, the nucleation of a conductive defect on the constriction layer is driven by a Bernoulli probability (P_n) for each bin of the GSSN matrix. On the other hand, the disappearance of a conductive defect in one bin is assumed to follow an exponential probabilistic distribution characterized by a time τ_d . Thus, conductive dots on the constriction zone are considered reversible and can thus switch back to a nonconductive state at every reading cycle. The vertical alignment of a column of conductive bins forms a conductive path, which is the condition chosen for reading endurance failure. To emulate the read cycling process, each position on the constriction zone is scanned at each

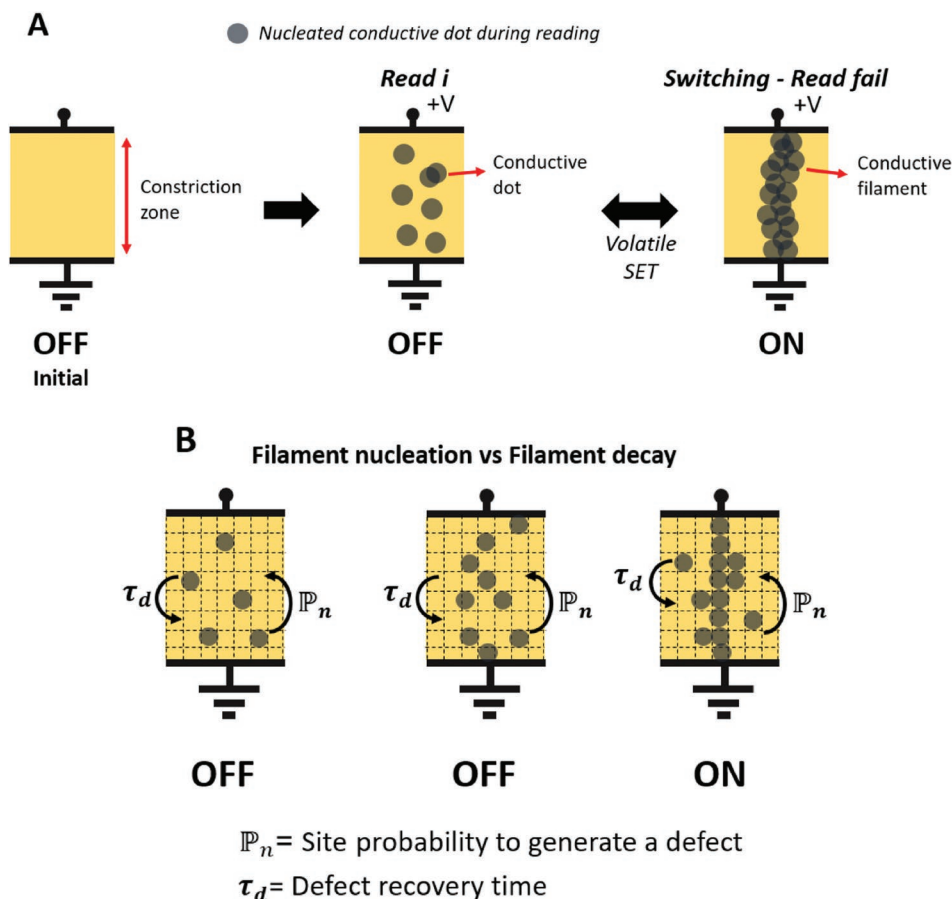


Figure 3. A) Conceptual description of the OTS switching operation considered in this work. The OTS switching operation is driven by an applied electrical field at the device terminals, which induces the nucleation of randomly spaced conductive dots on the bulk of the OTS.^[24] If the conductive defects are aligned, a conductive filament appears on the device and leads to the OTS switching. B) OTS Monte Carlo reading endurance model deployment. The nucleation of a conductive defect on the constriction layer is driven by a certain probability (P_n). The disappearance of a conductive dot on the constriction zone is driven by a certain recovery time (τ_d). The alignment of several defects implies the apparition of a conductive filament on the constriction zone and leads to a reading fail.

reading cycle, and its resistive state is updated based on P_n and τ_d simulation parameters. The first cycle at which a failure is detected corresponds to $\text{Cycle}_{1\text{st-fail}}$, whereas the number of failures over a given amount of cycles is used to estimate the switching probability.

3.3. OTS Switching Reliability Optimization

Figure 4A presents the $\text{Cycle}_{1\text{st-fail}}$ dependence with relaxation times and applied reading voltages on OTS devices. The box-plots represent the (25%; 75%) percentiles. Increasing V_{read} closer to the OTS intrinsic switching voltage lowers the median of $\text{Cycle}_{1\text{st-fail}}$. In addition, as the relaxation time between reading cycles increases, $\text{Cycle}_{1\text{st-fail}}$ median is shifted to higher values for a given V_{read} . This suggests a more robust OTS OFF state when a longer relaxation time between applied pulses is achieved. Moreover, **Figure 4B** presents experimental OTS $\text{Cycle}_{1\text{st-fail}}$ distributions for two distinct applied voltages for a corresponding t_{relax} of $\approx 1 \mu\text{s}$. One unique device is considered per distribution. Each point on the distribution corresponds to a unique sub-threshold reading disturb test until 10^8 reading cycles. In this context, the

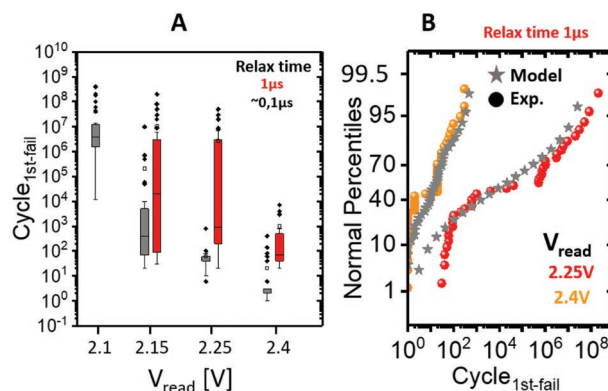


Figure 4. A) First measured cycle when the OTS switches under repeated applied pulses under threshold (called $\text{Cycle}_{1\text{st-fail}}$) as function as applied voltage. Two various relaxation times between reading cycles are considered. The higher V_{read} , the lower the median and maximal $\text{Cycle}_{1\text{st-fail}}$. For a given V_{read} , the larger the relaxation time between reading pulses and the higher appears $\text{Cycle}_{1\text{st-fail}}$. B) OTS $\text{Cycle}_{1\text{st-fail}}$ experimental and simulated distributions for $V_{\text{read}} = 2.25 \text{ V}$ and $V_{\text{read}} = 2.4 \text{ V}$. A fixed relaxation time of $1 \mu\text{s}$ between reading cycles is considered. High dispersion $\text{Cycle}_{1\text{st-fail}}$ experimental distributions are satisfactorily captured by adjusting the OTS global probability to switch on the Monte Carlo model.

Monte Carlo model for OTS switching operation is used to provide physical insights on $\text{Cycle}_{1\text{st-fail}}$ distributions. To capture the experimental data, the following dual behavior of the switching probability $\mathbb{P}_n$ along cycling is considered:

- Under mild reading conditions (either low reading voltage V_{read} or long rest time between subsequent reading cycles), the OTS is assumed to have sufficient time to relax between reading cycles so that the switching probability remains constant and the OTS state is unaltered. This is supported by the flat behavior of the OTS sub-threshold current, as provided in Figure 2D.
- Under aggressive reading conditions (either high reading voltage V_{read} or short rest time between subsequent reading cycles), gradual aging of the OTS is assumed and leads to a gradual increase of the switching probability. This is supported by the drift of the OTS sub-threshold current, as observed in Figure 2E.

Once the statistic of first switching is clarified, the next step consists in extending the study to subsequent OTS switching events through the device lifetime. Thus, OTS cycle-to-cycle switching probability has to be extracted. In this aim, the OTS current is measured ten times per decade when sub-threshold voltages are applied. The experience is repeated multiple times, and OTS switches are detected. Then, the OTS switching probability is calculated by quantifying the percentiles of switching events among all the experiments at a fixed reading cycle (Figure 5A)). The main goal here is to quantify the evolution of the OTS switching probability in the sub-threshold regime regarding the applied reading conditions (reading voltages and reading frequency). Figure 5B presents the evolution with the cycling of OTS switching probability at $V_{\text{read}} = 2.35\text{V}$, as a function of the relaxation time between reading operations. ≈ 110 tests are considered per condition. Again, shorter OTS relaxation time leads to a larger OTS switching probability.

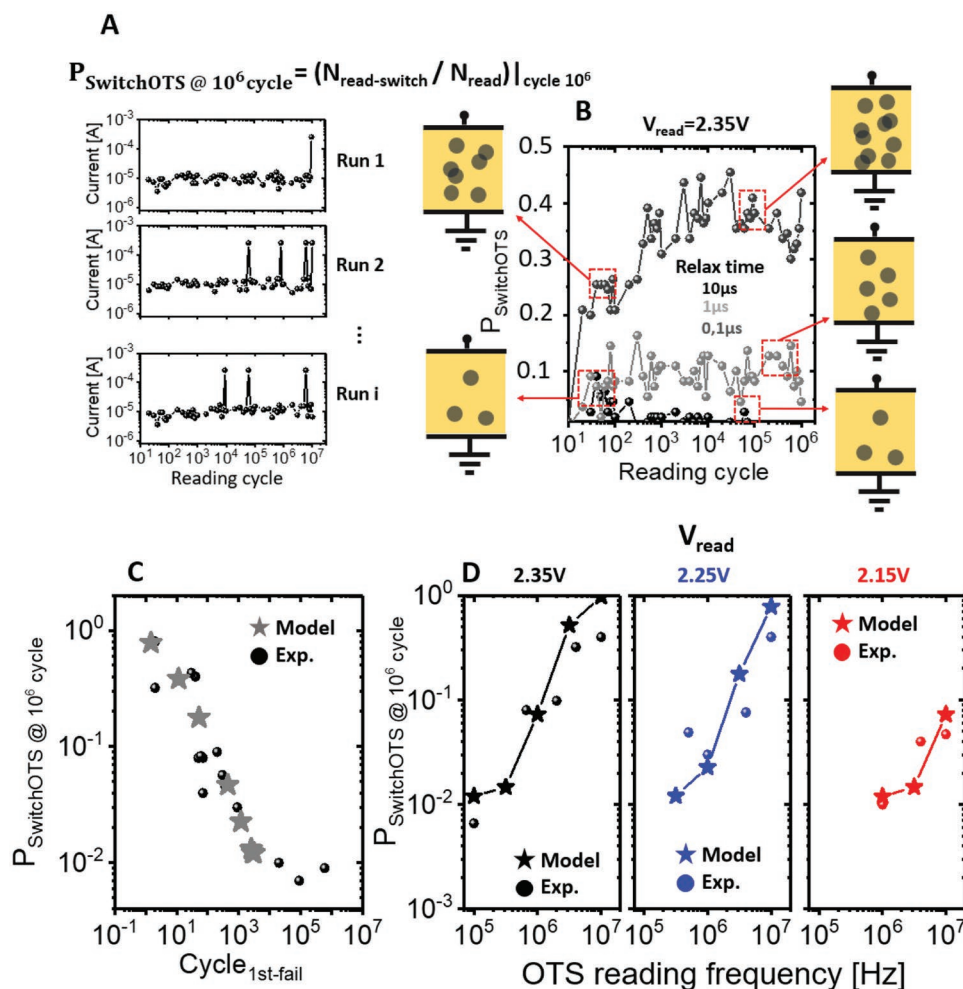


Figure 5. A) OTS switching probability description over cycling. Multiple pulses are applied in sub-threshold regime. By extracting the switching events at every cycle, the OTS switching probability is calculated. B) Evolution of the OTS switching probability with cycling at $V_{\text{read}} = 2.35\text{V}$, as a function of the relaxation time between reading operations. More frequent OTS switching is evidenced for shorter relaxation times between reading pulses. C) Experimental and simulated OTS switching probability after 10^6 reading cycles ($P_{\text{SwitchOTS}} @ 10^6 \text{ cycle}$) evolution with the median $\text{Cycle}_{1\text{st-fail}}$. $\text{Cycle}_{1\text{st-fail}}$ appears to be a direct image of the device $P_{\text{OTS}} @ 10^6 \text{ cycle}$. D) Experimental and simulated $P_{\text{SwitchOTS}} @ 10^6 \text{ cycle}$ evolution with the reading frequency. Three different reading voltages ($V_{\text{read}} = 2.15\text{V}$; 2.25V ; 2.35V) are considered. The higher the reading frequency, the larger the resulting switching probability on the devices for a certain V_{read} .

Moreover, the experimental and simulated OTS switching probabilities after 10^6 reading cycles ($P_{\text{SwitchOTS}} @ 10^6 \text{ cycle}$) as function as the median $\text{Cycle}_{1\text{st-fail}}$ is presented in Figure 5C. $\text{Cycle}_{1\text{st-fail}}$ appears to be a direct image of $P_{\text{SwitchOTS}} @ 10^6 \text{ cycle}$. In addition, the experimental and simulated $P_{\text{SwitchOTS}} @ 10^6 \text{ cycle}$ evolution with the reading frequency is illustrated in Figure 5D, together with its dependence with V_{read} . In agreement with precedent results, the higher the reading frequency, the larger the resulting OTS switching probability for a certain reading voltage. Thus, the lower V_{read} , the farther to the OTS intrinsic switching voltages, and so the minor the resulting switching probability for a given reading frequency. In this context, a similar drift in the probability P_n is required for the simulation to reproduce the experimental results. All in all, successful agreement between experimental and simulated data is evidenced.

3.4. 1S1R Reading Reliability Optimization

Based on the previous experimental and simulated OTS read operation reliability analysis, OTS switching statistics, under applied pulses in the sub-threshold regime, have been elucidated. In this section, we propose to estimate the overall 1S1R reading BER, which is directly impacted by OTS erratic switching events during the reading operation. Figure 6A illustrates the influence of the OTS reading voltages and reading frequency on the OTS switching probability after 10^6 reading cycles. The symbol size represents the resulting switching probability. This figure generalizes the trends described in the previous section and shows that a trade-off between reading frequency and voltage amplitude is required to prevent OTS erratic switching.

During 1S1R reading operation, the OTS behavior depends on the state of the OxRAM memory device: when the OxRAM is in LRS (resp. HRS), the OTS is expected to switch (resp. to remain in the OFF state) at V_{read} . On the one hand, 1S1R HRS BER is governed by the OTS probability to switch from OFF-to-ON state at V_{read} , which corresponds to the OTS switching probability studied in the previous sections. On the other hand, LRS BER represents the OTS probability to remain in the OFF state at V_{read} . Practically, LRS BER is obtained by flipping HRS BER ($P_{\text{OTS-ON}} = 1 - P_{\text{OTS-OFF}}$) and shifting it by the 1S1R window margin. Indeed, this shift corresponds to the read window margin (WM) and is equal to the voltage that drops on the memory in the HRS. The higher the memory resistance, the larger the threshold voltage shift ($\text{WM} = V_{\text{th-HRS}} - V_{\text{th-LRS}}$). It should be noticed that, for a first-order approximation, only the mean value for both OxRAM LRS and HRS resistive states is considered. The variability in the 1S1R memory window, which could be at the origin of a dissymmetrical behavior between LRS and HRS, is not taken into account here. In this context, high reading voltage increases the OTS switching probability and so degrades HRS BER while improving the LRS BER. Figure 6B,C) identify the optimal V_{read} allowing minimizing the overall 1S1R reading BER for both 10 and 4 MHz reading frequencies. A standard 1S1R read window margin of 700 mV is considered for the analysis.^[18] The lower the 1S1R reading frequency, the lower the optimal reading BER. Altogether, a 1S1R reading BER of $\approx 10^{-1}$ (resp. $\approx 4.5 \times 10^{-2}$) is estimated after 10^6

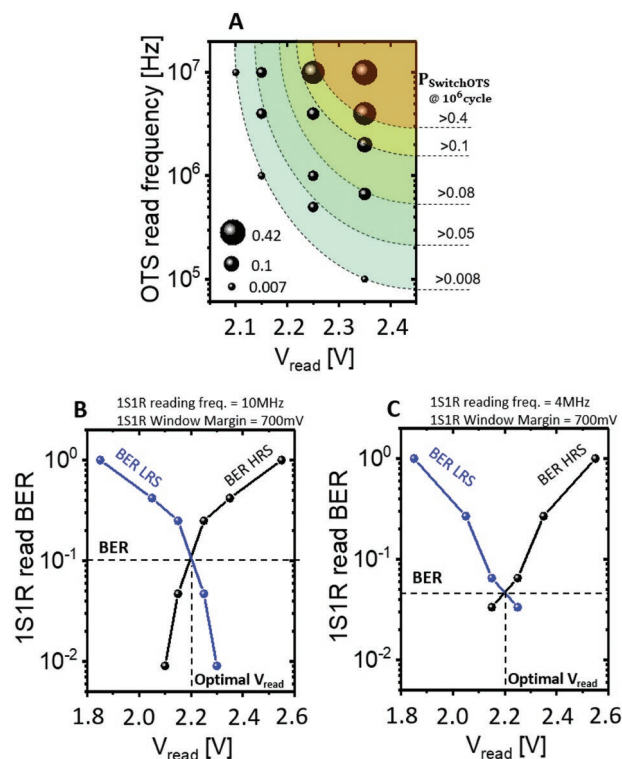


Figure 6. A) OTS reading voltages and reading frequency influence on the resulting OTS switching probability after 10^6 sub-threshold reading cycles. The BER is represented by the symbol size. B) Identification of the optimal reading voltages for 1S1R reading BER minimization for the devices of interest. The reading frequency is fixed at 10 MHz. A 1S1R read window margin of 700 mV is considered. C) Identification of the optimal reading voltages for 1S1R reading BER minimization for the devices of interest. The reading frequency is fixed at 4 MHz. A 1S1R read window margin of 700 mV is considered.

reading operations at 10 MHz (resp. 4 MHz) reading frequency. Improved performances could be expected for a larger window margin or lower reading frequency.

3.5. Binarized Spiking Neural Network Figures of Merit

To explore the benefit of a BSNN inference hardware implementation with 1S1R-based crossbar arrays, training simulations on a fully-connected BSNN with one hidden layer for an image classification task on the MNIST dataset are performed in this section. Figure 7A presents the considered BSNN topology, where the amount of neurons on the hidden layer is a variable of the study. Neurons of SNNs integrate spike inputs over time and thus are simulated with temporal dynamics. However, in our case, no significant network performance improvement is observed using several simulation timesteps (Figure 7B). Therefore, aiming to keep energy efficiency as high as possible, we chose to simulate our BSNN with a unique timestep. To perform the classification with our BSNN, the MNIST images are converted into spikes by using a pixel intensity threshold. If the pixel intensity is above (resp. below) the threshold, the input is set to +1 (a spike) (resp. 0). In this context, input sparsity is engineered by adapting the

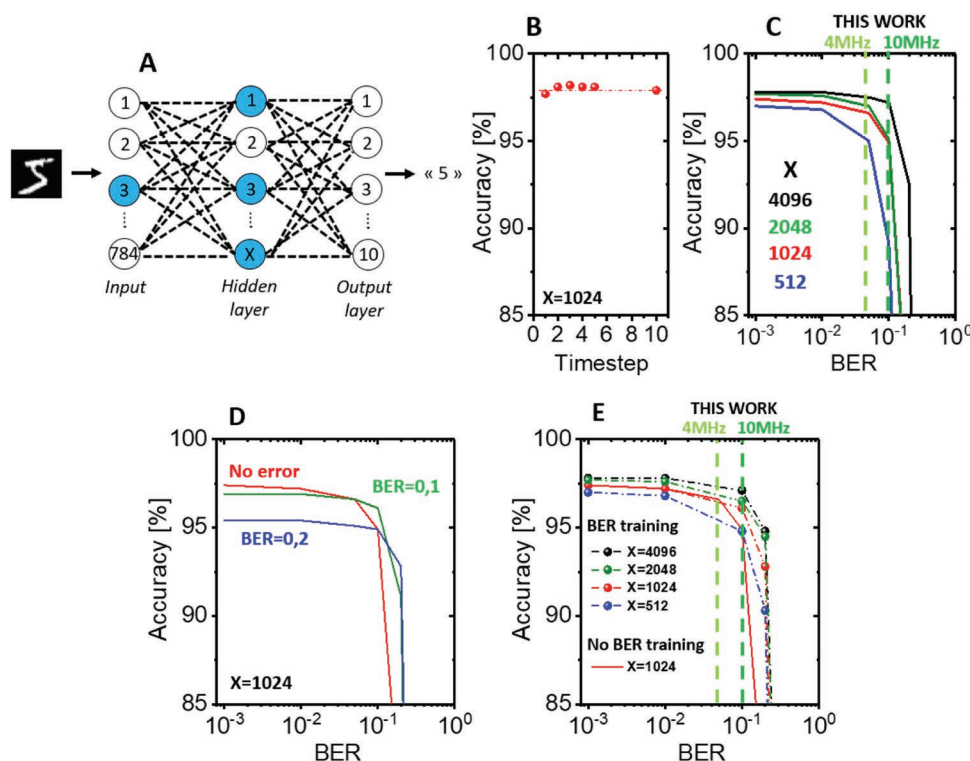


Figure 7. A) Binarized Spiking Neural Network (BSNN) considered in this work. Fully connected neural network with one hidden layer of X neurons for MNIST handwritten recognition. X values belong to [512, 1024, 2048, 4096]. B) BSNN network accuracy evolution with the number of inference timesteps, focusing on MNIST handwritten recognition task. No significant network performance improvement is observed using several simulation timesteps. C) Maximal attainable BSNN accuracy for a traditional training method as a function of the memory BER for the topologies of interest, considering one inference timestep. The 1S1R BERs are demonstrated to induce an important degradation in accuracy. D) An adapted BSNN training strategy is considered, where the memory BERs are included during the training process. 1024 neurons on the hidden layer are considered for this analysis. Again, one inference timestep is considered. The network accuracy strongly improves for the bit error of interest. However, the network becomes less performant for smaller BERs than the one used for training. E) Inference accuracy evolution with BER. Each point is obtained with a training strategy adapted to the considered BER. Again, one inference timestep is considered. No considerable accuracy degradation is observed for the 1S1R devices of interest (BER $\approx 10^{-1}$ at 10 MHz read frequency and BER $\approx 4.5 \times 10^{-2}$ at 4 MHz read frequency).

global pixel intensity threshold on the input images. The location of the pixel in the image is not considered for sparsity engineering. The BSNN is trained with standard deep learning techniques on the MNIST training dataset and evaluated on the validation dataset. The resulting maximal attainable BSNN accuracy evolution with memory BER is provided in Figure 7C. The number of neurons on the hidden layer is demonstrated to have an important influence on the network performance.^[18] The higher the number of neurons, the bigger the number of learnable parameters of the network (binarized weights), and so the higher the accuracy, up to some extent. Indeed, by increasing too much the number of neurons on the hidden layer, the network may be subject to the overfitting phenomenon. Unfortunately, the reading BER on the devices of interest is demonstrated to degrade the network performance. In this context, we propose to improve the BSNN tolerance to parasitic bit errors by adapting the network training strategy.^[26] To this aim, the BSNNs of interest are retrained, including bit errors in the weights during the training process, which allows anticipating the apparition of reading errors during the network inference. In particular, at each training iteration, the weights are randomly switched to their opposite with a probability equal to the target BER. Figure 7D presents the resulting maximal

attainable BSNN accuracy evolution with memory BER during testing for different target BER during training. The network performance improves for the BER introduced during network training. However, the network becomes less performant for smaller BERs than the one used for training. This trend is generalized in Figure 7E, which presents the optimized inference accuracy evolution with BER with a training strategy adapted to each case. The 1S1R experimental optimized BER characteristics in this work are now demonstrated to be perfectly tolerated by the networks and not to induce any important accuracy degradation.

Therefore, the ability to apply the optimal V_{read} on the devices in a Crossbar environment remains a key condition not to degrade the overall network accuracy. Based on 28 nm technological node resistive rules in metal lines, and while the optimal V_{read} is applied on the word-line extremity, Figure 8A provides the applied reading voltages on the 1S1R devices as a function of the device position on the word-line (column index n). The higher the number of devices per word-line on the Crossbar, the stronger the applied reading voltages degradation within the Crossbar due to IR voltage drop phenomenon. In this context, Figure 8B) provides the 1S1R reading BER evolution with the device position on the word-

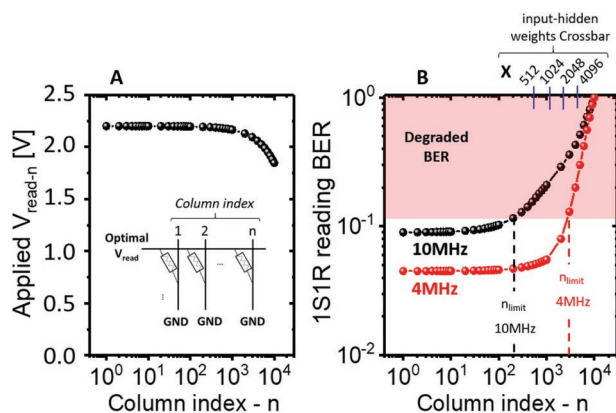


Figure 8. A) IR voltage drop on metal lines influence on the applied reading voltages on 1S1R devices as a function of their position on a Crossbar word-line (n value). 28 nm technological node resistive rules in metal lines are used for the analysis. The via resistivity is supposed to be negligible compared to metal line resistivity. The optimal V_{read} voltages identified in the previous section are applied in the word-line extremity. B) 1S1R reading BER evolution with the device position on the Crossbar word-line. Both 10 MHz and 4 MHz reading frequencies are considered for the analysis. The lower the 1S1R reading frequency, the lower the optimal reading BER and so the larger the readable Crossbar arrays due to better tolerance to IR drop phenomenon. Crossbar array sizes required to accommodate the input-hidden synaptic weights are indicated. The higher the amount of neurons on the hidden layer, the higher the required amount of columns on the Crossbar, and so the more the IR voltage drop phenomenon is important within the array. The Crossbar arrays required for hidden-output weights accommodation are not susceptible to lead to a significant IR voltage drop phenomenon, given a very low amount of columns on the array that corresponds to the amount of neurons on the output layer.

line. Both 10 MHz and 4 MHz reading frequencies are considered for the analysis. In order to preserve acceptable reading BER on the devices (Figure 7B), the maximal readable number of devices per word-line (n_{limit}) can be identified. Therefore, a trade-off exists between the 1S1R reading frequency and the maximal readable Crossbar size due to IR voltage drop

phenomenon on metal lines. In particular, the lower the 1S1R reading frequency, the lower the optimal reading BER, and so the larger the readable Crossbar arrays due to better tolerance to IR drop phenomenon. First, an adapted computation partitioning into smaller Crossbar arrays can be considered to prevent network performance degradation due to IR voltage drop on large Crossbar structures. Second, adapted network training strategies, including Crossbar IR drop phenomenon, allow preventing network accuracy degradation during inference.^[27–29] Third, including the IR voltage drop constraint when designing the circuit architecture, enhances its robustness to the phenomenon.^[29–31]

In this context, **Figure 9A** quantifies the trade-off between crossbar area and system accuracy. A 1S1R reading frequency of 4 MHz is considered for the analysis. The provided area is calculated by adding the respective area contributions from both input-hidden weights and hidden-output weights. First, assuming one driver transistor height per bit-line and word-line, the overall periphery area is calculated for 28 nm high-voltage CMOS for the network topologies of interest. The area contribution of the circuit dedicated to BSNN neuron implementation is not considered for the analysis. Second, assuming a CD_{min} metal width and space between metal lines, the 1S1R crossbar area is quantified for the various network topologies. Increasing linearly the number of neurons on the hidden layer leads to a quadratic increase in the number of synaptic weights and so of the overall crossbar area. Third, the equivalent 1T1R array area is estimated for the various network topologies.

In this context, the peripherals' footprint is demonstrated to be negligible compared to the actual crossbar area. In addition, one decade of area reduction is demonstrated for 1S1R-based Crossbar in comparison with 1T1R arrays. Moreover, network topologies with a large number of neurons on the hidden layer present degraded accuracy due to IR voltage drop phenomenon in the arrays. Altogether, the amount of neurons on the hidden layer of 1024 is observed to optimize the area-accuracy trade-off, promising high tolerance to BER while strongly reducing the overall system footprint. Given this topology, we propose to

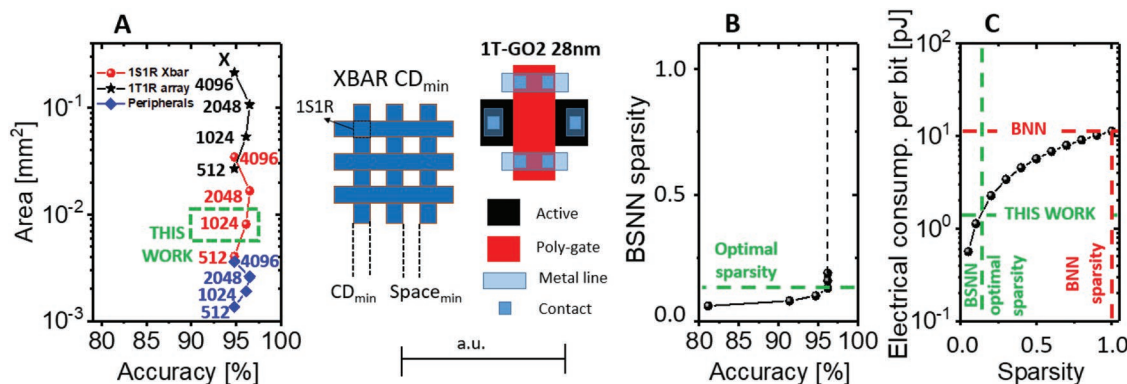


Figure 9. A) Evaluation of the trade-off between crossbar area and system accuracy for the 1S1R BER of interest, considering a 1S1R reading frequency of 4 MHz. Assuming one driver transistor height per bit-line and word-line, the overall periphery area is calculated for 28 nm high-voltage CMOS. The 1S1R crossbar area is provided (CD_{min} metal width and space between metal lines). The equivalent 1T1R area is estimated, demonstrating on order of magnitude improvement on system area for 1S1R crossbar architecture. An amount of neurons on the hidden layer of 1024 is observed to optimize the area-accuracy trade-off, showing high tolerance to BER while strongly reducing the overall system footprint. B) Optimal BSNN activations sparsity identification, preventing any network accuracy degradation on the MNIST recognition task. C) BSNN electrical consumption per read bit estimation. One order of magnitude improvement in comparison with standard BNN is estimated.

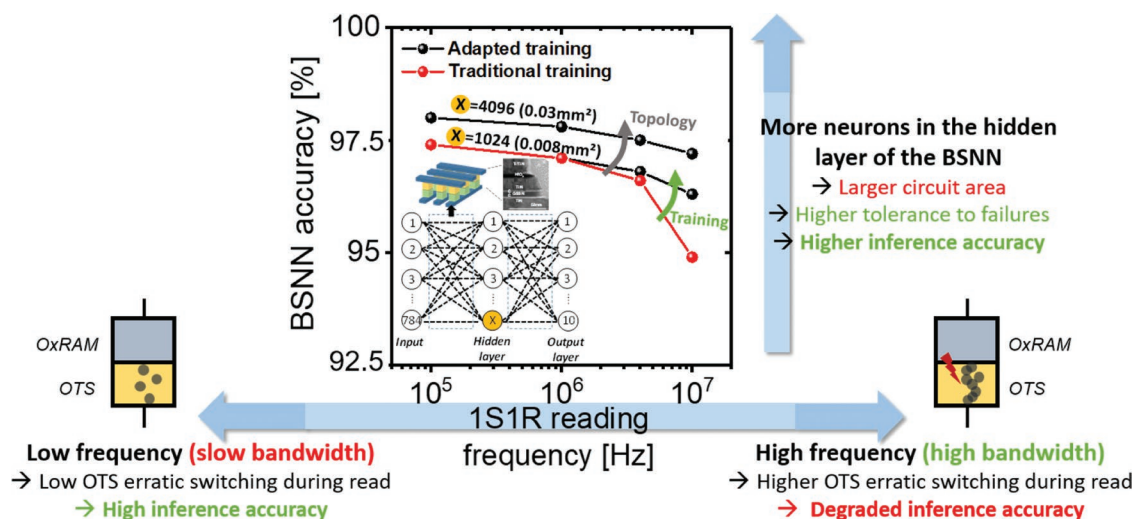


Figure 10. BSNN accuracy dependence with 1S1R inference frequency. Increasing too much the read frequency increases the OTS erratic switching probability, leading to degraded circuit inference accuracy. Increasing the number of neurons in the hidden layer on the BSNN (at the price of a larger circuit area) improves accuracy. Using adapted training protocol also increases tolerance to 1S1R failures and offers better overall BSNN accuracy.

optimize the overall BSNN electrical consumption by tuning the network activations sparsity. In particular, the activations sparsity is defined by the average number of spikes received by a synapse per image classification. Therefore, the sparsity corresponds here to the average number of weights that are read per inference. Thus, the lower the sparsity, the lower the number of weight readings per inference, and so the lower the overall electrical consumption. In our fully connected network, there are two contributions to the overall sparsity: the input-hidden synapses contribution and the hidden-output synapses contribution. Since the number of input-hidden weights is almost two orders of magnitude higher than the number of hidden-output weights in the given topology, the contribution of the hidden-output synapses to the overall sparsity is negligible. Particularly, the sparsity in the input-hidden synapses is the number of spikes fired by neurons in the input layer (previous layer in the network). In this context, the sparsity of our network is determined by the sparsity of the input layer, meaning the sparsity of the input image. To engineer the latter, the threshold that is used for the image binarization is modified. Figure 9B identifies the minimal BSNN sparsity, preventing the network performance degradation. The electrical consumption per reading bit evolution with the network sparsity is presented in Figure 9C). Approximately 1.4pJ electrical consumption per reading bit is estimated for this work. In comparison with standard Binarized Neural Networks (BNN), where all the weights are read at each inference, about one order of magnitude of energy consumption improvement is demonstrated for BSNN for an equivalent network accuracy on the task of interest.^[18]

Finally, general guidelines for BSNN figures of merit (area, operating frequency, and accuracy) optimization are provided in Figure 10. Assuming an adapted computation partitioning into smaller Crossbar arrays, IR voltage drop issues are not considered here. When the reading frequency is high, the erratic OTS switching probability (Section 3.3) increases due to the creation of a conductive path in the OTS, leading to degraded

1S1R BER (Section 3.4) what is detrimental for the neural network accuracy (Section 3.5). Adapted training and a larger circuit area (by increasing the number of neurons in the hidden layer) make the circuit more robust to failure.

4. Conclusion

1S1R capabilities for synaptic weight storage for Binarized Spiking Neural Network high-frequency inference hardware implementation are demonstrated. By crossing statistical experimental data on memory arrays with Monte Carlo simulations, the OTS switching dynamics are elucidated. Stochastic formation of local conductive dots in the selector leading to the formation of a conductive path allows catching the OTS switching probability distribution when repeated sub-threshold pulses are applied. Based on this analysis, 1S1R reading conditions are optimized for low reading BER during high-frequency inference. Focusing on the MNIST handwritten digit recognition task, general guidelines for system footprint and electrical consumption reduction and inference frequency and accuracy maximization network are provided. Overall, the 1S1R array of interest is demonstrated to perfectly sustain a 1 MHz inference with 97% accuracy, with an estimated circuit area lower than 0.01 mm². This opens the path to 1S1R exploitation for real-time image inference tasks.

Acknowledgements

This work was partially funded by the European project ANDANTE and StorAlge, the French IPCEI program, as well as MIAI @ Grenoble Alpes (ANR-19-P3IA-0003) program.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

1S1R, crossbar, ovonic threshold switch (OTS), resistive random-access memory (RRAM), spiking neural networks

Received: March 22, 2022

Revised: May 11, 2022

Published online:

- [1] A. Pedram, S. Richardson, M. Horowitz, S. Kvatinsky, S. Galal, *IEEE Design Test* **2017**, 34, 39.
- [2] V. Sze, presented at *NEURIPS*, Vancouver, Canada, Dec **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, M. F. Chang, *Nat. Electron.* **2019**, 2, 420.
- [4] M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, Q. Xia, J. P. Strachan, *Adv. Mater.* **2018**, 30, 1705912.
- [5] P. Yao, H. Wu, B. Gao, Q. Zhang, W. Zhang, J. J. Yang, H. Qian, *Nature* **2020**, 577, 641.
- [6] S. Yin, Y. Kim, X. Han, H. Barnaby, S. Yu, Y. Luo, W. He, X. Sun, J. J. Kim, J. S. Seo, *IEEE Micro* **2019**, 39, 54.
- [7] M. Le Gallo, A. Sebastian, R. Mathis, M. Manica, H. Giefels, T. Tuma, C. Bekas, A. Curioni, E. Eleftheriou, *Nat. Electron.* **2018**, 1, 246.
- [8] I. Boybat, M. Le Gallo, S. N. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian, E. Eleftheriou, *Nat. Commun.* **2018**, 9, 2514.
- [9] D. Ielmini, S. Ambrogio, *Nanotechnology* **2020**, 31, 092001.
- [10] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. di Nolfo, S. Sidler, M. Giordano, M. Bordini, N. C. P. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, G. W. Burr, *Nature* **2018**, 558, 60.
- [11] S. Yu, P. Y. Chen, Y. Cao, L. Xia, Y. Wang, H. Wu, presented at *IEDM*, Washington, USA, Dec **2015**, pp. 17.3.1–17.3.4.
- [12] M. Preziso, F. Merrik-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, D. B. Strukov, *Nature* **2015**, 521, 61.
- [13] P. Jain, U. Arslan, M. Sekhar, B. C. Lin, L. Wei, T. Sahu, J. Alzate-Vinasco, A. Vangapaty, M. Meterelliyo, N. Strutt, A. B. Chen, P. Hentges, P. A. Quintero, C. Connor, O. Golonzka, K. Fischer, F. Hamzaoglu, presented at *ISSCC*, San Francisco, CA, USA, Feb **2019**, pp. 212–214.
- [14] C. Chou, Z. Lin, C. Lai, C. Su, P. Tseng, W. Chen, W. Tsai, W. Chu, T. Ong, H. Chuang, Y. Chih, T. J. Chang, presented at *VLSI Circuits*, Honolulu, HI, USA, June **2020**, pp. 1–2.
- [15] C. Nail, G. Molas, P. Blaise, G. Piccolboni, B. Sklenard, C. Cagli, M. Bernard, A. Roule, M. Azzaz, E. Vianello, C. Carabasse, R. Berthier, D. Cooper, C. Pelisser, T. Magis, G. Ghibaudo, C. Vallée, D. Bedeau, O. Mosendz, B. de Salvo, L. Perniola, presented at *IEDM*, San Francisco, CA, USA, Dec **2016**, pp. 4.5.1–4.5.4.
- [16] G. Sassine, C. Nail, P. Blaise, B. Sklenard, M. Bernard, R. Gassilloud, A. Marty, M. Veillerot, C. Vallée, E. Nowak, G. Molas, *Advanced Electron Materials* **2018**, 5, 1800658.
- [17] A. Valentian, F. Rummens, E. Vianello, T. Mesquida, C. L. M. de Boissac, O. Bichler, C. Reita, presented at *IEDM*, San Francisco, Dec **2019**, pp. 14.3.1–14.3.4.
- [18] J. Minguet Lopez, T. Hirtzlin, M. Dampfhofer, L. Grenouillet, L. Reganaz, G. Navarro, C. Carabasse, E. Vianello, T. Magis, D. Deleruyelle, M. Bocquet, J. M. Portal, F. Andrieu, G. Molas, *Semi-cond. Science Tech.* **2022**, 37, 014001.
- [19] G. Molas, D. Alfaro Robayo, J. Minguet Lopez, L. Grenouillet, C. Carabasse, G. Navarro, C. Sabbione, M. Bernard, C. Cagli, N. Castellani, D. Deleruyelle, M. Bocquet, J. M. Portal, E. Nowak, presented at *IMW*, Dresden, Germany, May **2020**, pp. 1–4.
- [20] J. Minguet Lopez, D. Alfaro Robayo, L. Grenouillet, C. Carabasse, G. Navarro, R. Fournel, C. Sabbione, M. Bernard, O. Billoint, C. Cagli, L. Couture, D. Deleruyelle, M. Bocquet, J. M. Portal, E. Nowak, G. Molas, presented at *IMW*, Dresden, Germany, May **2020**, pp. 1–4.
- [21] D. Alfaro Robayo, G. Sassine, J. Minguet Lopez, L. Grenouillet, A. Verdy, G. Navarro, M. Bernard, E. Esmanhotto, C. Carabasse, D. Deleruyelle, E. Vianello, N. Castellani, L. Ciampolini, B. Giraud, C. Cagli, G. Ghibaudo, E. Nowak, G. Molas, presented at *IEDM*, San Francisco, CA, USA, Dec **2019**, pp. 35.3.1–35.3.4.
- [22] J. Minguet Lopez, L. Hudeley, L. Grenouillet, D. Alfaro Robayo, J. Sandrini, G. Navarro, M. Bernard, C. Carabasse, D. Deleruyelle, N. Castellani, M. Bocquet, J. M. Portal, E. Nowak, G. Molas, presented at *IRPS*, Monterey, CA, USA, March **2021**, pp. 1–6.
- [23] 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, F. Andrieu, presented at *IMW*, Dresden, Germany, May **2022**, Dresden Germany.
- [24] V. G. Karpov, Y. A. Kryukov, I. V. Karpov, M. Mitra, *Phys. Rev. B* **2008**, 78, 052201.
- [25] P. Noe, A. Verdy, F. D'acapo, J. B. Dory, M. Bernard, G. Navarro, J. B. Jager, J. Gaudin, J. Y. Raty, *Sci. Adv.* **2020**, 6, eaay2830.
- [26] T. Hirtzlin, M. Bocquet, J. O. Klein, E. Nowak, E. Vianello, J. M. Portal, D. Querlioz, presented at *AICAS*, Hsinchu, Taiwan, Mar **2019**, pp. 288–292.
- [27] M. E. Fouda, S. Lee, J. Lee, G. H. Kim, F. Kurdahi, A. M. Eltawi, *IEEE Access* **2020**, 8, 228392.
- [28] B. Liu, H. Li, Y. Chen, X. Li, T. Huang, Q. Wu, M. Barnell, presented at *ICCAD*, San Jose, CA, USA, Nov **2014**, pp. 63–70.
- [29] N. Lepri, M. Baldo, P. Mannocci, A. Glukhov, V. Milo, D. Ielmini, *IEEE Trans. Electron Devices* **2022**, 69, pp. 1575.
- [30] C. Huang, N. Xu, K. Qiu, Y. Zhu, D. Ma, L. Fang, *IEEE Journal of the Electron Devices Society* **2021**, 9, p. 645.
- [31] N. Lepri, A. Glukhov, D. Ielmini, presented at *IRPS*, Dallas, TX, USA, Mar **2022**, pp. 3C.2.1–3C.2.6.