IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 6, 2022, accepted April 27, 2022, date of publication May 3, 2022, date of current version May 12, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3172488

Time Domain Analog Neuromorphic Engine Based on High-Density Non-Volatile Memory in Single-Poly CMOS

TOMMASO RIZZO[1,2], SEBASTIANO STRANGIO[1], (Member, IEEE), AND GIUSEPPE IANNACCONE[1,2], (Fellow, IEEE)

[1] Department of Information Engineering, University of Pisa, 56122 Pisa, Italy
[2] Quantavis s.r.l., 56123 Pisa, Italy

Corresponding authors: Tommaso Rizzo (tommaso.rizzo@phd.unipi.it) and Giuseppe Iannaccone (giuseppe.iannaccone@unipi.it)

This work was supported in part by the EC Horizon 2020 Research and Innovation Programme under GA QUEFORMAL Grant 829035.

ABSTRACT Increasing the energy efficiency of deep learning systems is critical for improving the cognitive capability of edge devices, often battery operated, as well as for data centers, constrained by the total power envelope. Specialized architectures accelerated by analog vector-matrix multipliers (VMMs) can reduce by orders of magnitude the energy per operation, since the reduced precision of analog computation does not undermine the classification accuracy of the neural network. We show an analog vector-matrix multiplier fabricated with industry-standard 0.18 μm CMOS process, exploiting a single-transistor non-volatile analog memory cell and dedicated technology circuit co-design. The design is focused on implementation in neural networks performing offline training. The VMM performs the analog multiplication of a vector of inputs, encoded in the duration of time pulses, times a matrix of weights, encoded in the programmable currents of the memory cells. A 1.72 μm^2 memory cell is realized with a single transistor with floating gate, which can be operated as a two-terminal analog memristive device with more than 64 programmable current levels and high I_{high}/I_{low} ratio ($> 10^3$), tuned by the charge injected in the floating gate. A small-area charge amplifier is used to convert the multiply and accumulate operation result into a voltage. System-level projections based on our measurements and simulations provide a throughput of 333.17 GOps/s and an energy efficiency of 122.3 TOps/J, higher than comparable-precision VMMs reported in the literature, and an equivalent area per cell down to 2.15 μm^2, lower than any similar state-of-the-art solution. Of critical importance in view of translation to industry, our proposal uses in a new way an industry-standard low-cost single-poly CMOS process flow.

INDEX TERMS Analog computing, analog memory, analog neural networks, analog non-volatile memory, computing in-memory, neuromorphic computing.

I. INTRODUCTION

Classification accuracy and energy per task are the primary metrics for machine learning hardware, both at the level of edge devices, that typically must perform cognitive functions using the limited energy provided by batteries, and at the level of the data center, whose total computing capacity is constrained by the power envelope. This is the motivation of the vibrant research in dedicated architectures for Deep Neural Networks (DNNs), alternative to the Von Neumann model [1]–[3]. The most recurring computational task in a DNN is the multiplication of a vector of inputs of a network layer by a matrix of programmable weights, which explains the research attempts to develop specialized vector-matrix multipliers (VMMs) as dedicated hardware primitives with a high degree of parallelism, small latency, and energy-efficient operation [4]. In addition to parallel operation, the storage of the weights in local memory, or the usage of circuits in which computing elements and memory are intertwined as in "computing in-memory" solutions, are options to boost VMM performance and energy efficiency at the same time.

The associate editor coordinating the review of this manuscript and approving it for publication was Artur Antonyan.

In this scenario, dedicated analog VMMs [5]–[19] are promising for the possibility of exploiting technology-circuit co-design to provide an intrinsic and very high degree of parallelism and a quantum leap in energy efficiency. Analog computing has been historically phased away in the 80s for its sensitivity to noise, non-linearity, and process variations; however, deep neural networks have been proven to be capable of high classification accuracy also with limited arithmetic precision [20], [21], and to be resilient to noise and non-linearity [16], [22] in analog implementations.

Analog VMMs require analog non-volatile memories (NVMs) for weight storage, possibly close to the logic circuits to minimize the energy required for memory access [20], [23]. Emerging technologies such as Ferroelectric (Fe-) FETs [24] and Resistive Random Access Memories (RRAMs) [25], [26] still present technology challenges as dense analog NVMs: in particular, FeFETs, despite being able to achieve good linearity [24], [27], still have scalability issues; RRAMs, on the other side, are very promising from a scalability perspective, but often show an intrinsic bi-stable behavior [28], and limited analog multi-level programmability [29]. Single-poly floating gate (FG) cells represent an interesting alternative to industry standard double-poly FG flash memories [30]–[38], which require dedicated and expensive additional process steps. Different solutions for single-poly analog NVMs have been proposed for neuromorphic circuits also by some authors of the present paper [16], but typically require a large area due to the presence of a p-type MOSCAP between the control gate (surrounding n-well) and the FG (poly gate). Furthermore, they often need relatively high program voltage [39]–[41]. A few high-density proposals (area $\leq 3\ \mu\text{m}^2$) can still be found: [42], [43] present compact solutions, limited to single-bit logic; in [44], a Y-shaped two-transistor device cell is presented, used in [13] as analog memory for neuromorphic computing. A single-MOSFET device is proposed in [45], where electrons or holes can be injected in the spacer of non-overlapped channel regions, with limited analog capability.

In this paper, we show an analog VMM realized in an industry-standard, low-cost single-poly 0.18 μm CMOS process using a time-domain approach [9], [12], [17], namely where input signals are encoded in the duration of time pulses, achieving minimum area occupation and improved performance in terms of energy per task. The programmable weights are stored in a non-volatile memory matrix realized with novel single poly, single transistor, two-terminal FG cells, the *1T-FG cell*, using 3.3 V nMOS transistors with a relatively thick SiO_2 gate oxide ($\sim$ 7 μm) and a minimum area per analog cell of just 1.72 μm. We experimentally demonstrate the possibility to program and erase the cell with positive drain voltage and grounded source, with analog multi-level programmability corresponding to 6 equivalent bits and cyclability. Experiments and simulations are used to evaluate the performance metrics of the proposed time domain VMM: an energy efficiency of 122.3 TOps/J, a throughput of 333.17 GOps/s for a 500 × 500 VMM, with an overall silicon area of 0.537 mm^2 are obtained. System-level simulations of a 2-layer network processing the MNIST dataset [46], realized in a fully-analog approach using the proposed VMM with limited precision, result in a comparable inference accuracy with respect to floating-point precision operation ($\sim$ 1% at 6 bits).

II. 1T-FG CELL-BASED TIME-DOMAIN ANALOG VMM

A $M \times N$ time domain VMM (TD-VMM) architecture is depicted in Fig. 1(a), with details of the input vector of M elements, the $M \times N$ 1T-FG memory matrix where the computation takes place, and the N amplifiers converting the results into N output voltages. Each of the M inputs ($i = 1, \dots, M$) is encoded in the duration t_i of a voltage pulse of constant amplitude $V_{DS,ON}$ applied to the i-th row of the array. Each input pulse activates all cells in the row for its time duration, which ranges from 0 to a maximum pulse width T. The programmable weights $w_{i,j}$ are encoded in the currents $I_{i,j}$ of 1T-FG cells, which can be programmed by means of charge injection in the FG of each cell. The sources of all cells in the same column are connected to the same bitline (BL) node at virtual ground. The net charge $Q_{i,j}$ injected by cell (i, j) into the bitline is

$$Q_{i,j} = I_{i,j} \cdot t_i. \tag{1}$$

The total charge injected in the bitline j is converted into a voltage by means of a so-called charge amplifier, realized using a purposedly designed operational transconductance amplifier (OTA) with a feedback capacitor. The output voltage is therefore:

$$V_{out,j} = \frac{1}{C}\sum_{i}^{M} Q_{i,j} = \frac{1}{C}\sum_{i}^{M} I_{i,j} \cdot t_i. \tag{2}$$

As the VMM implementation heavily impacts the overall performance of a neural network, in terms of precision, area occupation and power consumption, several challenges arise for an effective design. To enable a comparison with other analog and digital options, we consider the following figures of merit (FOMs):

- Throughput (THR), i.e. number of elementary operations performed per unit time (Ops/s);
- Energy Efficiency (EE), defined as the number of elementary operations per unit energy (Ops/J);
- Effective Number of Bits (ENOB), indicating the equivalent number of bits of an analog-to-digital conversion only affected by quantization noise, considering the signal-to-noise ratio (SNR) and non-linearity (expressed by total harmonic distortion (THD)) of the real implementation.

An elementary operation (''Op'') is a scalar sum or multiplication: the multiplication of a vector of M elements times a ($M \times N$) matrix involves ($2M - 1$) × N elementary operations [16]. More details on the FOMs are in Appendix A.

A. SINGLE-TRANSISTOR ANALOG MEMORY (1T-FG CELL)

The layout of a 4×8 subset of the 1T-FG memory cell array is shown in Fig. 1(b). A single cell consists of a minimum-size 3.3 V n-type MOSFET with 7 nm-thick silicon oxide, where the gate terminal is floating. Thus, cell addressing in a two dimensional array is done by selecting drain (D) and source (S), which are respectively connected to the wordline (WL) and to the bitline (BL). The cell pitch of 1.12 μm and 1.54 μm results in a cell area of 1.72 μm^2.

Due to the lack of cell selectors in the 1T-FG array, the possible unintentional selection of a cell must be accurately addressed. To limit half selection issues, different configurations should be adopted when performing a program/erase (i.e. write) or a read operation. When writing, it is important to avoid that the non-selected cells are disturbed: Fig. 1(c) shows the implemented voltage scheme for such operation. When a write (program/erase) voltage V_{wri} is applied between D and S of the target cell, the non-selected lines are kept in a high-impedance (Hi-Z) state. In this way, a minimum of three cells in series will be exposed to the same V_{wri}, which is not enough to produce the unintentional writing of half-selected cells. A read operation can be performed by forcing a voltage V_{source} between D and S, and reading the resulting current I_{sense}, as illustrated in Fig. 1(d). Non-selected BLs are put in Hi-Z, whereas non-selected WLs are forced to 0 V.

The proposed architecture can be subject to sneak-path leakage currents due to half-selected cells when either WL or BL (or both) are in Hi-Z, especially for extremely large memory array dimensions. This is not a practical issue for reduced dimension memory arrays exploited in our time-domain VMM application, and we believe the resulting small leakage can be tolerated considering that the cells are programmed only once right after the training phase, while they are accessed only in read-mode during normal inference operations. In addition, in the proposed VMM, the whole array is read in parallel, with all the BL voltages forced to 0 V by the OTA virtual grounds, and the WLs being either at V_{DD} or 0 V depending on the corresponding input pulse: in each cell the current can only flow from the WL (D) toward the BL (S), and never in the opposite direction. For this reason, unintentional cell activation cannot take place during the inference operation.

The program and erase schemes for a single cell are illustrated in Fig. 1(e) and (f), respectively. By relying on oxide tunneling and different gate injection phenomena [32], it is possible to program or erase each cell to target a specific current when the wordline is activated with a fine granularity, using a series of voltage pulses. It is worth to clarify that, for both program and erase operations, no negative voltage is needed, and the intensity and sign of the injected charge is tuned by modulating the amplitude of the V_{DS}. Appendix B provides more detail of the underlying carrier injection phenomena involved in cell programming.

It is worth to mention that device-to-device variations (surely one of the major problems of analog design) can be almost entirely mitigated with program and erase operations: indeed, memory programming stops when the desired level of current in a 1T-FG cell is reached, independent of the initial threshold voltage of the transistor.

B. EXPERIMENTS ON ANALOG PROGRAM AND ERASE

We have performed the electrical characterization of single 1T-FG cells to demonstrate program and erase operations and multi-level analog programmability. Writing a cell results in a shift of its threshold voltage: Fig. 2(a) shows several I-V characteristics of the same cell at different programming stages: after the cell is first erased with a 6.2 V pulse, the leftmost transfer characteristic is measured (lowest V_{th}). Then, single 80 ms program pulses are applied to the cell, and a new I-V curve is captured each time, so that the resulting variation of the threshold voltage can be verified. Different programming schemes can be used to obtain the curves: in Fig. 2(b) and (c), we investigate two possible approaches, respectively Constant-Pulse Programming (CPP) and Incremental-Step Pulse Programming (ISPP) [47]. Each plot shows the variation of the read voltage V_{read}, defined as the V_{DS} when a current $I_D = 1$ nA is forced to flow through the cell, after a CPP or ISPP programming pulse. Using a CPP approach, programming occurs with 80 ms pulses at a constant value of 4.8 V, whereas the cell is erased with a single 80 ms pulse at 6.2 V. With ISPP, every state can be reached using pulses with progressively increasing amplitude. The programming voltage varies from 4.5 V to 6 V in steps of 5 mV, enabling a fine granularity. A similar result could be likely achieved by increasing the voltage amplitude and decreasing the pulse time. By comparing the evolution of the read voltage, it is clear that ISPP presents better linearity with respect to CPP, allowing us to reach all the desired states within the selected V_{read} range and with a better predictable number of steps. Additionally, linearity reduces the stress on the device gate oxide caused by the high electric field, as explained in Appendix C, where an additional level of detail regarding CPP and ISPP comparison is provided.

The benefits of ISPP are even better explored in Fig. 2(d), where several ISPP programming cycles were performed on the same cell, setting a stop condition to analyze only the rising part. In particular, increasing step pulses were used to bring the cell from a reset state to an upper read voltage threshold of $V_{high} = 1.4$ V, whereas a 6.2 V pulse was used to quickly erase the cell and get it ready for another cycle. The V_{read} measured points show high linearity and analog capability, spanning more than 64 intermediate levels within the explored operating range, corresponding to a weight ENOB greater than 6 bits. A preliminary endurance test of 100 cycles, corresponding to more than 10 thousand pulses, shows no evident variation in programmability. It is important to point out that the value of 80 ms for a single program/erase pulse is a limit imposed by the test setup: fast pulses can be practically used in the integrated memory, with further benefits in terms of both ENOB and cyclability. For the same

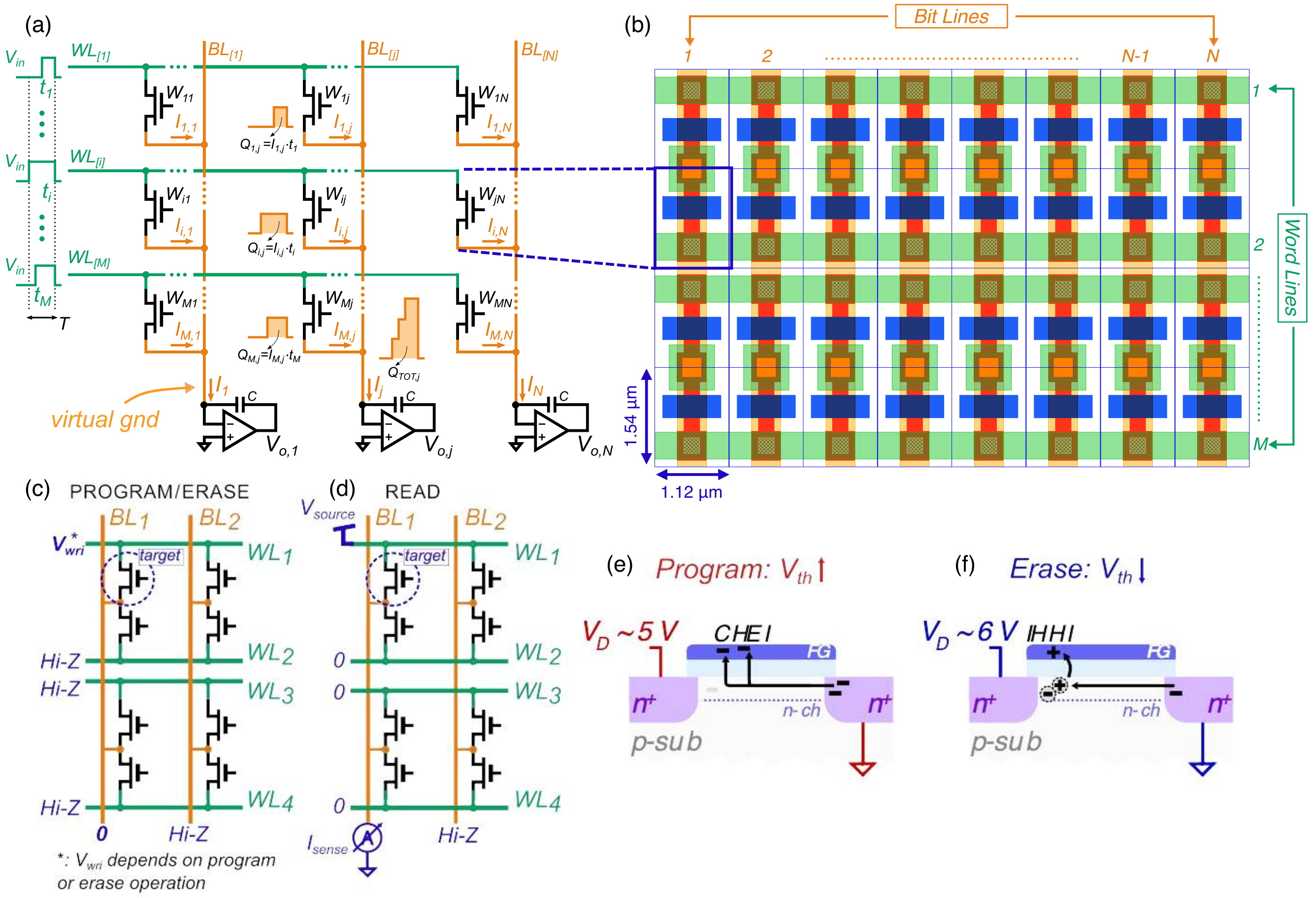

FIGURE 1. Analog time-domain VMM and 1T-FG cell. (a) Architecture of the proposed $M \times N$ analog time-domain VMM: the input signals $t_{in} \in [0, T]$ are highlighted in green, the matrix cells, realized with single transistor floating gate devices (1T- FG cells), and the integrator blocks (operational amplifiers with feedback capacitors) performing the charge-to-voltage conversion are depicted in black, with BL connections in orange; (b) layout of a 4 × 8 1T-FG cell array composed of several pairs of mirrored cells, sharing the BL contact; electrical scheme for (c) write and (d) read operations with detail of the applied voltages for selected and non-selected WLs and BLs; sketch of the injection phenomena involved in (e) program and (f) erase operations.

reasons, while an increment of the value of a cell can be finely realized with 80 ms programming pulses, a decrementing operation is done by erasing the cell and re-programming it to reach the target value. Shorter time pulses can enable gradual erasing and possibly successive approximation programming schemes.

Finally, retention test measurements (Fig. 2(e)) show no degradation after a 12-hour time frame, considering 16 programmed levels. Results are therefore promising, but they need to be improved for industrial level applications, especially if data must be retained for much longer time with similar granularity achieved right after the cell programming. Higher retention can be obtained using devices with thicker gate oxide, such as for example 5 V MOSFETs, which have an oxide thickness close to 10 nm.

C. DESIGN OF THE CHARGE AMPLIFIER

The charge amplifier, or Miller integrator, integrates the current gathered at its input node coming from its column matrix cells. Its design is crucial since it heavily impacts the performance of the VMM, affecting throughput, energy efficiency, area occupation and accuracy of the operation. In Appendix D we describe the methodology used to design the block, composed by a current-mirror OTA and a feedback capacitor. The transistor-level schematic of the OTA is shown in Fig. 3(a), whereas its layout is shown in Fig. 3(b), resulting in an area of $15.91 \times 13.26 = 211\ \mu\text{m}^2$. Voltage supply is set at $V_{DD} = 1.2\,\text{V}$ and $V_{SS} = -1.2\,\text{V}$, whereas the voltage at the non inverting node is $V_{ref} = 0\,\text{V}$. With a bias current I_{bias} of $1.25\ \mu\text{A}$, all transistors are in saturation, and the resulting DC gain, static power consumption and GBW are of 56.9 dB, $3.79\ \mu\text{W}$ and 9.77 MHz, respectively. By varying the bias current value, with a lower and an upper limit set by the saturation of the transistors, it is possible to increase the GBW of the device, at the cost of a higher power consumption, as shown by Fig. 3(c).

Fig. 3(d) shows the circuit representation of the integrator block, composed by the current-mirror OTA, the feedback capacitor $\text{C} = 0.6\,\text{pF}$, and the reset transistor which is activated right before starting a new integration cycle.

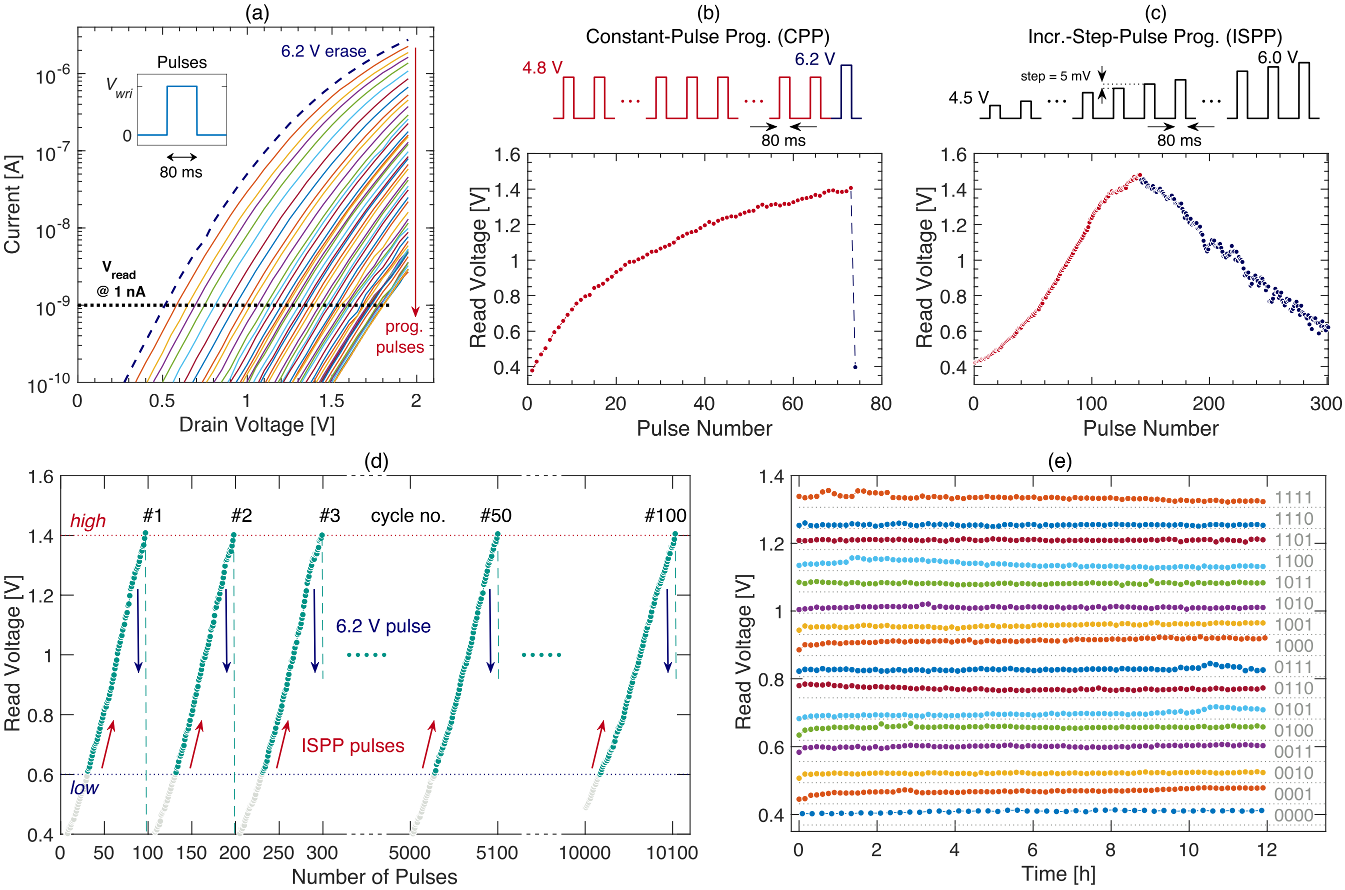

FIGURE 2. Measurements on a single 1T-FG cell. (a) I–V characteristics of the cell: starting from a condition where the cell has been erased with a 6.2 V pulse, the various curves are obtained performing a V_D sweep after the cell has been programmed using single 80 ms pulses with CPP or ISPP techniques. A more detailed comparison between the two techniques, displaying the read voltage (measured using a 1 nA current source) after every programming pulse, is shown in: (b) CPP technique, where 4.8 V pulses are used to program the cell, whereas a single 6.2 V pulse is sufficient for the erase operation; (c) ISPP technique, where programming is performed using incremental voltage pulses, from 4.5 V with incremental steps of 5 mV. (d) Cell cycling of ISPP programming operations – note that even after 100 full program-and-erase cycles, all analog levels between the chosen interval can still be reached. In (e), the retention characteristic of 16 different analog levels is evaluated in a 12 hour time frame.

The capacitor is realized using a MIM (Metal-Insulator-Metal) structure, featuring high linearity at the cost of large area with respect to MOS capacitors. However, given that it makes only use of the top metal layers, it can be placed on top of the other elements, without adding area overhead.

To separately evaluate the constraints on precision due to the integrator block, we use as input to the integrator an ideal current source (Fig. 3(d)), in place of the BL gathering the currents from the column of 1T-FG cells. The pulsed current source provides an on current I_{in} for a duration t_{in}, following the timing diagram of Fig. 3(e), where also the reset signal is shown.

With the objective of evaluating the precision of the network (using the ENOB definition given in Appendix A), we can define the following simulation setup. We consider a sinusoidal input signal, applying $K + 1$ ($K = 128$) ideal current pulses with a constant amplitude and a duration calculated as follows:

$$t_{in}(k) = \frac{T}{2}\left(1 + \sin 2\pi \frac{k}{K}\right), \tag{3}$$

with $k = 0, \ldots, K \Rightarrow t_{in}(k) = 0, \ldots, T$. The amplitude of the current pulses is kept at a constant level $I_{in} = \Delta V_{MAX} C/T$, which represents the maximum weight value. Therefore, the corresponding output voltage V_{out} is expected to vary with a practically sinusoidal rule (from sample to sample) between 0 and ΔV_{MAX}, the latter chosen equal to 1 V. For each one of the K simulations, the output voltage is sampled at the time $(T + T/4)$, and thus the V_{out}-k curve represents the sinusoidal output over which the ENOB is evaluated. Recalling Fig. 3(c), by changing the I_{bias} of the OTA, it is possible to improve its GBW, at the cost of an increased static power dissipation. Similarly, since the GBW sets the minimum period T of the VMM, it is possible to evaluate the minimum T that, for a given power consumption, guarantees a certain ENOB, shown in Fig. 3(f): we can observe that the P_{diss}-T curves have a nearly hyperbolic behavior, meaning that it is possible to move along a fixed ENOB curve while keeping constant the $P_{diss} \times T$ product, and thus the energy required for the integration. These considerations set theoretical limits for the THR and EE achievable with a VMM

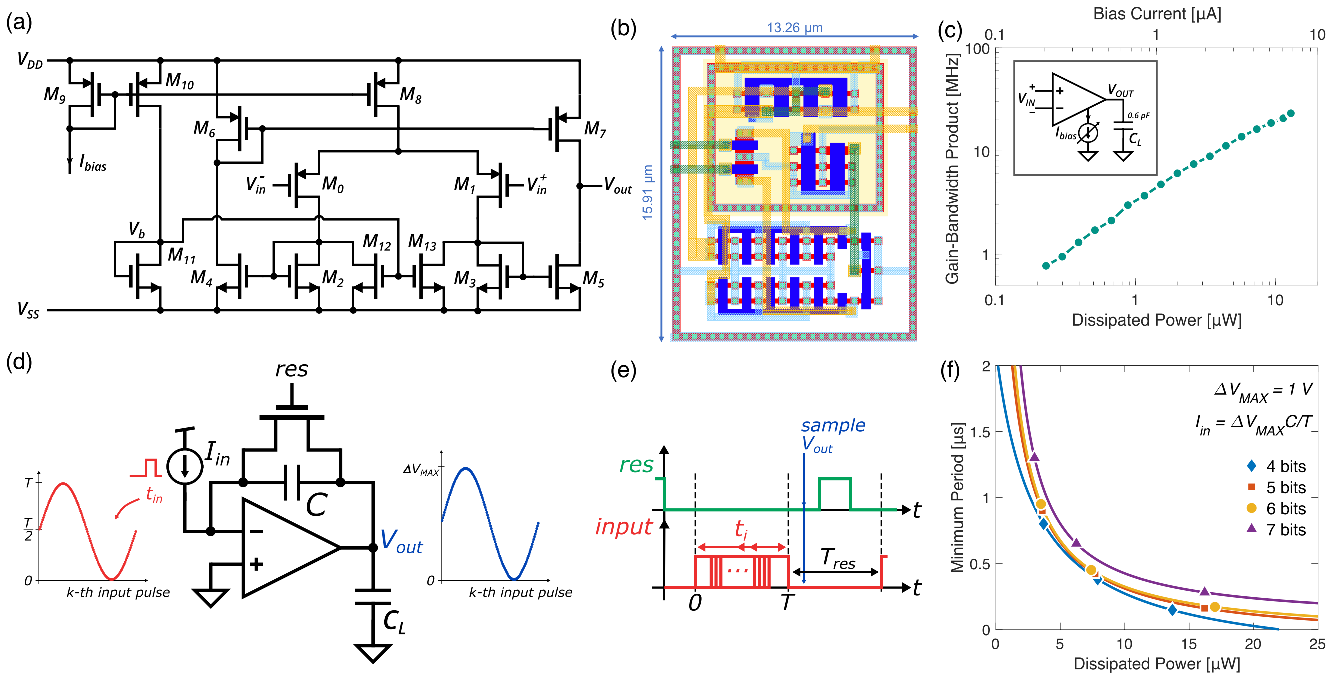

FIGURE 3. Integrator block. (a) Transistor level design and (b) layout of the gain-enhanced current-mirror OTA; (c) OTA static power versus GBW, obtained for different bias current; (d) simulation setup of the integrator block with a current input provided by an ideal current source, with (e) showing a timing diagram of the input and reset signals, and (f) showing the simulated results of the minimum period T needed to reach a target precision (in bits), for a given power (set by I_{bias}).

based on the proposed OTA. From Fig. 3(f) one can see that a precision of 6 bits – which is an acceptable value for most DNNs [16] – is reached with $T = 0.45\ \mu$s, obtained for $I_{bias} = 1.25\ \mu$A, i.e. for a dissipated power of 7.3 μW.

When the actual bitline is posed at the input of the OTA instead of the ideal current source, the non-ideal virtual ground due to the finite DC gain of the OTA has a significant impact on the performance of the VMM: the voltage at the inverting node connected to the BL column of the memory array is not exactly constant, and the resulting current pulse through the cells during the integration period will be distorted with respect to the ideal rectangular waveform.

Starting from the theoretical limit corresponding to $T = 0.45\ \mu$s and OTA $I_{bias} = 1.25\ \mu$A , the whole VMM is simulated with the same input applied to the 1T-FG cells of the same columns and the resulting ENOB accuracy is extracted. For 1T-FG cells programming, we have emulated the charge injection in the FG by means of an ideal pulse-current source, which charges the FG at the beginning of each transient simulation to emulate the programmed weight: in an actual implementation, a CPP approach will be used with two different voltage values for program ($V_{prg} = 4.8$ V) and erase ($V_{ers} = 6.2$ V) operation. During VMM operation, the wordline voltage pulse will be set at $V_{DD} = 1.2$ V to properly drive the corresponding cells. Cells are programmed at the maximum $I_{i,j}$ weight so that the maximum current (full scale) is provided as input to the integrator:

$$I_{column} = \frac{\Delta V_{MAX} C}{T} \tag{4}$$

where $\Delta V_{MAX} = 1$ V. In this case, if $t_{in} = T$, the output voltage of the integrator corresponds exactly to ΔV_{MAX}. When a series of t_{in} inputs sampled from a sine shape spanning the full scale $(0 \ldots T)$ is provided to the VMM, the sinusoidal sequence of resulting outputs is affected by non-linearity, with a THD $= -26$ dB (noise is a second order effect, with SNR $=$ 38 dB), therefore the ENOB is limited to 4 bits. This degradation is due to the current pulse which causes a deviation of the virtual ground voltage from 0 V, because of the finite GBW of the OTA. The 1T-FG cell source voltage variation leads to a distortion of the $I_{i,j}$ waveform. This effect can be mitigated by lowering the current. Looking at (4), a lower I_{column} can be obtained by either lowering the output voltage range ΔV_{MAX} or increasing the integration period T. Note that, since I_{column} is the current corresponding to the maximum weight, all other current values corresponding to different weights will be lower than I_{column}, and therefore will not introduce any further undesired effects.

III. VMM PERFORMANCE ANALYSIS

In order to optimize the real VMM operated with the 1T-FG cells, we have extracted the ENOB as a function of three parameters: the maximum output voltage ΔV_{MAX}, the maximum period T and the 1T-FG column array size M. As a first analysis, we have considered a single cell VMM ($M = 1$) in Fig. 4(a): for a long period (e.g. $T = 2\ \mu$s), an ENOB of 6 bits can be reached in a wide range of ΔV_{MAX}, between 0.48 V and 1.1 V. In Fig. 4(b), the same analysis is repeated for a fixed T of 2 μs and for an increasing M up to 1000.

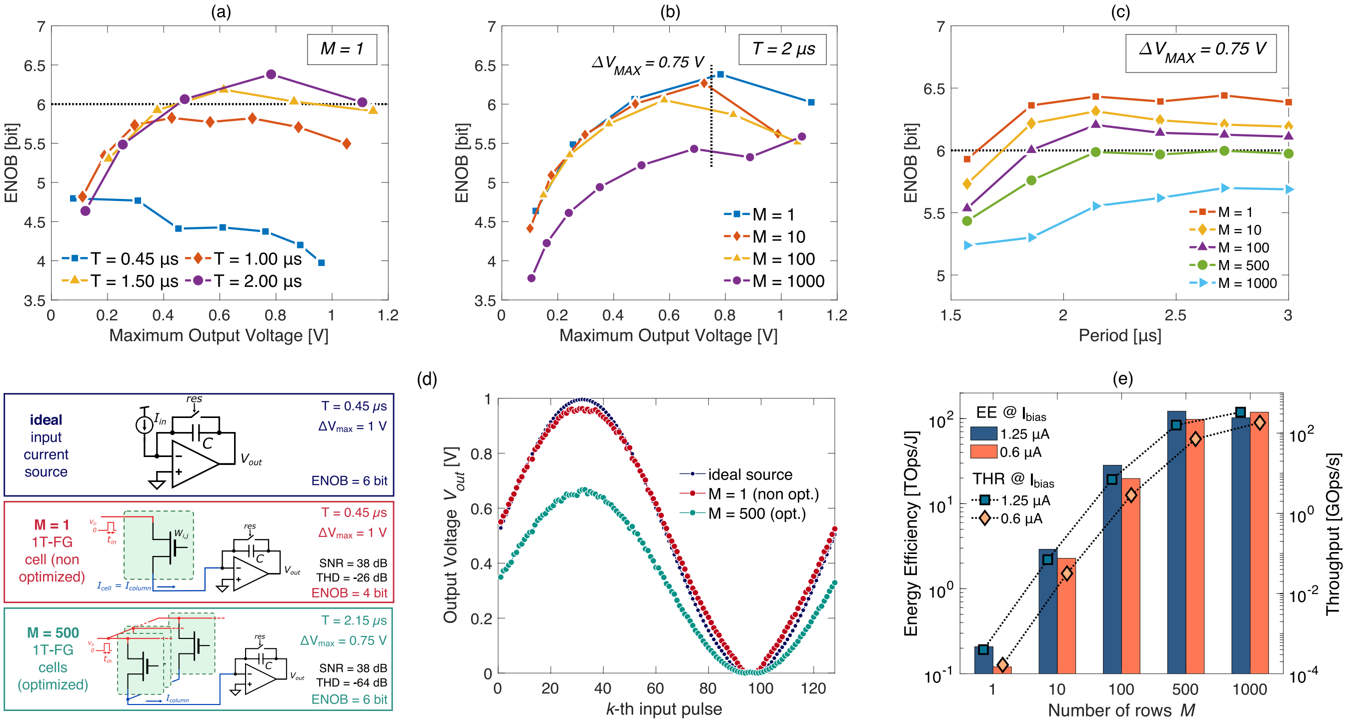

FIGURE 4. TD-VMM performance. (a) Single cell VMM (M = 1) ENOB as a function of the period T for decreasing maximum output voltage ΔV_{MAX}; (b) ENOB as a function of ΔV_{MAX} for a fixed $T = 2$ μs and for different sized VMMs (M from 1 to 1000); (c) ENOB as a function of T for a fixed $\Delta V_{MAX} = 0.75$ V and for different M, from 1 to 1000; (d) Output voltage sampled for an input pulse width modulated with a sinusoidal rule for various VMM settings (ideal current source, non optimized single memory cell, $M = 500$ optimized cells); (e) VMM Energy Efficiency and Throughput, as a function of M, for a target precision of 6 bits, considering two different biasing condition (i.e. different dissipated power values) of the OTA.

Finally, by selecting a ΔV_{MAX} of 0.75 V as an intermediate point where a close-to-maximum ENOB is achieved, we have increased the period with a fine granularity in Fig. 4(c): although for $M = 1$ a 6-bit ENOB is practically achieved with a period close to 1.5 μs, for increasing M the period must be relaxed up to 2.15 μs for a size of 500, while the case with $M = 1000$ saturates to ENOB $= 5.7$ in the investigated range. A detailed discussion about the limiting factors (i.e. THD and SNR) of the ENOB precision in Appendix E.

The effect of the VMM optimization, namely varying T and ΔV_{MAX} to reach the target ENOB, is illustrated in Fig. 4(d), which shows the sequence of output voltages sampled at the end of the integration, for a sequence of input pulse widths sampled from a sine shape spanning the whole pulse-width range $0 \ldots T$. An almost perfect sinusoidal shape is obtained when the OTA integrator is biased with an ideal current pulse generator, with $T = 0.45$ μs and $\Delta V_{MAX} = 1$ V, while the waveform appears quite distorted when a 1T-FG cell is considered (note the "clipped" output values close to the peak). However, after fine optimization ($\Delta V_{MAX} = 750$ mV, $T = 2.15$ μs), even the reported case with $M = 500$ reaches the required precision, as also qualitatively confirmed by the good sinusoidal shape of the output samples depicted in green.

The resulting EE and THR (defined in Appendix V-A) for a target accuracy of 6 bits are shown in Fig. 4(e), for different VMM sizes. The EE is evaluated for two different biasing conditions of the OTA, which in turn determine two different values of P_{diss}. A lower $I_{bias} = 0.6$ μA leads to a lower dissipated power but, recalling Fig. 3(f), a higher T needs to be used to ensure adequate precision. This leads to lower EE for the lower I_{bias}, but at $M = 1000$ an inversion occurs: the E_{diss} evaluated at $I_{bias} = 0.6$ μA is lower than the one evaluated at $I_{bias} = 1.25$ μA, meaning that, even if a higher T is needed, the lower P_{diss} makes it more convenient to bias the OTA with a smaller current. On the other hand, the THR is always higher at $I_{bias} = 1.25$ μA, as Fig. 4(e) clearly shows. Note that the THR is evaluated considering a squared matrix for the VMM, i.e. $N = M$.

IV. VMM AND NEURAL NETWORK BENCHMARKING

Table 1 shows a comparison of the TD-VMM with the other proposed analog VMMs in terms of accuracy, area occupation, EE and THR. The comparison regarding the accuracy is based on the ENOB, which provides information on the scalar product multiplication accuracy and would also enable a comparison with dedicated digital solutions. This is very useful since we treat the VMM as a standalone building block, that can be included in many different types of neural network. The classification accuracy of a complete neural network depends of course on both the accuracy of the employed VMMs and on the overall architecture. All performance parameters of the TD-VMM are assessed considering

TABLE 1. Comparison between state-of-the-art CMOS analog VMMs.

References	[15]	[10]	[9]	[12]	[17]	[13]	[16]	This Work
Approach	CM	CM	TD	TD	TD	CM	CM	**TD**
Tech. Node	180 nm	180 nm	14 nm	55 nm	55 nm	180 nm	180 nm	**180 nm**
Memory Type	Digital	Embedded NOR	ReRAM	Embedded NOR	1T1R	Single-Poly FG	Single-Poly FG	**Single-Poly FG**
ENOB [bit]	4	~ 5	< 8	6	6	6	6	**6**
VMM size	196×100 100×50 50×10	784×64 64×10	1024×1024	500×500	200×200	12×8	100×10	$\mathbf{500 \times 500}$
Equivalent single cell area (no ADC) [μm^2]	N/A	1.68	0.066	4.33	24	3	85.5	**2.15**
Equivalent single cell area (with ADC) [μm^2]	N/A	N/A	0.07**	N/A	N/A	N/A	N/A	**82.15***
EE (no ADC) [TOps/J]	~ 8	5	259	135	123.1	N/A	26.4	**122.3**
EE (with ADC) [TOps/J]	N/A	N/A	164**	N/A	N/A	N/A	N/A	**102.3***
THR [GOps/s]	1.7	N/A	5.46×10^3	20×10^3	630	N/A	19.9×10^{-3}	**333.17**
Results	Simulation	Experiment	Simulation	Simulation	Simulation	Sim. and Experiment	Sim. and Experiment	**Sim. and Exp.**

*: area and energy consumption of the ADC taken from [48].
**: area and energy consumption of the ADC are only partially included.

a 500×500 VMM: the VMM area normalized to a single cell (including the pure memory area and a fraction of the charge amplifier area, shared within the cells of the same column) is extracted as the ratio of the total area of a 500×500 memory array and 500 integrators to the total 250 000 cells. The result of 2.15 μm^2 is significantly lower than any other single poly solutions, and even tops the 55 nm commercial NOR flash memory proposed in [12], while it is worse than the 180 nm NOR flash [10] or the 14 nm ReRAM solution [9], that, however, exploits a more advanced and non-standard technology node with simulation-based FOMs. In terms of energy efficiency, the resulting EE is 122.3 TOps/J, almost one order of magnitude higher than most of the other solutions, and still comparable with the best ones which rely on non-standard processes. The corresponding THR results in 333.17 GOps/s. Although our VMM has been designed for a target precision of 6 bits, as most of the other solutions, higher ENOB values can be reached with a further optimization, at the cost of a worse tradeoff between the VMM area overhead and energy efficiency.

We designed the TD-VMM as a general purpose building block to be used in different neuromorphic applications: in order to fully exploit the advantages of the analog domain, and avoid additional conversions to the digital world, it could be possible to implement it in a fully analog neural network (i.e. cascaded layers of analog neurons realized by VMMs followed by analog activation functions [49]). However, analog to digital converters (ADCs) are sometimes mandatory, especially in convolutional networks that rely on VMMs to improve efficiency and computation time, but still need to be designed to work in a digital environment. In such scenarios, the presence of ADCs will add energy and area overhead to the network, since one ADC per column would be needed to convert each VMM output into a digital signal. We refer to state-of-the-art ADCs [48], [50] to evaluate their impact on area and energy efficiency: for example, the 9-bits, high-speed, low-power ADC presented in [48] contributes with an additional energy of 1.6 pJ to the E_{diss} term in (9), resulting in a slightly decreased EE of 102.27 TOps/J for the 500×500 VMM. On the other side, the 0.04 mm^2 ADC area, to be added to our value of 2.15 μm^2 by dividing it by the number of VMM rows, would degrade the normalized area value to 82.15 μm^2.

To assess the performance of the proposed analog VMM in a full neural network, we consider a 2 layer network (Fig. 5) for image recognition of the MNIST database. Fig. 5(a) shows a top level view of the network: it takes as input a 28×28 image of a handwritten digit and processes it through a Fully-Connected (FC) layer, followed by a ReLU (Rectified Linear Unit) activation function. Then, another FC layer operates on the resulting data, and a Softmax activation function performs the 10-digit final classification. A fully analog implementation is shown in Fig. 5(b): Voltage-Controlled Oscillators (VCOs) perform voltage to

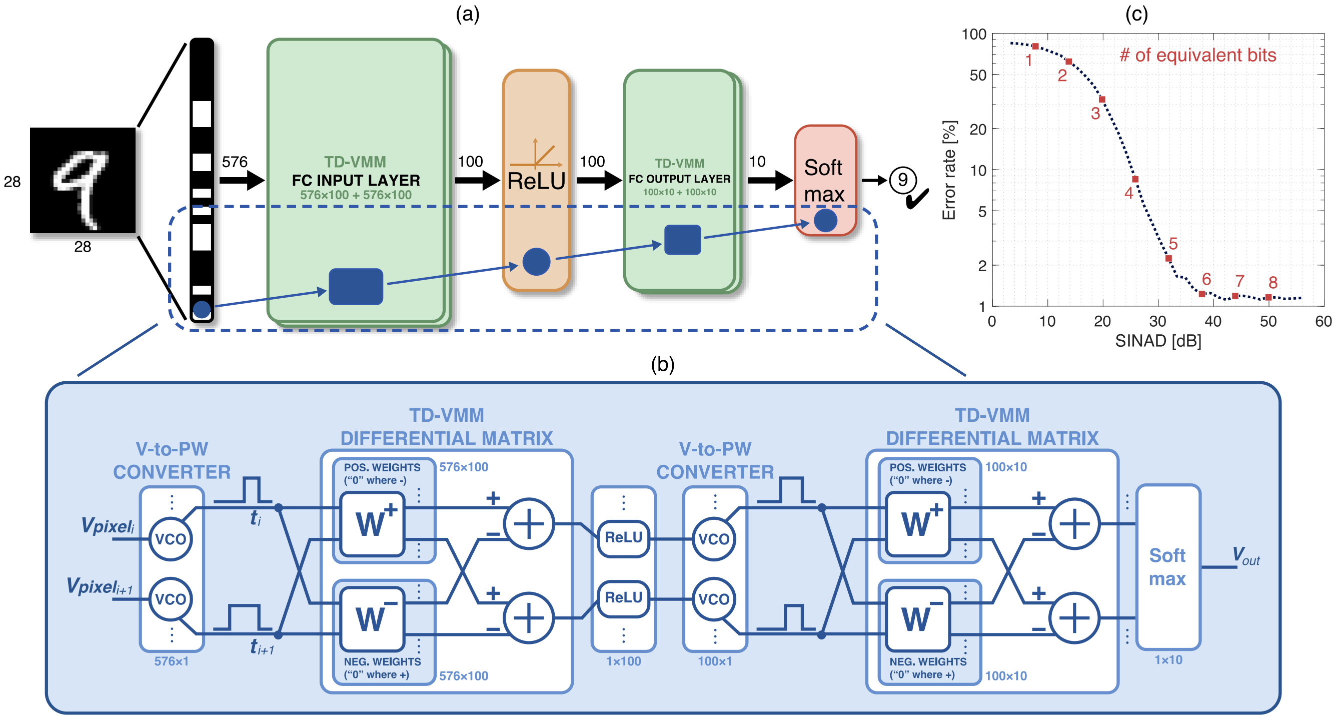

FIGURE 5. Assessment of a full neural network including the proposed VMM. (a) high-level design of the neural network resolving the MNIST dataset: the network is composed by two differential Fully-Connected (FC) layers, followed by non-linear activation functions (ReLU and Softmax respectively); a more detailed view of the operation of the network is shown in (b), where two TD-VMMs for each FC layers are used to implement positive and negative weights. In (c), the classification error rate of the analog network is shown as a function of SINAD and ENOB.

pulse-width conversions, so that the data is encoded in time pulses. Two pairs of TD-VMMs are used for each FC layer, implementing both positive and negative weights: the corresponding outputs of the two VMMs are then subtracted to obtain the final result. An analog ReLU can be easily realized (e.g. comparator controlling a pass gate), while analog Softmax designs can be found in literature [49], as mentioned. The assessment of the network was performed via software: training was done on a set of 8000 images, using a back propagation algorithm with floating point precision. Optimum weights were then extracted after a training of 50 epochs: tests on a set of 2000 images showed a classification accuracy of 98.9%. In order to evaluate the performance of an analog neural network when operating with a reduced precision and in the presence of noise and non-linearity, we emulated via software the operation performed by the TD-VMMs using fixed point operation and adding a white noise disturb corresponding to the SINAD at the output of each multiplication stage [16]:

$$y_j = W_{i,j} \cdot x_j + \alpha \frac{FS}{2} 10^{\frac{-SINAD}{20}}, \tag{5}$$

with α being a random gaussian variable (mean 0, standard deviation 1). Fig. 5(c) shows the error rate of the classification using the analog network when varying the SINAD (and therefore the ENOB): it is clear how a reduced precision can give satisfactory classification results, with a $<$ 3% error when using an ENOB $=$ 5, and an error of 1.23% with ENOB $=$ 6. This analysis validates our reduced precision assumptions and optimization of the TD-VMM for 6-bit operation.

V. CONCLUSION

In this paper, we have demonstrated a time-domain analog VMM realized using a commercial 0.18 μm CMOS process. The proposed VMM is suitable for neural networks performing offline training. The programmable weights are stored in single-poly, single-transistor, two-terminal FG cells, representing the smallest single-poly NVM cell available in the literature, with a size of $1.54 \times 1.12 = 1.72$ μm^2. Experiments demonstrate simple program and erase operations, requiring positive and relatively low voltage levels. Multi-time and multi-level analog programmability have been successfully verified, with a I_{high}/I_{low} ratio of more than three orders of magnitude, no visible read disturb in its range of operation, and over 64 threshold voltage levels. Retention can be further improved considering thicker gate oxide options. A dedicated compact gain-enhanced OTA has been designed with a tunable GBW that can be traded off with dissipated power, enabling a performance optimization flow involving required arithmetic precision, array size, power consumption and throughput. We show that an optimized 500×500 VMM with ENOB $=$ 6 reaches an EE of 122.3 TOps/J and a THR of 333.17 GOps/s, occupying a layout area of only

0.0537 mm^2. Reported FOMs are better than those of most other solutions, and comparable with respect to architectures based on non-standard technology nodes. The implementation of an image recognition neural network for the MNIST database based on the proposed analog VMM shows that an ENOB of 6 can provide a classification accuracy of 98.77%. These results are noteworthy, also considering that they are achieved by exploiting a reliable industry-standard low-cost single-poly CMOS process flow in an unconventional way, with FG devices not being part of the process design kit delivered by the foundry.

APPENDIX

A. VMM FIGURE-OF-MERIT DEFINITION

Definitions of ENOB, EE and THR are given below, based on the architecture and operation of the proposed TD-VMM.

1) EFFECTIVE NUMBER OF BITS (ENOB)

We follow the classical method inherited by the analog-to-digital converters of applying a sine input and calculating the Signal to Noise and Distortion Ratio (SINAD) at the output, which in turn is a combination of non-linearity (THD) and noise (SNR), according to:

$$10^{\frac{\mathrm{SINAD}}{10}} = 10^{\frac{\mathrm{SNR}}{10}} + 10^{\frac{-\mathrm{THD}}{10}}. \tag{6}$$

Then, the conversion between SINAD and ENOB is given by:

$$\mathrm{ENOB} = \frac{\mathrm{SINAD} - 1.76}{6.02}. \tag{7}$$

2) TROUGHPUT (THR)

We consider the $N \times (2M - 1)$ elementary operations, which are performed in parallel by a $M \times N$ VMM and can be repeated after a time equal to $T + T_{res}$ (where T_{res} is the time needed to reset the charge amplifier), therefore we get:

$$\mathrm{THR} = \frac{N(2M - 1)}{T + T_{res}}. \tag{8}$$

3) ENERGY EFFICIENCY (EE)

We consider the $(2M - 1)$ elementary operations performed by a single column (and a single integrator block): the energy per operation is computed by integrating the dissipated power $P_{diss}(t)$ over the operation period T (E_{diss}), by considering the contribution of the integrator and of the current flowing through the cells programmed with realistic weights. All the N columns operate in parallel during the period T, and each of them dissipates (on average) an energy equal to E_{diss}. Therefore, the EE can be evaluated as follows:

$$\mathrm{EE} = \frac{N(2M - 1)}{N \cdot E_{diss}}. \tag{9}$$

B. CELL PROGRAM AND ERASE

The carrier injection phenomena enabling cell program and erase operations, as illustrated respectively in Fig. 1(e) and (f), are discussed below.

1) PROGRAM

The program operation exploits channel hot electron injection (CHEI): with large V_{DS}, electrons coming from the source are accelerated by the longitudinal electric field, and undergo scattering in the vicinity of the drain, after which they can emerge with adequate direction and kinetic energy to have a finite probability to tunnel to the FG through the gate oxide (even if the electric field between drain and gate is unfavorable) and to be trapped in the FG. The net result is an increase of the equivalent threshold voltage V_{th} of the cell transistor.

2) ERASE

For higher V_{DS}, electrons in the channel are sufficiently accelerated to trigger impact ionization in the vicinity of the drain, with generation of electron-hole pairs. In this case, the dominant mechanism of charge injection in the FG becomes impact-ionized hot hole injections (IHHI), since holes are favored, with respect to electrons, by the accelerating electric field induced between drain and gate by the drain bias. The injection of positive charge in the FG causes a decrease of the cell transistor threshold voltage.

C. CPP AND ISPP PROGRAM SCHEME

A more detailed comparison between CPP and ISPP programming scheme is reported below.

With CPP, starting from a completely erased cell, the channel current is maximum and the transverse electric field (directed from the drain side channel to the FG, counteracting the injection of electrons) is minimum for the first pulse. For subsequent pulses, the available channel current decreases, as negative charge accumulates on the FG, leading to increased V_{th}, whereas the transverse electric field increases. Therefore, the increase of V_{read} is maximum for the first pulse and decreases at each successive one. On the other hand, the gradual increase of the programming voltage V_D performed by ISPP is transferred by capacitive coupling to the gate and partially compensates the decrease of the gate voltage V_G due to the accumulation of electrons in the floating gate, keeping the net charge injected into the FG per pulse almost constant in a wide range of the pulse train. This results in better linearity of program and erase operations with the number of pulses and finer granularity. The maximum value is achieved with a V_{prg} of 5.2 V, after $\sim$ 140 pulses, against the $\sim$ 75 of CPP for the same 80 ms pulse width. However, with ISPP the transverse electric field see a twofold increase at each subsequent pulse (decreasing FG potential, increasing V_D pulse amplitude), that explains why after 140 pulses the program dynamic is reversed and V_{read} decreases for subsequent pulses.

D. GAIN-ENHANCED OTA DESIGN METHODOLOGY

A differential amplifier for charge amplifier applications must satisfy the following conditions:

- high DC gain to keep the voltage at the inverting input node as constant as possible – in order to offer a constant bias voltage for the memory cells connected to the node;

- high Gain Bandwidth Product (GBW), that will determine the minimum input time (and therefore the integration period T) that guarantees the desired precision;
- minimize the area overhead to the VMM.

The transistor-level design of the OTA is shown in Fig. 3(a): with respect to a conventional two stage amplifier, the choice of a current mirror OTA topology helps improving the trade-off between power consumption and GBW, since the only high impedance node is the output node and therefore there is no internal low frequency pole and no need for compensation [51]. On the other side, the OTA typically suffers from a low DC gain A_0, given by:

$$A_0 = g_{m1} S r_{d5}, \tag{10}$$

where g_{m1} is the transconductance of the input pair, S is the current ratio of mirrors $M_2 - M_4$ and $M_3 - M_5$ and r_{d5} is the resistance seen from the output node V_{out}. In order to increase the DC gain, transistors M_{12} and M_{13} are added to shunt portions of the input stage current [52]. This results in lower current delivered to the output transistors, translating in higher output resistance values, while the gain of the input pair remains not affected by this change: the net result is an increase of A_0 of about $15/20$ dB [52].

In addition, although by varying the bias current value it is possible to increase the GBW of the device, the DC gain A_0 remains constant, since a different I_{bias} has the same but opposite effect on g_{m1} transconductance and r_{d5} output resistance.

E. VMM PRECISION OPTIMIZATION METHODOLOGY

In this section we discuss the dependence of the VMM precision (ENOB) on noise (SNR) and distortion (THD), as a result of several factors, among which we recall: unideal virtual ground of the OTA affecting the BL voltage (source of the 1T-FG cells) during the current integration, non-linear characteristics of the OTA, impact of the noise on both the 1T-FG cells and on the integrator operation.

The main trends can be understood by considering simulation results performed on a single cell VMM ($M = 1$), in Fig. 4(a), but can also be applied to higher values of M as investigated Fig. 4(b). For a given T, there is an optimum ΔV_{MAX} allowing to maximize the ENOB, as a result of two different limiting causes at the boundaries: higher ΔV_{MAX} requires higher I_{column}, leading to an increased variation of the virtual ground voltage due to limited GBW of the OTA, which results in a reduced THD because the 1T-FG cell $I_{DS,ON}$ is not constant; on the other hand, for low values of ΔV_{MAX}, noise becomes comparable with the reduced I_{column} signal and the reduced SNR becomes the limiting factor for the ENOB. A similar discussion can explain the behavior for a fixed ΔV_{MAX}, when the maximum period T is varied (Fig. 4(c)): a small pulse width requires a high current to charge the 0.6 pF capacitor to the same ΔV_{MAX}, modulating the virtual ground voltage and then the effective current $I_{DS,ON}$ provided by the 1T-FG cell in on state. On the other hand, while an extremely long pulse width requires small currents which can possibly become comparable to 1T-FG cell noise, noise averaging over time represents an effective filtering action preventing ENOB degradation, if ΔV_{MAX} is reasonably large. In addition, one should not that an extreme value for ΔV_{MAX}, that is close to 0 V or close to the supply voltage V_{DD}, can produce an increase of the THD due to the OTA transfer-characteristics offset and saturation, respectively.

The impact of the number of cells in the BL (i.e. M) shown in Fig. 4(b) and (c), can be understood if one consider that, for a fixed pulse width T and output swing ΔV_{MAX}, the total current charging the capacitor is equal to I_{column} (from (4)), meaning that each cell provides a fraction of the total current equal to $I_{cell} = I_{column}/M$. As the VMM size increases, the capacitive load seen at the input of the OTA is also higher, and the amplifier GBW is not large enough to retrieve immediately the BL node voltage to ground after a deviation caused by the rising edge of the input pulse: this leads to a degradation of the THD, that in turn limits the accuracy of the network.

REFERENCES

[1] V. Sze, "Designing hardware for machine learning: The important role played by circuit designers," *IEEE Solid State Circuits Mag.*, vol. 9, no. 4, pp. 46–54, Nov. 2017.

[2] X. Xu, Y. Ding, S. X. Hu, M. Niemier, J. Cong, Y. Hu, and Y. Shi, "Scaling for edge inference of deep neural networks," *Nature Electron.*, vol. 1, no. 4, pp. 216–222, Apr. 2018.

[3] K. Berggren *et al.*, "Roadmap on emerging hardware and technology for machine learning," *Nanotechnol.*, vol. 32, Jan. 2021, Art. no. 012002.

[4] M. R. Azghadi, C. Lammie, J. K. Eshraghian, M. Payvand, E. Donati, B. Linares-Barranco, and G. Indiveri, "Hardware implementation of deep network accelerators towards healthcare and biomedical applications," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 6, pp. 1138–1159, Dec. 2020.

[5] J. Lu, S. Young, I. Arel, and J. Holleman, "A 1 TOPS/W analog deep machine-learning engine with floating-gate storage in 0.13 μm CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 270–281, Jan. 2015.

[6] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. D. Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, and G. W. Burr, "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature*, vol. 558, no. 7708, pp. 60–67, Jun. 2018.

[7] M. M. Hasan and J. Holleman, "Implementation of linear discriminant classifier in 130 nm silicon process," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2018, pp. 1–5.

[8] M. Judy, N. C. Poore, P. Liu, T. Yang, C. Britton, D. S. Bolme, A. K. Mikkilineni, and J. Holleman, "A digitally interfaced analog correlation filter system for object tracking applications," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 9, pp. 2764–2773, Sep. 2018.

[9] M. J. Marinella, S. Agarwal, A. Hsia, I. Richter, R. Jacobs-Gedrim, J. Niroula, S. J. Plimpton, E. Ipek, and C. D. James, "Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 1, pp. 86–101, Mar. 2018.

[10] F. Merrikh-Bayat, X. Guo, M. Klachko, M. Prezioso, K. K. Likharev, and D. B. Strukov, "High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4782–4790, Oct. 2018.

[11] Z. Wang, Y. Chen, A. Patil, J. Jayabalan, X. Zhang, C.-H. H. Chang, and A. Basu, "Current mirror array: A novel circuit topology for combining physical unclonable function and machine learning," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 4, pp. 1314–1326, Apr. 2018.

[12] M. Bavandpour, M. R. Mahmoodi, and D. B. Strukov, "Energy-efficient time-domain vector-by-matrix multiplier for neurocomputing and beyond," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 66, no. 9, pp. 1512–1516, Sep. 2019.

[13] L. Danial, E. Pikhay, E. Herbelin, N. Wainstein, V. Gupta, N. Wald, Y. Roizin, R. Daniel, and S. Kvatinsky, "Two-terminal floating-gate transistors with a low-power memristive operation mode for analogue neuromorphic computing," *Nature Electron.*, vol. 2, no. 12, pp. 596–605, Dec. 2019.

[14] Y.-H. Kim, J.-M. Choi, J.-J. Woo, E.-J. Park, S.-W. Kim, and K.-W. Kwon, "A 16×16 programmable analog vector matrix multiplier using CMOS compatible Floating Gate device," in *Proc. Int. Conf. Electron. Inf. Commun. (ICEIC)*, Jan. 2019, pp. 1–4.

[15] J. Binas, D. Neil, G. Indiveri, S.-C. Liu, and M. Pfeiffer, "Precise neural network computation with imprecise analog devices," 2016, *arXiv:1606.07786*.

[16] M. Paliy, S. Strangio, P. Ruiu, T. Rizzo, and G. Iannaccone, "Analog vector-matrix multiplier based on programmable current mirrors for neural network integrated circuits," *IEEE Access*, vol. 8, pp. 203525–203537, 2020.

[17] S. Sahay, M. Bavandpour, M. R. Mahmoodi, and D. Strukov, "Energy-efficient moderate precision time-domain mixed-signal vector-by-matrix multiplier exploiting 1T-1R arrays," *IEEE J. Explor. Solid-State Comput. Devices Circuits*, vol. 6, pp. 18–26, 2020.

[18] W. Wan, R. Kubendran, S. B. Eryilmaz, W. Zhang, Y. Liao, D. Wu, S. Deiss, B. Gao, P. Raina, S. Joshi, H. Wu, G. Cauwenberghs, and H.-S.-P. Wong, "33.1 A 74 TMACS/W CMOS-RRAM neurosynaptic core with dynamically reconfigurable dataflow and *in-situ* transposable weights for probabilistic graphical models," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 498–500.

[19] M. Yamaguchi, G. Iwamoto, Y. Nishimura, H. Tamukoh, and T. Morie, "An energy-efficient time-domain analog CMOS BinaryConnect neural network processor based on a pulse-width modulation approach," *IEEE Access*, vol. 9, pp. 2644–2654, 2021.

[20] A. Chen, "A review of emerging non-volatile memory (NVM) technologies and applications," *Solid-State Electron.*, vol. 125, pp. 25–38, Nov. 2016.

[21] I. Hubara, M. Courbariaux, and D. Soudry, "Quantized neural networks: Training neural networks with low precision weights and activations," *J. Mach. Learn. Res.*, vol. 18, pp. 1–30, Jan. 2018.

[22] W. Haensch, T. Gokmen, and R. Puri, "The next generation of deep learning hardware: Analog computing," *Proc. IEEE*, vol. 107, no. 1, pp. 108–122, Jan. 2019.

[23] A. Sebastian, M. L. Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnol.*, vol. 15, no. 7, pp. 529–544, 2020.

[24] X. Sun, P. Wang, K. Ni, S. Datta, and S. Yu, "Exploiting hybrid precision for training and inference: A 2T-1FeFET based analog synaptic weight cell," in *IEDM Tech. Dig.*, Dec. 2018, pp. 3.1.1–3.1.4.

[25] A. Ash-Saki, M. N. I. Khan, and S. Ghosh, "Reconfigurable and dense analog circuit design using two terminal resistive memory," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 3, pp. 1596–1608, Jul. 2021.

[26] V. Saxena, "Mixed-signal neuromorphic computing circuits using hybrid CMOS-RRAM integration," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 2, pp. 581–586, Feb. 2021.

[27] M. Jerry, P.-Y. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu, and S. Datta, "Ferroelectric FET analog synapse for acceleration of deep neural network training," in *IEDM Tech. Dig.*, Jan. 2017, p. 6.2.1–6.2.4.

[28] I.-T. Wang, Y.-C. Lin, Y.-F. Wang, C.-W. Hsu, and T.-H. Hou, "3D synaptic architecture with ultralow sub-10 fJ energy per spike for neuromorphic computation," in *IEDM Tech. Dig.*, Dec. 2014, p. 28.

[29] E. Perez, C. Zambelli, M. K. Mahadevaiah, P. Olivo, and C. Wenger, "Toward reliable multi-level operation in RRAM arrays: Improving post-algorithm stability and assessing endurance/data retention," *IEEE J. Electron Devices Soc.*, vol. 7, pp. 740–747, 2019.

[30] A. Di Bartolomeo, H. Rücker, P. Schley, A. Fox, S. Lischke, and K.-Y. Na, "A single-poly EEPROM cell for embedded memory applications," *Solid-State Electron.*, vol. 53, no. 6, pp. 644–648, Jun. 2009.

[31] B. Rumberg and D. W. Graham, "A floating-gate memory cell for continuous-time programming," in *Proc. IEEE Int. Midwest Symp. Circuits Syst.*, Aug. 2012, pp. 214–217.

[32] Y.-D. Wu, K.-C. Cheng, C.-C. Lu, and H. Chen, "Embedded analog nonvolatile memory with bidirectional and linear programmability," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 59, no. 2, pp. 88–92, Feb. 2012.

[33] S.-H. Song, K. C. Chun, and C. H. Kim, "A bit-by-bit re-writable eflash in a generic 65 nm logic process for moderate-density nonvolatile memory applications," *IEEE J. Solid-State Circuits*, vol. 49, no. 8, pp. 1861–1871, Aug. 2014.

[34] F. M. Bayat, X. Guo, H. A. Om'mani, N. Do, K. K. Likharev, and D. B. Strukov, "Redesigning commercial floating-gate memory for analog computing applications," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2015, pp. 1921–1924.

[35] C. Li, J.-C. Li, J. Shang, W.-X. Li, and S.-Q. Xu, "Multitime programmable memory cell with improved MOS capacitor in standard CMOS process," *IEEE Trans. Electron Devices*, vol. 62, no. 8, pp. 2517–2523, Aug. 2015.

[36] D. Basford, M. Judy, P. Liu, and J. Holleman, "A sub-1 V-read flash memory in a standard 130 nm CMOS process," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2018, pp. 1–4.

[37] S. Xu, H. Wang, J. Wu, L. Zheng, and J. Diao, "A new multitime programmable non-volatile memory cell using high voltage NMOS," *Microelectron. Rel.*, vols. 88–90, pp. 169–172, Sep. 2018.

[38] J.-M. Choi, E.-J. Park, J.-J. Woo, and K.-W. Kwon, "A highly linear neuromorphic synaptic device based on regulated charge trap/detrap," *IEEE Electron Device Lett.*, vol. 40, no. 11, pp. 1848–1851, Nov. 2019.

[39] C.-P. Chung and K.-S. Chang-Liao, "A highly scalable single poly-silicon embedded electrically erasable programmable read only memory with tungsten control gate by full CMOS process," *IEEE Electron Device Lett.*, vol. 36, no. 4, pp. 336–338, Apr. 2015.

[40] L. Milani, F. Torricelli, and Z. M. Kovács-Vajna, "Single-poly-EEPROM cell in standard CMOS process for medium-density applications," *IEEE Trans. Electron Devices*, vol. 62, no. 10, pp. 3237–3243, Oct. 2015.

[41] P. Song, Q. Sun, M. Qi, D. Qiao, and C. Bai, "Cost-effective reliable EEPROM cell based on single-poly structure," in *Proc. Int. Conf. IC Design Technol. (ICICDT)*, Jun. 2019, pp. 1–4.

[42] S. Shukuri, S. Shimizu, N. Ajika, T. Ogura, M. Mihara, Y. Kawajiri, K. Kobayashi, and M. Nakashima, "A 10 k-cycling reliable 90 nm logic NVM 'eCFlash' (embedded CMOS flash) technology," in *Proc. IEEE Int. Memory Workshop*, May 2011, pp. 1–2.

[43] S.-K. Park, K.-I. Choi, and N.-Y. Kim, "Improved performance of novel vertical assist operating select gate lateral coupling cell for logic non-volatile memory," *IEEE Electron Device Lett.*, vol. 37, no. 4, pp. 412–415, Apr. 2016.

[44] Y. Roizin, E. Pikhay, V. Dayan, and A. Heiman, "High density MTP logic NVM for power management applications," in *Proc. IEEE IMW*, May 2009, pp. 1–2.

[45] Y.-F. Chen, J. Gong, W.-J. Tung, S.-W. Chou, and E. S. Jeng, "Characteristics of n-channel MOSFETs with tailored source/drain extension for mask ROM and EEPROM applications," *IEEE Trans. Electron Devices*, vol. 56, no. 9, pp. 2099–2106, Sep. 2009.

[46] *The MNIST Database of Handwritten Digits*. Accessed: Oct. 10, 2021. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[47] K.-D. Suh, B.-H. Suh, Y.-H. Lim, J.-K. Kim, Y.-J. Choi, Y.-N. Koh, S.-S. Lee, S.-C. Kwon, B.-S. Choi, J.-S. Yum, J.-H. Choi, J.-R. Kim, and H.-K. Lim, "A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme," *IEEE J. Solid-State Circuits*, vol. 30, no. 11, pp. 1149–1156, Nov. 1995.

[48] H. S. Bindra, A.-J. Annema, S. M. Louwsma, and B. Nauta, "A 0.2–8 MS/s 10b flexible SAR ADC achieving 0.35–2.5 fJ/conv-step and using self-quenched dynamic bias comparator," in *Proc. Symp. VLSI Circuits*, Jun. 2019, pp. C74–C75.

[49] M. Vatalaro, T. Moposita, S. Strangio, L. Trojman, A. Vladimirescu, M. Lanuzza, and F. Crupi, "A low-voltage, low-power reconfigurable current-mode softmax circuit for analog neural networks," *Electronics*, vol. 10, no. 9, p. 1004, Apr. 2021.

[50] S. Jeong, W. Jung, D. Jeon, O. Berenfeld, H. Oral, G. Kruger, D. Blaauw, and D. Sylvester, "A 120nW 8b sub-ranging SAR ADC with signal-dependent charge recycling for biomedical applications," in *2015 Symp. VLSI Circuits*, Jun. 2015, pp. C60–C61.

[51] T.-H. Lin, C.-K. Wu, and M.-C. Tsai, "A 0.8-V 0.25-mW current-mirror OTA with 160-MHz GBW in 0.18-μm CMOS," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 54, Feb. 2007, pp. 131–135.

[52] L. Yao, M. S. J. Steyaert, and W. Sansen, "A 1-V 140-μW 88-dB audio sigma-delta modulator in 90-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 39, no. 11, pp. 1809–1818, Nov. 2004.

TOMMASO RIZZO received the B.S. and M.S. degrees *(cum laude)* in EE from the University of Pisa, in 2017 and 2019, respectively. He is currently pursuing the Ph.D. degree in electronics in a joint program with the University of Pisa and Quantavis s.r.l., in the field of analog and mixed-signal IC design using standard and non-standard CMOS technologies. From 2014 to 2019, he was an "Allievo Ordinario" at the Sant'Anna School of Advanced Studies, Pisa. In 2017, he was at Fermilab, Batavia, IL, USA, as a Visiting Student, and in 2019, he joined imec, Eindhoven, The Netherlands, working on a wireless powering receiver system for deep implants as his master's thesis project. His research interests include the design of CMOS analog blocks for DNNs and the development of wireless power transfer solutions for IMDs.

SEBASTIANO STRANGIO (Member, IEEE) was at the University of Udine, Udine, Italy, as a Temporary Research Associate, from 2013 to 2016, and at Forschungszentrum Jülich, Jülich, Germany, as a Visiting Researcher, in 2015, researching on TCAD simulations, design, and characterization of TFET-based circuits. From 2016 to 2019, he was at LFoundry, Avezzano, Italy, where he worked as a Research and Development Process Integration and Device/TCAD Engineer, with main focus on the development of a CMOS image sensor technology platform. He is a Researcher of electronics with the University of Pisa, Pisa, Italy. He has authored and coauthored over 30 articles, most of them published in IEEE journals and conference proceedings. His research interests include technologies for innovative devices (e.g. TFETs) and circuits for innovative applications [CMOS analog building blocks for deep neural networks (DNNs)], CMOS image sensors, power devices, and circuits based on wide-bandgap materials.

GIUSEPPE IANNACCONE (Fellow, IEEE) received the M.S. and Ph.D. degrees in electronic engineering from the University of Pisa, Pisa, Italy, in 1992 and 1996, respectively. He is a Professor of electronics with the University of Pisa. He has authored and coauthored more than 230 articles published in peer-reviewed journals and more than 160 papers in proceedings of international conferences, gathering more than 8500 citations on the Scopus database. His research interests include quantum transport and noise in nanoelectronic and mesoscopic devices, development of device modeling tools, new device concepts and circuits beyond CMOS technology for artificial intelligence, cybersecurity, implantable biomedical sensors, and the internet of things. He is a fellow of the American Physical Society.