Hybrid FeRAM/RRAM synapse circuit for on-chip inference and learning at the edge

Michele Martemucci, François Rummens, Tifenn Hirtzlin, Simon Martin, Olivier Guille, Tarcisius Januel, Catherine Carabasse, Olivier Billoint, Julie Laguerre, Jean Coignus, et al.

Hybrid FeRAM/RRAM Synapse Circuit for On-Chip Inference and Learning at the Edge

M. Martemucci[1, 2], F. Rummens[1], T. Hirtzlin[2], S. Martin[2], O. Guille[2], T. Januel[2], C. Carabasse[2], O. Billoint[2], J. Laguerre[2], J. Coignus[2], A. F. Vincent[3], D. Querlioz[4], L. Grenouillet[2], S. Saïghi[3], E. Vianello[2]
[1]CEA-List, Univ. Grenoble Alpes, France, email: michele.martemucci@cea.fr, elisa.vianello@cea.fr
[2]CEA-Leti, Univ. Grenoble Alpes, France, [3]IMS, Univ. Bordeaux, Bordeaux INP, CNRS, France,
[4]Univ. Paris-Saclay, CNRS, France

Abstract — This article presents an experimental demonstration of a hybrid FeRAM/RRAM synapse circuit implemented at the 130nm node. The circuit incorporates Metal-Ferroelectric-Metal (MFM) stacks, which exhibit FeRAM capacitor behavior when no filament formation occurs and function as RRAM after undergoing a forming operation. By leveraging the unique advantages of FeRAMs, such as ultra-low power consumption, in combination with the non-disruptive (infinite) reading capability of RRAMs, this circuit enables efficient on-chip inference and learning at the Edge.

I. Introduction

The rapid advancement of artificial intelligence has fueled the development of powerful algorithms that enable machines to learn from experience and interact autonomously with their environment [1]. However, implementing intelligent machines is hindered by hardware constraints like energy consumption and memory size. To overcome these challenges, researchers have explored in-memory computing architectures that utilize embedded memory devices for both processing and data storage. Recent progress has been made in on-chip inference [2][3], but achieving on-chip learning requires an ultra-low switching energy and exceptional endurance. Additionally, the vast amount of data to be processed by an inference chip implies the need of quasi infinitive reading endurance and non-destructive reading operations. Despite extensive efforts in the past decade, attaining such memory remains elusive.

In the pursuit of an embedded memory solution with ultra-low power consumption and outstanding endurance for on-chip learning, one transistor-one capacitor (1T-1C) structures based on ferroelectric materials (FeRAMs) have emerged as promising candidates [4]. However, the data-destructive reading operation of FeRAMs poses challenges for implementing inference applications. On the other hand, resistive switching devices (RRAM) offer non-disruptive reading operations, making them highly suitable for inference engines [5]. However, limited writing endurance and increased programming power limit RRAM's effectiveness for learning [6].

In this paper, we introduce a novel memory stack based on ferroelectric Si-doped hafnium oxide, enabling the co-integration of RRAM and FeRAM memories in the same Back End of Line (BEOL) of 130nm CMOS technology. This integration is achieved without the need for additional masks, simplifying the manufacturing process by combining both technologies into a single memory stack [7]. Leveraging this innovative technology, we designed, fabricated, and tested a new hybrid FeRAM/OxRAM synapse circuit. This circuit enables on-chip learning of Binarized Neural Networks (BNNs), where each weight is associated with a *hidden* value used only during the training phase and a binary value for the inference phase [8]. FeRAMs store the *hidden* weights, while RRAMs store the binary weights. We apply this approach to perform heartbeat arrhythmia detection, demonstrating compatibility with hardware constraints. This advancement paves the way for energy-efficient learning and inference at the edge, addressing the challenges posed by existing memory technologies.

II. Metal-Ferroelectric-Metal Stacks

TiN/Ti/Si:HfO_2/TiN Metal-Ferroelectric-Metal (MFM) structures were successfully integrated into the BEOL of a 130nm CMOS technology, positioned between M4 and M5 (Fig.1b,c) [9]. A Ti scavenging layer was deposited by PVD at the top interface, without any air break between the Ti and TiN top electrode (Fig.1d). The presence of the Ti layer serves a dual purpose: it enhances the ferroelectricity of the structure, resulting in a higher remanent polarization in the ferroelectric capacitor, and increases the concentration of oxygen vacancies (V_o) at the interface. This enables the creation of a purposeful V_o-filament within the ferroelectric layer, facilitating the operation of a resistive memory.

III. 1T-1C FeRAM Array Performances

We fabricated 16kbit 1T-1C FeRAM arrays with $0.36\,\mu\mathrm{m}^2$ MFM capacitors and sense amplifiers [10]. Our experimental results demonstrated a fully opened memory window of $120\,\mathrm{mV}$ (Fig.2). To evaluate the switching efficiency of the 1T-1C cells, we employed $3\,\mathrm{V}$-$2\,\mu\mathrm{s}$ reading pulses with a fixed reference voltage (V_{ref} =$0.56\,\mathrm{V}$). Fig.3 illustrates the trade-off between pulse width and amplitude for programming a 0 (SL pulse) or a 1 (BL pulse) in a cell that was initially set to the opposite state. Based on this investigation, we identified three programming conditions (A, B, and C) for evaluating power consumption and endurance. Notably, programming condition C achieved a switching energy lower than $200\,\mathrm{fJ/bit}$ (Fig.4). Higher programming voltages and longer programming times (condition A) enhance the Memory Window (MW) (Fig.5),

they come at the expense of increased Bit Error Rate (BER) after cycling (Fig.6). Therefore, for the remainder of this paper, condition C has been selected for programming the FeRAMs.

IV. 1T-1R Ferroelectric RRAM array

We integrated the same $0.36\,\mu m^2$ MFM stack into 16kbit 1T-1R arrays on the same wafer, along with 1T-1C FeRAMs. Following the forming step, the array underwent cycling under various programming conditions to minimize Bit Error Rate (BER) and enhance MW (Fig.7). The optimal RESET condition was achieved at V_{BL} =3.0 V and V_{WL} =3.1 V, while for binary operation, the optimal SET condition utilized a programming current of I_{cc} =67μA. For Multi-Level Cell (MLC) operation, a range of I_{cc} from approximately 26μA to 67μA can be employed (Fig.10). Furthermore, Fig.8 demonstrates the distributions of the Low Conductance State (LCS) and High Conductance State (HCS) after multiple cycling phases, providing the first array-level evidence supporting the use of a ferroelectric material as a resistive switching layer in RRAMs. To address the conductance instability relaxation issue observed at low programming currents [11], we verified the stability over time of the array programmed with 67μA (Fig.9). MLC programming was achieved using an iterative programming scheme [11], with the programming conditions depicted in Fig.11. The results of MLC programming for 4 and 8 levels per cell are shown in Fig.12.

V. Hybrid FeRAM/RRAM Synapse for on-chip learning and inference

We present a novel hybrid FeRAM/RRAM synaptic array that combines the strengths of both technologies, enabling efficient edge inference and learning. Our design incorporates the Metal-Ferroelectric-Metal stack technology discussed earlier, utilizing ultra-low power FeRAMs for training and non-disruptive RRAMs for inference. Our hybrid synapse consists of a group of n 1T-1C cells, where the bitline (BL) is connected to the wordline (WL) of a single 1T-1R cell. This connection forms a Transfer Line (TL), enabling direct and analog data transfer from n FeRAMs to a single RRAM without the need for intermediate circuits (see Fig.13). The parallel reading of all FeRAM elements within the hybrid synapse involves loading the parasitic capacitance of the TL to an analog voltage, representing the sum of '1' stored in the 1T-1C cells. This FeRAM-data-dependent voltage is then utilized to set the RRAM compliance current, effectively adjusting the device's conductance. We implemented an array of N=128 TLs, with each TL connecting N FeRAMs to M=16 RRAMs. Ideally, each TL is utilized for 16 synapses, with n=8 (see Fig.14). An optical microscopy photograph of the fabricated hybrid array and its peripheral circuitry is shown in Fig.15. Experimental results (Fig.16) demonstrate the relationship between the TL voltage and the number of activated FeRAM wordlines (WL_{Fe}) when the corresponding source lines (SL_{Fe}) are pulsed. The voltage difference of approximately 200mV between devices at 0 and 1 states aligns with the required programming range for multi-level operation in RRAMs (Fig.7). For a given n, the voltage level of the TL is determined by the number of programmed '1' states (N_{SW}) in the FeRAMs and the amplitude of the SL_{Fe} reading pulse. Additionaly, by fixing the RRAM BL voltage during the transfer operation, the gate-source voltage of the RRAM access transistor can be controlled, thereby matching the required RRAM programming voltages (Fig.18). Fig.19 illustrates the measured data transfer from the FeRAM data to the RRAM cells for n=8. The programmed RRAM conductance shows a direct proportionality to the number of 1 FeRAM states.

We implemented a 2-layer fully connected Binarized Neural Network (BNN) with on-chip learning for heart arrhythmia detection using the proposed FeRAM/OxRAM synapse [12]. The BNN training process involves assigning a real value to each synapse, which accumulates loss gradients using binary weights and a Stochastic Gradient Descent-like algorithm [7]. The real value, known as the hidden value (W_h), remains unused during inference, where only its sign is employed to obtain the binary weight ($W_b = \pm 1$) used during inference. To address this, we utilize FeRAM cells to store the W_h values as an integer number of 2 (n=4), 3 (n=8) or 4 bits (n=16) and RRAM to store the W_b values. During learning, W_b values are updated based on W_h values using our analog transfer technique (Fig.19). The RRAM stored data is binarized through a low-power binary reading operation to obtain binary weights. Fig. 20b presents a schematic representation of the training algorithm. During forward and backward propagation for each input sample, the binarized weights stored in the RRAM array are read to evaluate activations and gradients. These calculated gradients are then used to update the hidden weights stored in the FeRAMs accordingly. The FeRAM update is performed using a probabilistic programming scheme (Fig.21). FeRAMs are updated for each input sample, while the transfer operation into binary weights stored in the RRAMs is performed every $K = 100$ samples. Once training is completed, only the binarized weights are used for model evaluation and infer on previously unseen samples. The network achieves an inference accuracy of 88% with n=8 (Fig.22). These results validate the potential of the proposed circuit in enabling on-chip learning and inference.

VI. Conclusions

Our novel hybrid FeRAM/RRAM synapse circuit successfully combines the ultra-low power consumption of FeRAMs for on-chip learning with the non-disruptive reading capability of RRAMs for in-memory computing during inference. This technological breakthrough holds significant promise for neural network implementation and design. When coupled with advanced algorithms that address issues like catastrophic forgetting and enable continuous learning [13], this technology will enable the development of future *intelligent* devices and applications.

Acknowledgment. This work is supported by the ERC consolidator grant DIVERSE (101043854) and from a France 2030 government grant (ANR-22-PEEL-0010).

INTRODUCTION

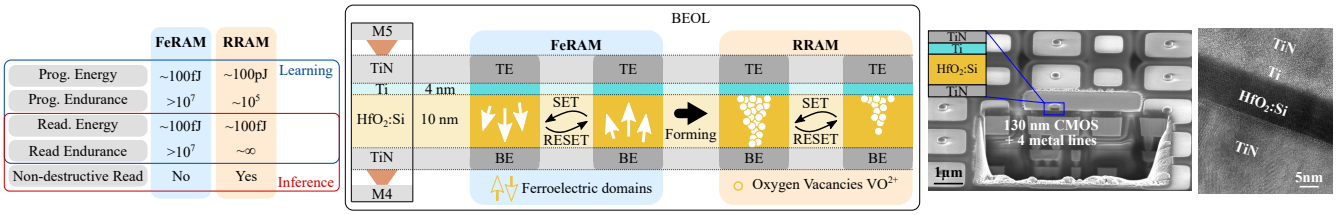

Fig. 1. (a) Performance comparison of state of the art HfO_2 based FeRAMs and RRAMs. (b) The Metal-Ferroelectric-Metal (TiN/HfO_2:Si/Ti/TiN) structure, integrated into the BEOL of 130nm CMOS, functions as a FeRAM when integrated into 1T1C bit-cells, and as a RRAM when integrated as a 1T1R structure. (c) Tilted SEM view with FIB cross section of $0.36\,\mu m^2$ TiN/HfO_2:Si/Ti/TiN stacks integrated above M4 of 130 nm CMOS, after memory stack etch. (d) HRTEM cross section illustrating 10nm Si:HfO2 film crystallization in the presence of Ti scavenging layer.

1T1C FeRAM ARRAY PERFORMANCES

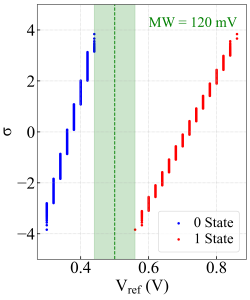

Fig. 2. Distribution of 0 and 1 states in a 16kbit 1T1C FeRAM array, measured at 3 V after wake-up cycling with 10^3 pulses (3 V-2 µs).

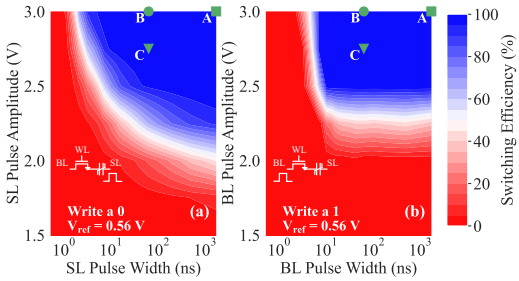

Fig. 3. Array-level switching efficiency measured for programming 0 (a) and 1 (b) states using different pulse widths and amplitudes. A fixed reference voltage (V_{ref} =0.56 V) is utilized for reading. Three programming conditions (A, B, and C) are highlighted in green.

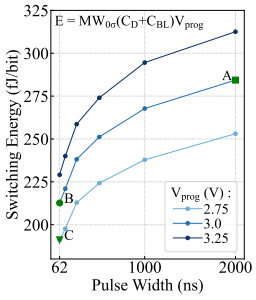

Fig. 4. Switching energy per bit. $MW_{0,\sigma}$ is the Memory Window at 0σ, C_D =9.5 fF the dielectric capacitance, and C_{BL} =188 fF.

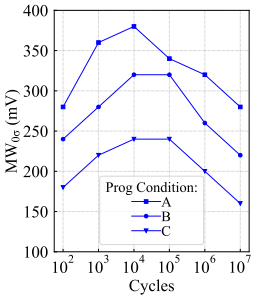

Fig. 5. Measured Memory Window at 0σ as a function of cycling. The same conditions are utilized for reading.

1T1R FERROELECTRIC RRAM ARRAY PERFORMANCES

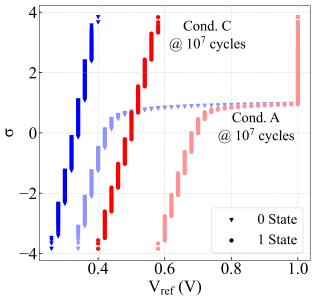

Fig. 6. Distributions of 0 and 1 states in 16kbit 1T1C FeRAM arrays, measured after 10^7 endurance cycles, for programming conditions A and C.

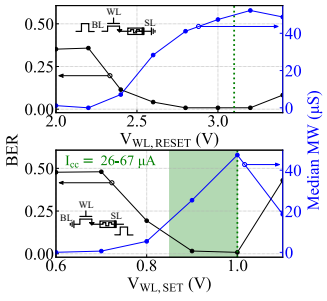

Fig. 7. BER and MW of a 1T1R ferroelectric RRAM array for different programming conditions. Green lines represent the conditions for binary programming, while the green area represents the conditions for MLC.

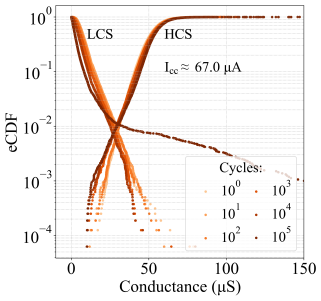

Fig. 8. Measured High Conductance State (HCS) and Low Conductance State (LCS) distributions of 16kbit RRAM array after several SET-RESET cycling phases.

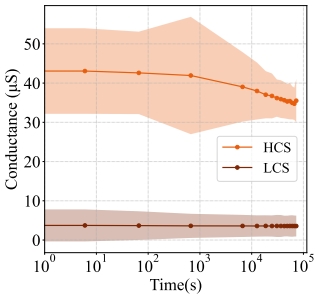

Fig. 9. Measured mean LCS and HCS evolution of RRAM array over time. Shaded area corresponds to standard deviation at 1σ.

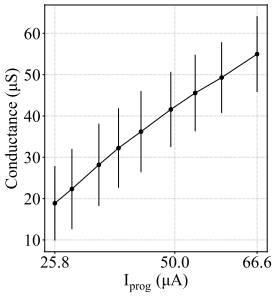

Fig. 10. Average RRAM conductance and standard deviation at 1σ for increasing programming currents.

Level		I_{prog} (µA)		G_{min} (µS)		G_{max} (µS)	
1	1	30.3	24.4	20	18	25	20
	2		30.8		24		26
2	3	39.5	37.3	30	30	35	32
	4		43.8		36		38
3	5	48.7	50.3	40	42	45	44
	6		56.8		48		50
4	7	57.9	63.2	50	54	55	56
	8		69.7		60		62

Fig. 11. 4 and 8 levels per cell allocation and corresponding programming conditions.

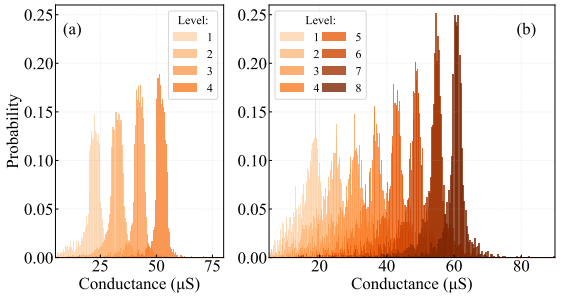

Fig. 12. Experimental conductance distributions for programming 4 (a) and 8 (b) bits per cell using the programming conditions shown in Fig.11.

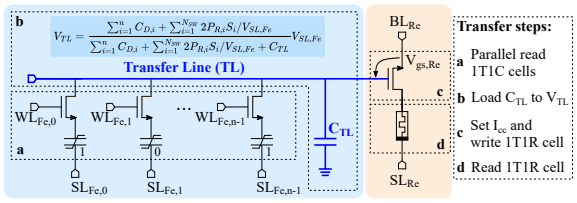

Fig. 13. Schematic view of the hybrid FeRAM/RRAM synapse circuit. During transfer, V_{TL} increases according to the expression shown in the figure, where $C_{D,i}$ represents the dielectric capacitance of a single FeRAM cell, $P_{R,i}$ the remanent polarization, S_i the surface, $V_{SL,Fe}$ the read voltage, and N_{SW} the number of capacitors at a ratio of 1 over n.

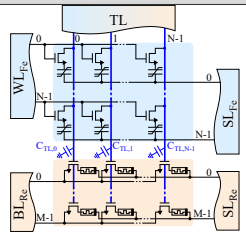

Fig. 14. Array-level organization of the Hybrid Synapses. A N by N FeRAM array is connected to a N by M RRAM array via TLs. Here, N=128 and M=16.

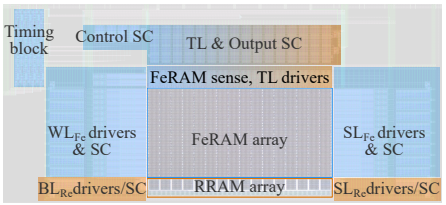

Fig. 15. Optical microscopy photograph of the die of the hybrid synapse array and peripheral circuitry, which includes line scan chains, drivers, sense amplifiers and output scan chain for the FeRAM array, control scan chain and timing block for the internal pulse generator.

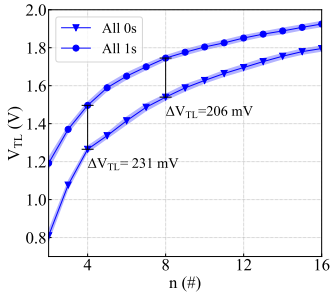

Fig. 16. Average TL voltage measured as a function of the number of activated FeRAM WLs (n). The measurements have been repeated over 126 synapse circuits.

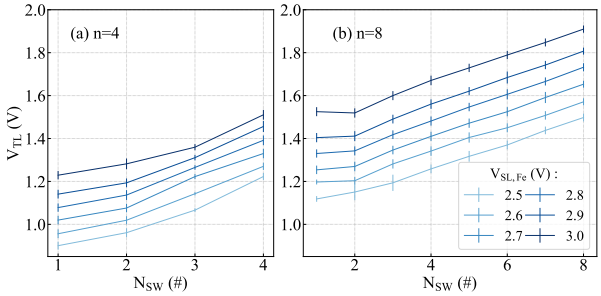

Fig. 17. Average TL voltage for n=4 (a) and n=8 (b) as a function of the number of 1 states programmed in the FeRAM, N_{SW}. The measurements have been repeated for different SL pulse amplitudes. Vertical lines correspond to the standard deviation at 1σ measured over 126 synaptic circuits.

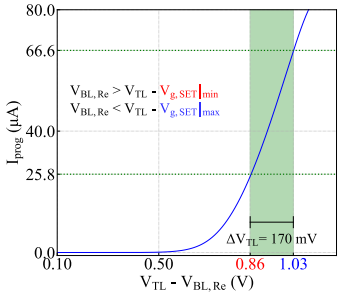

Fig. 18. Simulated RRAM programming current as a function of its gate-source voltage, defined as V_{TL}-$V_{BL,Re}$. $V_{BL,Re}$ can be adjusted to match the RRAM requirements and facilitate the data transfer from FeRAMs to RRAM.

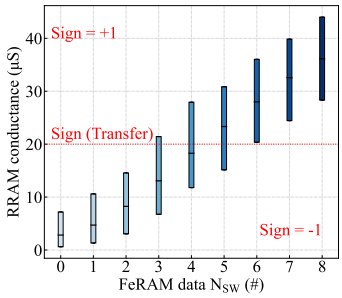

Fig. 19. Measured data transfer from 8 FeRAM to one RRAM cell. Median and inter-quartile ranges are shown as boxplots for 107 synaptic circuits. By thresholding the RRAM conductance, the transfer implements a sign fuction.

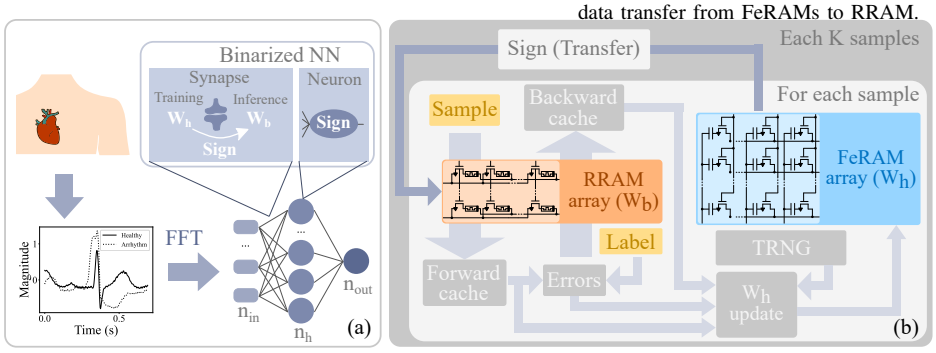

Fig. 20. (a) Example of an ECG signal converted into 64 features through Fast Fourier Transform (FFT), which are used as input for a two-layer fully connected neural network (64-512-1). (b) Detailed schematic of the training algorithm implementation using the hybrid synapse array concept shown in Fig.14: the FeRAM array is utilized to store the hidden values (W_h) used during training, while the RRAM array is used for the binary weights (W_b).

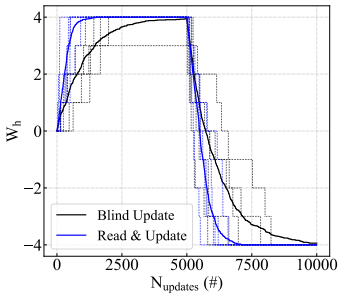

Fig. 21. Simulated synapse plasticity achieved on n=8 FeRAM cells using a probabilistic programming scheme. The blue lines represent the weight read before being updated, while the black lines represent the weight updated blindly.

Training strategy	Test Accuracy (%)		
	n=4	n=8	n=16
Online (Read & Update)	$87,48 \pm 1,89$	$89,03 \pm 1,08$	$89,30 \pm 0,89$
Online (Blind Update)	$85,23 \pm 3,82$	$88,04 \pm 1,42$	$89,15 \pm 1,22$

Fig. 22. Summary of the mean and standard deviation of the simulated test accuracies over 10 runs.

REFERENCES

[1] S. Bianchi et al., Nat. Commun., 14, 2023
[2] W. Wan et al., Nature, 608, 2022
[3] R. Khaddam-Aljameh et al., VLSI, 2021
[4] J. Okuno et al., IMW, 2021
[5] F. Moro et al., Nat. Commun., 13, 2022
[6] L. Grenouillet et al., IMW, 2021
[7] B. Max et al., J. Appl. Phys., 123(13), 2018
[8] M. Courbariaux et al., arXiv:1602.02830, 2016
[9] L. Grenouillet et al., VLSI, 2020
[10] T. Francois et al., IEDM, 2021
[11] E. Esmanhotto et al., Adv. Intell. Syst., 4, 2022
[12] G. Moody et al., IEEE Eng. Med. Biol. Mag., 20, 2001
[13] A. Laborieux et al., Nat. Commun., 12, 2021