IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication **unknown**, date of current version **unknown**.

Digital Object Identifier ***unknown***

Analog Vector-Matrix Multiplier based on Programmable Current Mirrors for Neural Network Integrated Circuits

Maksym Paliy[1], Sebastiano Strangio[1], Member, IEEE, Piero Ruiu[1,2], Tommaso Rizzo[1], Giuseppe Iannaccone[1,2], Fellow, IEEE

[1]Department of Information Engineering, University of Pisa, Pisa, 56122 Italy
[2]Quantavis s.r.l., Largo Padre Renzo Spadoni snc, Pisa, 56126 Italy

Corresponding author: Sebastiano Strangio (e-mail: sebastiano.strangio@unipi.it).

This work was partially supported by the European Commission through the QUEFORMAL H2020 Project (contract no. 829035) and jointly by the European Commission and the Italian Ministry of Industry and Economic Development (MISE) through the ECSEL CHARM Project (contract no. 876362).

ABSTRACT In this paper we propose a CMOS Analog Vector-Matrix Multiplier for Deep Neural Networks, implemented in a standard single-poly 180 nm CMOS technology. The learning weights are stored in analog floating-gate memory cells embedded in current mirrors implementing the multiplication operations. We experimentally verify the analog storage capability of designed single-poly floating-gate cells, the accuracy of the multiplying function of proposed tunable current mirrors, and the effective number of bits of the analog operation. We perform system-level simulations to show that an analog deep neural network based on the proposed vector-matrix multiplier can achieve an inference accuracy comparable to digital solutions with an energy efficiency of 26.4 TOPs/J, a layer latency close to 100 µs and an intrinsically high degree of parallelism. Our proposed design has also a cost advantage, considering that it can be implemented in a standard single-poly CMOS process flow.

INDEX TERMS Analog Neural Network, CMOS, Current-Mirror, DNN, Floating-Gate

I. INTRODUCTION

The increasing requirements of cognitive capabilities in electronic systems is driving research toward highly efficient and dense specialized hardware to implement Deep Neural Networks (DNNs). Migration toward architectures beyond the Von Neumann paradigm and towards in-memory computation may lead to an improvement in terms of Energy Efficiency (EE), defined as the ratio of the number of elementary operations to the energy consumed to perform these operations, and of throughput, i.e. the number of performed elementary operations per unit time. In the implementation of a DNN, the most recurring complex operation is the vector-matrix multiplication, i.e. the multiplication of a vector of features (e.g. input of a layer) with a matrix of learning weights, that are constant quantities during the inference phase. The large number of multi-bit elementary arithmetic operations performed by the vector-matrix multiplier (VMM) and the heavy data exchange between the memory and logic elements represent the main limiting factors of both EE and throughput in conventional digital CPU architectures [1], [2], [3], [4]. The recurring nature of these arithmetic operations can be exploited by taking advantage of the parallel computing capability of GPUs [5] and of embedded ASIC accelerators [6], [7]. Parallelism in computation and in memory access can be better exploited through in-memory computing architectures, consisting of a large number of modularized processing elements distributed in space and operating in parallel, where each processing element contains both the logic and the memory to perform the assigned partial processing task.

In this context, analog circuits enable the implementation of in-memory computing architectures where analog computations are performed by exploiting fundamental circuital laws and devices properties. Analog processing blocks are usually affected by circuit nonidealities such as noise, non-linearity and process variations. However, their finite precision can be well tolerated by the inherent capabilities of neuromorphic networks, which feature high tolerance of functional parameter variations [8] and to limited precision [9].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037017, IEEE Access

In this paper we focus on the design, operation and experimental validation of an analog VMM realized by means of an array of tunable conversion-factor Current Mirrors (CMs) based on single-poly floating-gate (FG) cells, as illustrated in Fig.1. In each tunable CM, the current conversion ratio I_{out}/I_{in} can be interpreted as the weight associated to the charge stored in the FG. The FG cell is obtained by an n-type MOSFET (nMOS) and a p-type MOSCAP (pCAP) sharing an isolated polysilicon-gate. The multiplier is realized in a standard 180 nm single-poly CMOS technology, by using devices with 3.3 V nominal voltage domain realized with a thick gate oxide (~7 nm), typically required to achieve the ten-year retention time adequate for a non-volatile memory.

Single-poly FG cells have been designed and fabricated. In particular, we have experimentally verified the possibility to program an analog weight with a current conversion ratio equivalent to a nominal 8-bit integer. We have performed system-level simulations of trained DNNs, using parallel VMMs to implement both fully-connected and convolutional layers. The inference accuracy of the same network operated either with floating-point precision or with reduced bitwidth fixed-point precision was compared. This analysis has been repeated for a simple DNN purposely designed to classify the MNIST dataset [10], as well as for AlexNet [11] employed for ImageNet [12] dataset classifications. We have verified that a reduced bitwidth might allow for comparable inference accuracy as the original network, with a minimum number of equivalent bits that is a function of the particular application (dataset and DNN architecture). Then, we have selected a 6-bit specification to design an analog CM-based VMM and have proposed a general design flow applicable to different CM topologies. We demonstrate with experiments and simulations the operation and performance of CM-VMM. The best option exhibits an energy efficiency of 26.4 TOPs/J and a layer latency of 100 µs. A 100×10 VMM has an area of 0.868 mm^2 and a throughput of 19.9 MOPs/s, with each multiplying cell of the matrix occupying a layout area of 85.5 μm^2.

The remainder of this paper is organized as follows. In section II we present a discussion on the background of this work, by reviewing approaches using CMOS analog circuits to implement neuromorphic building blocks. In section III we present the CM-VMM basic principle and we introduce its main figures of merit (FOMs). Experimental results measured on silicon demonstrators are shown in Section IV, proving the analog multi-level storage capability of single-poly FG cells. Measurements on an experimental proof-of-concept of a programmable CM multiplier are also shown. Then, in Section V, four possible implementations of FG CM-VMM are designed and compared, in order to choose the best CM topology for the implementation of a FG-cell CM for a given ENOB specification. Our chosen design is finally benchmarked against state-of-the-art VMMs in Section VI. The conclusions of the paper are drawn in Section VII.

II. BACKGROUND

As DNNs are concerned, it has been shown that digital approaches with fixed-point data representation can provide comparable classification accuracy to a floating-point computation [13]. In addition, due to the intrinsic resilience of DNN algorithms to noise and uncertainty [8], data representation based on a limited number of bits reduces the arithmetic complexity of processing elements, leading to an improvement of both power consumption and computing time, possibly without losing classification accuracy [9]. In this regard, different approaches have been proposed, for instance relying on reduced bitwidth of the weight [14], of

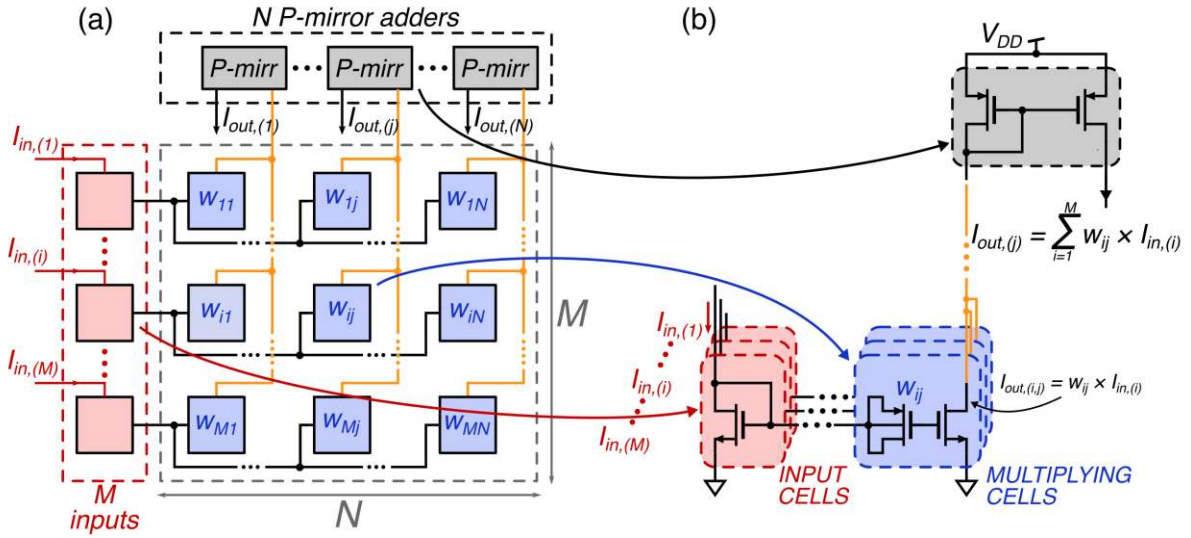

FIGURE 1. (a) Architecture of the analog M×N Vector-Matrix-Multiplier based on an array of "M" input cells (in red), a "M×N" matrix of multiplying cells (in blue) and an array of "M" p-type current-mirrors (in grey). (b) Possible circuital realization of the current-mirrors implementing the input cells and multiplying cells (multiplier blocks), and of the p-mirror (summation block).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037017, IEEE Access

both weights and activation function [15], or by implementing the entire network with a limited data precision [16].

As discussed in the introduction, this consideration opened the opportunity to exploit analog computing circuits in implementing DNN blocks. Several papers have proven the capability of analog computing elements to achieve an acceptable trade-off between algorithmic accuracy and numerical precision. Analog solutions are also suitable to be implemented with an in-memory circuit architecture [17], [18], avoiding costly memory access.

In addition, analog data might be stored in an analog non-volatile memory. Innovative non-volatile memory solutions such as the Resistive Random Access Memories (RRAMs) have been proposed in the literature for this tasks, such as oxygen vacancy memory (OxRAM) [19], conductive bridging memory (CBRAM) [20], and spin-transfer torque magnetic memory (STT-MRAM) [21]. However, the intrinsic variability of OxRAMs and CBRAMs makes them not suitable for very large-scale integration; on the other hand, despite the high industrial maturity of STT-MRAMs, they are intrinsically bistable and are therefore not suitable as analog non-volatile memories, which would require continuous tuning. In fact, simulations of DNNs based on RRAMs have been recently proposed [22], [23], [24], [25], [26], but the lack of experimental demonstrators suggests that viable alternatives must be investigated. A worthy option is the industry-standard double-poly embedded FG memory cells, which have been proposed for similar applications [27], [28]. In fact, they can rely on the fine tuning of stored charge (up to 4-bit single transistor memory cells have reached the market with a tunable 16-level threshold voltage and 10-year retention time [29]). However, the double-poly process flow is relatively expensive and the geometry of each single cell cannot be independently modified by designers, since the layout of an FG array is generally provided as foundry intellectual property [27], [30], [31]. An interesting option is to use single-poly embedded non-volatile cells, where the FG can be realized with a floating polysilicon area among two planar MOS devices, at the cost of larger area occupation [18], [32], [33] with respect to the double-poly case.

Different techniques have been proposed to perform a vector-matrix multiplication in the analog domain: time-domain approaches [22], [34] and current-mode sum operation [18], [27], [32], [33], [35]. Current-mode operation can be implemented by relying on the addition performed using Kirchhoff's current law; currents resulting from weight multiplication of different inputs are added by letting all currents flow to the same node.

III. CURRENT MIRROR VMM BASIC PRINCIPLE

The basic principle of an analog VMM implemented with CMs and the representative FOMs used in this paper are discussed in this section. In subsection III-A, the concept of CMs with tunable conversion factors used as current multipliers is introduced. A discussion on the VMM operation is proposed, emphasizing nonidealities in terms of both linearity and noise immunity level, and their impact on the maximum achievable accuracy. In subsection III-B, FOMs normally used for generic analog-to-digital converters (ADCs), such as the Signal-to-Noise And Distortion ratio (SINAD) and the Equivalent Number Of Bits (ENOB) are introduced and matched to the particular VMM design parameters.

A. CURRENT-MIRROR VMM BASIC PRINCIPLE

Fig.1(a) sketches the architecture of an analog current-mode M×N VMM, with M input currents ($I_{in,(i)}$ is a generic input, for i = 1…M), M×N weights and multiplying blocks in the matrix (each element is $w_{(i,j)}$), and N output currents (a generic output is $I_{out,(i,j)}$, for j = 1…N). Each input signal is applied to all the matrix cells in the same row, where the multiply operation is performed between each row input and the corresponding weight in the cell, according to

$$I_{out,(i,j)} = I_{in,(i)} \times w_{(i,j)}. \quad (1)$$

The output of the column is then obtained by summing over all terms to implement the scalar product operation

$$I_{out,(j)} = \sum_i^M I_{out(i,j)} = \sum_i^M I_{in,(i)} \times w_{(i,j)}. \quad (2)$$

The VMM basic implementation proposed in this paper is detailed in Fig.1(b), which show the CM approach where the current entering in an "input cell" (block in red) is multiplied by a scaling-factor by a "multiplying cell" (in blue) and provided as an output current, while all the currents of the same column are summed at the same circuit node. An additional p-type CM (in grey) is also added at the top of each column to provide the $I_{out,(j)}$ with the appropriate direction.

The storage capability of the multiplying cell associated to a generic $w_{(i,j)}$ is obtained via a FG cell, implemented by an nMOS sharing an FG with a pCAP. By relying on specific programming and erasing schemes, charge can be added to or removed from the FG. The net charge in the FG results in a shift $\Delta V_{th,(i,j)}$ of the threshold voltage determining the current magnification factor (i.e. the weight). For a given input current $I_{in,(i)}$, if the nMOS is operated in the subthreshold region, the corresponding output current $I_{out,(i,j)}$ depends exponentially on the $\Delta V_{th,(i,j)}$. Ideally, we have:

$$I_{out,(j)} = \sum_i^M I_{in,(i)} \times e^{\frac{\Delta Vth_{(i,j)}}{\eta V_T}} \quad (3)$$

where the exponential represents ideal weight,

$$w_{(i,j)}^{ideal} = e^{\frac{\Delta Vth_{(i,j)}}{\eta V_T}}. \quad (4)$$

Beyond enabling a wide range of variations of the output of the multiplying operation, sub-threshold operation regime is also beneficial to reduce power consumption [27], [31], [32], [35], [36].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037017, IEEE Access

Practical CMs do not exhibit the ideal behavior described by Eq.(3). Indeed, one should note the different V_{DS} of the input and multiplying cells. The small output resistance of short channel devices can thus degrade the linearity. This weakness can be worsened if devices with poor electrostatics are used, due to finite pCAP capacitance. An additional degradation arises if the input current becomes too low, due to poor transistor saturation when the V_{DS} of the diode-connected nMOS input cell becomes comparable with V_T. The non-linearity can be described in terms of Total Harmonic Distortion (THD).

Another root cause of precision degradation comes from intrinsic noise sources of the devices implementing the CM. The Signal-to-Noise Ratio (SNR) of the CM output current increases with the input current, and it inversely depends on the bandwidth [35]. Furthermore, for short channel devices, it decreases with the square of the channel length [37].

Provided that in analog circuits both noise and nonlinearity can severely impact the accuracy of the analog function, distortion and noise nonidealities are normally considered together within the Signal-to-Noise And Distortion ratio (SINAD) [38], which depends on SNR and THD as Eq.(5):

$$10^{-\frac{SINAD}{10}} = 10^{-\frac{SNR}{10}} + 10^{\frac{THD}{10}} \tag{5}$$

THD, SNR and SINAD are all expressed in dB and their definition are given in Appendix A.

B. FIGURES OF MERIT FOR ANALOG MULTIPLIERS

When DNNs consisting of multiple layers are considered (e.g. AlexNet [11]), the VMM arrays become the dominant functional blocks in the system, the main factor determining total area occupation and power consumption [11]. The design of an efficient analog VMM then involves different trade-offs among performance (throughput), EE, computation accuracy, and area occupation.

To provide a comparison with DNN implemented in digital architectures, FOMs for analog VMMs are normally expressed in terms of elementary operations, such as P-bit (where P is the bit-width) multiplications and additions. An M×N VMM includes N columns of M-sized multiply-and-accumulate (MAC) operations as shown in Eq. (2). We consider a number of M multiplications and M-1 additions per each MAC, corresponding to a total number of (2M-1)×N elementary operations in a VMM.

The (2M-1)×N elementary operations are performed in parallel in a VMM, then the throughput is given by the ratio of (2M-1)×N to the worst case time T_{op} needed by the CM multiplier to provide an output current corresponding to the expected result (within a confidence interval dependent on the assumed accuracy) in response to an input current step.

The EE is the calculated as the ratio of the (2M-1)×N parallel operations to the average energy consumed by the VMM to perform a vector-matrix multiplication (i.e. the consumed power integrated over the T_{op}). The energy is extracted using actual trained weights and it is averaged over a number of operations, each corresponding to an input array related to an actual input of the tested database (i.e. test images in case of MNIST or ImageNet).

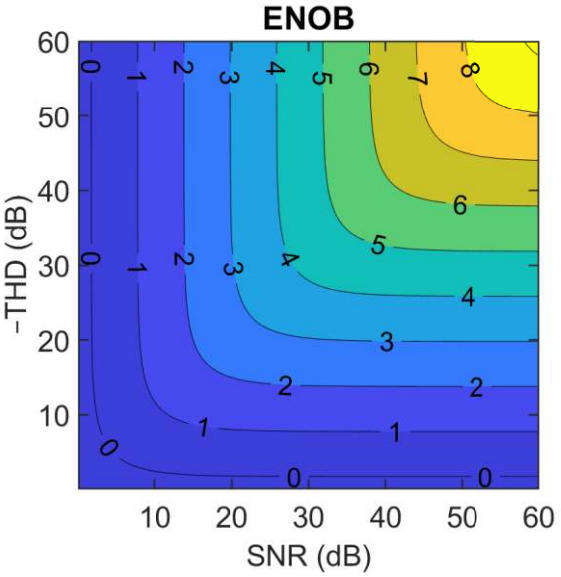

FIGURE 2. Contour plot of ENOB as a function of THD and SNR.

The accuracy of an analog VMM can be described by the SINAD, which can be related to linearity and noise immunity. In order to enable an intuitive comparison between the precision of an analog function and its digital counterpart, we can use the Effective Number of Bits (ENOB) linked to SINAD as [38]:

$$ENOB = \frac{SINAD - 1.76}{6.02} \tag{6}$$

Fig.2 depicts a contour plot of the ENOB as defined in Eq.(6), as a function of SNR and THD. SINAD, and therefore ENOB, is generally limited by the smaller between SNR and –THD. This plot is relevant in the choice of design trade-offs, since in several cases both SNR and THD play a role in the accuracy of an analog function. In fact, in most cases they should be balanced up in order to get a fine-grain optimization of the ENOB. In the definition of the ENOB given by Eq.(6), it is assumed that a sinusoidal input signal spanning the full-scale of the ADC input swing is used. Similarly, these definitions can be adapted to an analog circuit, where SNR includes all kind of noises which affects the circuit, while THD accounts for the nonlinear behavior of its transfer function. For consistency, in our study we use a sine waveform for the input current spanning between 0 and the target maximum input current $I_{in,MAX}$, also referred to as full-scale (FS) hereafter. When the ENOB characterization is performed for a unitary weight (i.e. $I_{out,MAX} = I_{in,MAX}$), the FS current levels as for I_{in} are also spanned by the I_{out} waveform. On the other hand, when the ENOB characterization is performed with a weight < 1, the resulting peak-to-peak value of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037017, IEEE Access

sinusoidal output current is $I_{out,MAX} = I_{in,MAX} \times w < FS$. To account for this partial sweep of the assumed full-scale of the output, a "$-\log_2(w)$" correction term is added in Eq.(6) to extract the equivalent full-scale ENOB by a projection.

IV. EXPERIMENTS ON SINGLE-POLY FG CURRENT MIRROR VMM

In this section, the electrical characterization of single-poly FG cells fabricated with UMC 0.18 µm CMOS technology is discussed. The analog storage capability with a possible current resolution larger than 8 bits (i.e. $I_{out,(i,j)}/I_{in,(i)} < 256^{-1}$) is first demonstrated. Then, a simple CM multiplier implemented with these cells is measured at different stored weight conditions. A good matching between experiments and simulations is demonstrated.

With reference to the CM implementation shown in Fig.1(b), the non-ideal coupling between the pCAP and the nMOS enhances the asymmetry between the input cell and the multiplying cell, which is already in place due to difference in V_{DS}. This asymmetry leads to a linearity degradation, which could be avoided by using a very large pCAP (so that $A_{pCAP} \gg A_{nMOS}$) resulting in an almost ideal coupling factor, but this cannot be appointed as a recommended solution for obvious reasons. A better option to increase the symmetry can be the use of an additional pCAP in the input cell. In this case, the FG on the input cell is not used to store data but just for electrostatic symmetry.

Experimental data for a symmetric CM are reported in Fig.3. The CM is realized with 0.5 µm long nMOS transistors sharing the floating poly with a pCAP area 49 times larger than the nMOS gate area, while the control gate (CG) is the N-well hosting the pCAP shorted with the P-diffusions implementing its S/D regions. All transistors have a 3.3V nominal voltage.

Fig.3(a) and (b) report the voltage levels used for the program and erase operations, which are both possible by applying positive voltage pulses on the CG and D terminals, activating different gate injection phenomena in agreement with [39]: for a V_{DS} in the range 4.5 V ~ 6.5 V, at high V_{CG-S} voltages (>3V) both channel hot electron injection (CHEI) and impact-ionized hot-electron injection (IHEI) lead to an increase of the equivalent V_{th}, while the impact-ionized hot-hole injection (IHHI) is the dominant mechanism at relatively low V_{CG-S} voltage (e.g. 1V ~ 1.5V) leading to a V_{th} variation in the opposite direction. This means that the threshold voltage can be moved in both directions without the need to design complicated circuitry to generate the negative voltage levels normally needed to reset FG memory cells. It is important to highlight that this flexibility is possible only on the

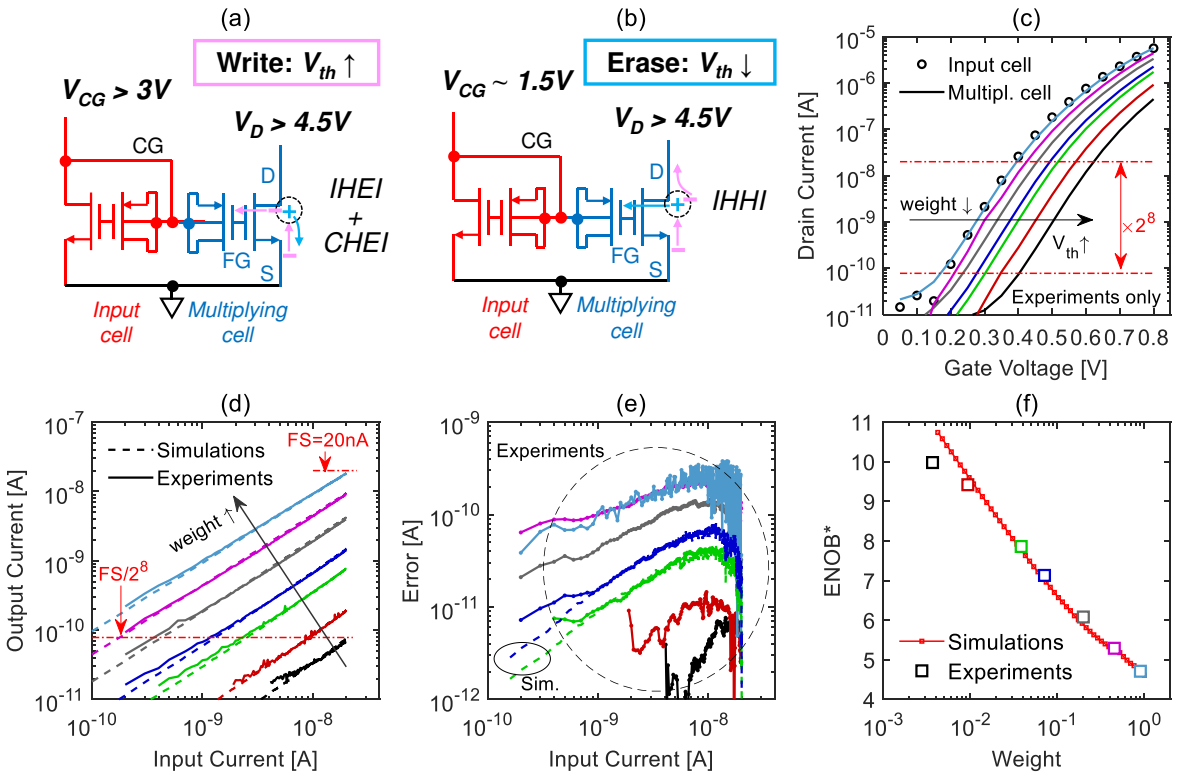

FIGURE 3. **(a) Write conditions to increase the V_{th} by means of impact-ionized and channel-hot-electron-injections (IHEI and CHEI, respectively). (b) Erase conditions to decrease the V_{th} through impact-ionized hot-hole-injection (IHHI). (c) Experimental I_D-V_{CG-S} transfer-characteristics of the diode-connected input cell (symbol) and output cell at V_{DS} = 400 mV (lines). Simple symmetric current mirror: (d) experimental and simulated output current as a function of the input current at different weight conditions; (e) experimental and simulated Error = abs(I_{out}-w_{ideal}×I_{in}) as extracted from (d); (f) effective number of bits calculated from the error in (e), where ENOB*=$\log_2$(FS/(2·max(Error))). $W/L_{(NMOS)}$ = 1 µm / 0.5 µm, $A_{(P\text{-}CAP)}/A_{(NMOS)} \approx 49$.**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037017, IEEE Access

multiplying cell, given that in the input cell the CG and the D are short-circuited. This issue is not really critical since the current conversion ratio (i.e. the weight of the CM) is dependent on the V_{th} difference between the input and output cell, thus a possible charge in the input cell FG can be compensated by offsetting the charge to be added in the multiplying one.

Measurements for a typical cell are shown in Fig.3(c), on both the input and multiplying cell. The I_D-V_{GS} (the gate is the CG, since the FG is not accessible) transfer characteristics were measured. The input-cell has been measured with the FG discharged (symbols), while the multiplying-cell has been characterized at different stored charge conditions (lines). A possible threshold voltage shift ΔV_{th} larger 500 mV has been verified, although few hundred of mV are enough to enable a sufficient conversion factor considering an average inverse subthreshold slope of 90 mV/dec in the current range upper-limited by 20 nA (e.g. ΔVth ~215 mV for a weight of 256^{-1}, i.e. 8 equivalent bits).

For the same weights as programmed in Fig.3(c), the CM has been tested by providing an input current swept in the range 0.2 nA ↔ 20 nA. The resulting output current is shown in Fig.3(d) and post-processed to calculate the error and the corresponding ENOB in Fig.3(e) and (f), respectively. Similar data have been extracted from transient noise simulations performed with UMC 0.18 µm PDK models. The matching between theoretical data and experiments is quite good. Since gate current is not implemented in the transistor models, in the simulations we have used an ideal pulsed current source to inject the needed charge in the FG.

Finally, Fig.4 demonstrates the operation of a 2x1 CM-VMM, implemented with two separated input cells driving two multiplying cells whose weight is independently set to various conditions and whose currents are summed, by implementing the $I_{OUT} = w_A \times I_{in,A} + w_B \times I_{in,B}$ operation.

V. OPTIMIZATION OF FG CURRENT MIRRORS

After the demonstration of an experimental proof-of-concept of analog programmable CM multiplier, there is the need to better understand how to optimize the design of the CM in order to meet a desired precision specification. We have selected an ENOB of 6 bit as a reference specification for the remainder of this study, considering it a good trade-off between precision and cost of the VMM function (in terms of silicon area and power consumption). In section VI, by considering a simple DNN case study trained with the MNIST database, we have verified that with 6 bits the inference accuracy loss is almost negligible compared to higher resolution. However, the choice of a 6-bit ENOB does not affect the generality of our analysis.

Input current FS, transistor sizing and CM topology are design knobs that determine the final performance of the VMM. Concerning the topology, we have already suggested the possible improvement provided by a symmetric CM. In addition, feedback can be also exploited in order to improve

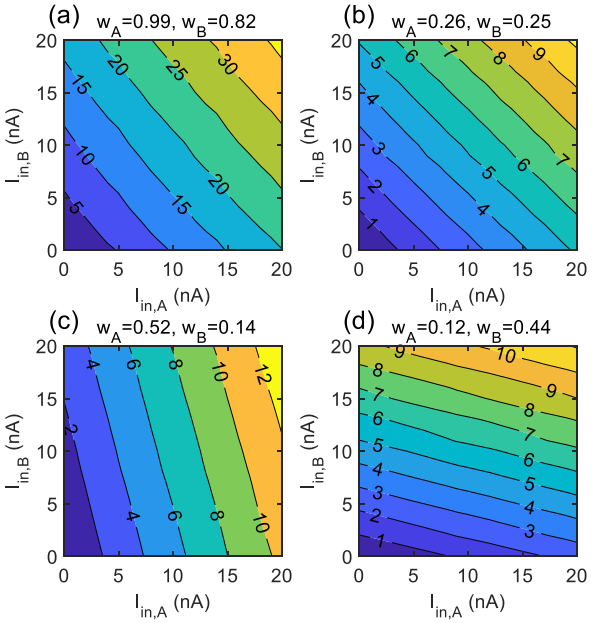

FIGURE 4. 2x1 VMM transfer-function contour plot ($I_{OUT} = w_A \times I_{in,A} + w_B \times I_{in,B}$) for different w_A and w_B conditions: (a) w_A = 0.99, w_B = 0.82; (b) w_A = 0.26, w_B = 0.25; (c) w_A = 0.52, w_B = 0.14; (d) w_A = 0.12, w_B = 0.44. Simple symmetric CM with: $W/L_{(nMOS)}$ = 1 µm / 0.4 µm, $A_{(P\text{-}CAP)}/A_{(nMOS)} \approx 49$.

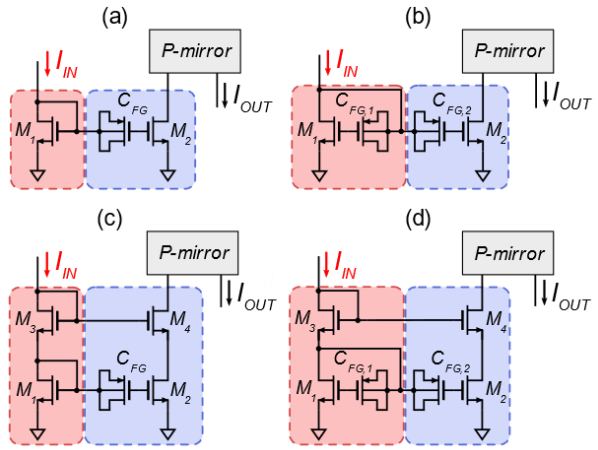

FIGURE 5. Schematics of (a) Asymmetric and (b) Symmetric Simple Current Mirror (ASCM and SSCM, respectively), and of (c) Asymmetric and (d) Symmetric Cascode Current Mirror (ACCM and SCCM, respectively).

the linearity of the analog multiplier. For instance, a cascode CM topology relies on two additional transistors to regulate the V_{DS} of the multiplying cell, by forcing it to follow the one of the input cell.

The four topology options reported in Fig.5 have been considered, consisting in the asymmetric and symmetric versions of simple CMs ((a) ASCM and (b) SSCM, respectively) and of cascode CMs ((c) ACCM and (d) SCCM, respectively). Symmetric and cascode solutions require additional transistors to be implemented. For instance, in a fixed M×N VMM, there will be M additional pCAPs in the input cell array for a symmetric solution with respect to the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037017, IEEE Access

asymmetric one, or M×(N+1) additional nMOS transistors for a cascode CM topology with respect to the simple one. It is important to highlight that it is not obvious that the additional transistors required by more complicated topologies will result in a larger are occupation, if we consider that solutions with a reduced number of transistors will likely require a different sizing of the cell in order to compensate for the reduced linearity performance (for instance a much longer channel transistor L).

A detailed discussion on the linearity (THD), on the noise immunity (SNR), and on the resulting ENOB trends as a function of input current full-scale, supply-voltage V_{DD}, and transistor sizing, as well as a suggested design flow to properly set the W and L sizes of CM transistors, can be found in the Appendix B.

In Fig.6(a) and (c), THD and SNR were extracted at the input current FS I_{MAX} value of 5 nA, V_{DD} = 1.5 V, W/L = 1 µm/2 µm, for different pCAP/nMOS coupling ratios, for both symmetric and asymmetric versions of both simple (SSCM and ASCM) and cascode (SCCM and ACCM) topologies. Symmetric versions show much better linearity for smaller pCAP/nMOS ratio compared to the asymmetric counterpart. In addition, SNR depicted in Fig.6(c) is almost constant for the symmetric solution (~43 dB) down to the minimum considered point of pCAP/nMOS area ratio, while it shows a sudden degradation with reducing pCAP/nMOS ratio for the asymmetric options. From Fig.6(a) we have extracted the minimum pCAP/nMOS ratio (with a margin) which features a THD value of ~ -40 dB for each topology: 49 for ASCM, 36 for ACCM, 25 for SSCM, and 9 for SCCM. Starting from these 4 conditions, we have plotted in Fig.6(b) and (d) the THD and SNR degradation with L scaling. Curves depicted in this plot have been obtained at fixed normalized input current (with respect to the width-to-length ratio, i.e. $I_{norm} = I \times L/W$), basically meaning that when the L is halved the corresponding current is doubled, so that the transistor working point is maintained in similar sub-threshold operating condition (and similar linearity in case of long channel devices). As regards the THD trends, both asymmetric options require a longer channel device compared to the respective symmetric counterparts, despite a much larger pCAP/nMOS ratio initially selected. Then, the area advantage of using a symmetric solution is twofold (i.e. smaller pCAP/nMOS ratio and shorter L), although a pCAP is needed also in the input cell. In addition, if we focus on the symmetric options, it can be observed that SSCM features a small degree of linearity degradation at extremely short length, while the SCCM features a THD value which is optimum at L_{MIN}. This result is attributed to the intrinsic feedback property of the cascode topology, whose action in enforcing similar V_{DS} to the input and multiplying cell transistors results in an effective workaround for reduced output resistance of short-channel devices. As regards the SNR shown in Fig.6(d), a similar behavior is observed for all the configurations, with SNR degrading as the L is reduced. However, one should note that

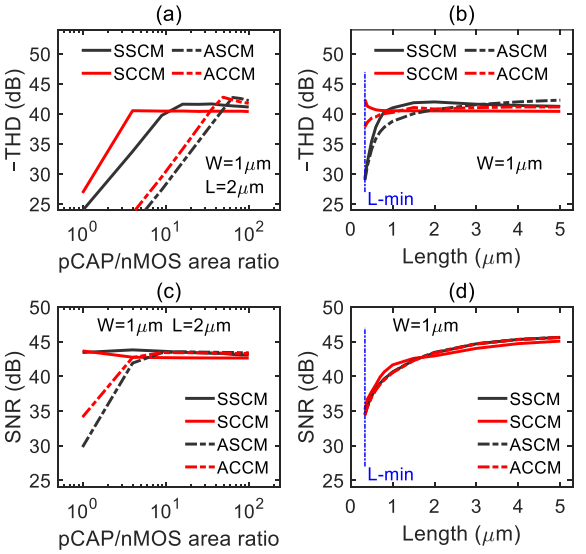

FIGURE 6. Linearity (a) -THD and (c) SNR degradation as a function of pCAP/nMOS coupling ratio scaling reported for symmetric (solid lines) and asymmetric (dashed lines), simple (black) and cascode (gray) current mirrors. Linearity (b) -THD and (d) SNR degradation as a function of the transistor length scaling, reported for different pCAP/nMOS ratio for each implementation in order to have similar THD values (~-41 dB) at L = 2µm. I_{MAX}×L = 10 nA×µm. Area ratios (pCAP/nMOS): simple symmetric (25), simple asymmetric (49), cascode symmetric (9), cascode asymmetric (36).

TABLE I. Comparison of proposed CM topologies

Type	W (µm)	L (µm)	pCAP/ nMOS Ratio	Single Mult.Cell Gate Area (µm²)	Single Mult. Cell LayoutArea (µm²)	100×10 VMM Gate Area (mm²)	100×10 VMM Layout Area (mm²)
ASCM	1	2	49	100	280	1.001	2.814
SSCM	1	1	25	26	136	0.263	1.383
ACCM	1	1.5	36	57	199	0.571	2.005
SCCM	**1**	**0.8**	**9**	**8.8**	**85.5**	**0.089**	**0.868**

SNR can be independently adjusted by proportionally increasing the transistor width and the input operating current (i.e. at fixed I_{MAX}/W) without impacting the THD (see related discussion in Appendix B).

In Table I, the final transistor sizing and occupied areas of each topology, independently designed in order to meet a 6-bit ENOB specification (i.e. SNR & –THD > 40 dB, according to Fig.2) are listed. Asymmetric multiplying cells occupy from ~3.8× to ~6.5× more gate area compared to the one of symmetric multipliers. In particular, SCCM is the best solution in terms of ENOB per unit area (with a single multiplying cell gate area equal to 33.8% and 15.4% the ones of SSCM and ACCM, respectively), due to the smallest required coupling ratio and transistor length needed to reach the ENOB target, despite the fact that such topology needs additional transistors compared to the simple CM. In case of a 100×10 VMM (i.e. one column array of 100 input cells, a 100×10 multiplying cell matrix, and 10 P-mirror adders) the advantage of SCCM persists, with an overall gate area equal to 33.9% and 15.6% the ones of SSCM and ACCM, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037017, IEEE Access

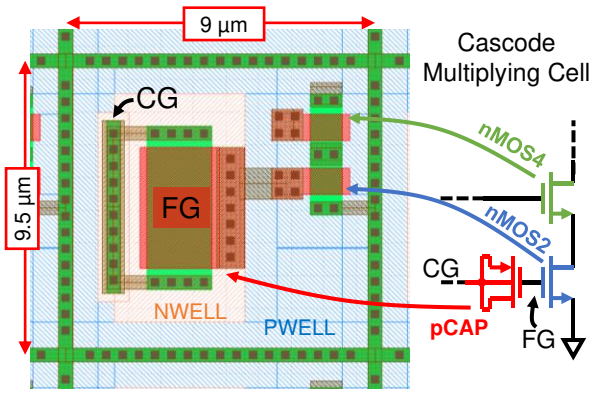

FIGURE 7. Layout of a cascode multiplying cell with W = 1 µm, L = 0.8 µm , pCAP/nMOS area ratio = 9.

The example layout of an SCCM multiplying cell is depicted in Fig.7. We want to clarify that the overall area on the layout is much bigger than the one estimated by using the gate area. This is mainly due to the spacing needed to avoid the turn-on of PNP and NPN parasitic transistors (e.g. n+ diffusions of the nMOS S/D (emitter) / pwell of the nMOS (base) / n-well of the pCAP (collector)). In the reported layout, we have used 2 µm spacing for well-to-well parasitic bipolar paths, and at least 1 µm spacing for diffusion-to-well cases. One should however consider that standard design rules available in the PDK are not intended for such a specific design, then we can speculate that there is some margin to scale the overall layout, e.g. after a specific characterization of any of these paths with dedicated test-structures. Due to this extra area, the overall layout of a multiplying cell of the SCCM is 9.7× larger than the one extracted considering the gate area only (see Table I). However, the advantage of symmetric multipliers is still verified, and the best solution, which is SCCM cell, occupy much less layout area than the SSCM (- 59%) and the ACCM (-132%) multiplying cells.

VI. SYSTEM-LEVEL ASSESSMENT ON ANALOG DNNS

This section is dedicated to a system-level assessment of DNNs, using MATLAB, in order to link the behavior and the FOMs of analog VMMs to the system-level performance of a complete DNN. Two DNNs have been trained and simulated by relying on two different datasets in order to be used as test benches. The grey-scale MNIST [10] dataset has been used to train a purposely designed network ("Net A" in the following) depicted in Fig.8(a), while a subset of classes from ImageNet [12] has been used to train AlexNet [11], as sketched in Fig.8(b). The training has been performed by relying on floating-point data precision.

The designed DNN Net A operates as follow: the input 28×28 pixels gray-scale image is filtered by a convolutional layer with 20 filters on 9×9 kernels. The extracted features are then passed to the activation function, which is a Rectified Linear Unit (ReLU). Then the Maxpooling layer halves the

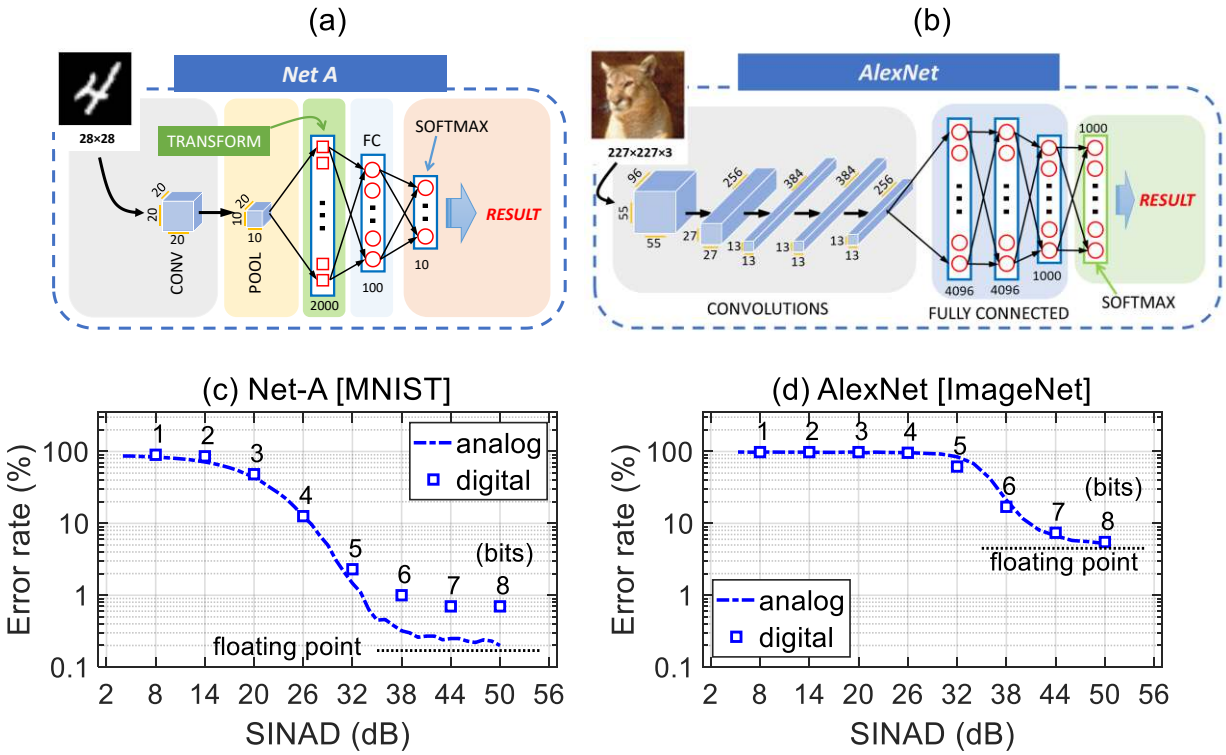

FIGURE 8. (a) Implemented neural network which has been trained for the classification of the MNIST database. (b) AlexNet used to classify a subset of ImageNet. Comparison of the error rate between "digital" and "analog" approximations is shown for Net A classifying MNIST (c) and for AlexNet classifying a subset of 50 classes of ImageNet (d).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037017, IEEE Access

overall number of coefficients by extracting the biggest elements in the 2×2 submatrices. Processed features are passed to the transform level, whose coefficients are trainable, in order to convert the two-dimensional image into a vector. This vector is the input of the fully-connected layer, containing 100 nodes with ReLU. The output layer has 10 nodes and a softmax activation function for the final 10-digits classification.

Details of AlexNet architecture will not be discussed here since they can be easily found in the literature [11].

When the DNNs are used to perform predictions, with floating-point precision, we have found an inference accuracy of 99.8% and 95% for the MNIST and ImageNet datasets, respectively.

Beyond extracting the inference accuracy for the original network, we have artificially derived reduced-precision networks from Net A and AlexNet. Two approximation cases were considered, a "digital" and an "analog" version: a) in the "digital case", floating-point numbers have been replaced with integers with different number of bits; b) in the "analog case", floating-point precision has been maintained, but a white noise has been added to the output of each multiplication based on the assumed SINAD value (and therefore ENOB), according to the following expression:

$$y_{i,j} = w_{i,j}x_{i,j} + \alpha\left(\frac{FS}{2}\right) \times 10^{\frac{-SINAD}{20}} \qquad (7)$$

where α is a random value with a Gaussian distribution of average zero and standard deviation 1.

The inference error rate of the tested DNNs as a function of the corresponding SINAD and ENOB is reported in Fig.8(c) and (d) for the MNIST and ImageNet cases, respectively. As for the digital case, simulations were run on the complete validation dataset of 2000 images for MNIST and almost 1000 images for ImageNet. Instead, for the analog case, the inference on validation dataset was repeated for 5 times and the mean value of the inference accuracy was extracted.

The similarity between the inference capability of a "digital" and of an "analog" network for similar number of bits and ENOB validates the FOMs used in this study. In addition, it also confirms that 6 equivalent bits represent a reasonable value to provide an almost maximum accuracy for MNIST classification by the Net A, while at least 7 bits would be required in the case of ImageNet tested with AlexNet. As a result, we can conclude that the ENOB which must be targeted when designing an analog VMM is dependent on the specific DNN architecture and dataset applications, as expected.

In order to provide a dependable estimation of the energy efficiency of the designed symmetric cascode current-mirror VMM, featuring 6 equivalent bits, we have extracted a 100×10 weight matrix from a trained fully-connected layer of Net A. The estimation was performed by assuming to operate the VMM at V_{DD} = 1.5 V, I_{MAX} = 12.5 nA and to perform 1990 elementary operations in parallel (100 multiplications and 99 summations per each of the 10 columns of the VMM) in a T_{op} close to 100 µs (with a resulting throughput of 19.9 MOPs/s). The energy has been measured for 100 different input vectors, resulting in an average energy efficiency of 26.4 TOPs/J.

Finally, in Table II we have benchmarked our proposal against state-of-art analog VMMs, by selecting the analog VMMs executing arithmetic operation in either current mode or time domain, implemented with memristors [22], embedded FG arrays [27], [30], [31], [36], or single poly FG memories [18], [32], [33]. Both gate and layout areas of a VMM cell, as well as the energy efficiency, are compared to the other design solutions. A single VMM multiplying cell occupies a total gate area of 8.8 µm^2, while the estimated layout area is 85.5 µm^2. Although other single-poly FG solutions are implemented with a more scaled 130 nm technology, the area of our VMM cell is almost one order of magnitude smaller than other proposals based on a similar

TABLE II. Comparison with state-of-art CMOS VMM solutions

Reference	[18]	[22]	[27]	[30],[36]	[31]	[32]	[33]	**This work**
Approach	CM	TD	CM	CM	TD	CM	CM	**CM**
Tech. Node	180 nm	55 nm	55 nm	180 nm	55 nm	130 nm	130 nm	**180nm**
Mem. Type	Digital	1T1R	Embedded NOR	Embedded NOR	Embedded NOR	Single-poly FG	Single-poly FG	**Single-poly FG**
ENOB (bit)	4	4	2	~5	6	7	8	**6**
Single cell gate area (µm^2)	8.1	~3	N/A	N/A	N/A	N/A	120	**8.8**
Single cell layout area (µm^2)	N/A	N/A	0.33	18.5	4.33	792	N/A	**85.5**
EE (TOPs/J)	~8	1305	N/A	5.7	85	1	6.32	**26.4**
Results	Meas.	Sim.	Meas.	Meas.	Meas.	Meas.	Meas.	**Sim./Meas.**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037017, IEEE Access

process technology. The 6-bit ENOB precision is lower than other single-poly multipliers, but similar precision can be matched with a trimming of the design. Compared to the double poly embedded FG array based multipliers, our solution is much bigger, but it has to be considered that the counterpart can rely on the advantages of the double poly and of the more scaled technology node (55 nm). On the other hand, one should note that with double poly technologies it is not possible to modify the geometry of a single cell, thus the optimization of transistor size aiming at increasing the accuracy of the cell is not feasible. Another weakness is that the CMOS double poly process is much more expensive than the single poly one. As regards to the energy efficiency, our multiplier reaches 26.4 TOPs/J, which is better than all the other single-poly VMM counterparts, but worse than the one proposed in [31] (55 nm embedded NOR solution) and the one based on memristors in [22] (only simulations, no experimental data are provided).

VII. CONCLUSION

We have demonstrated an in-memory analog VMM based on current mirrors realized in a commercial 180 nm CMOS technology platform with experiments, circuit-level and system-level simulations. Single-poly floating-gate memory cells provide the possibility to implement the in-memory computing approach. FG cell programming/erasing methods and storing capability have been validated by experimental measurements showing the possibility to set a single poly FG current mirror with a current scaling factor corresponding to more than 256 levels (i.e. >8-bit). Measurements on a symmetric simple current-mirror multiplier resulted to be well matched to circuit level simulations. With the validated simulation deck, a design optimization has been performed for four current mirror topologies, by relying on a proposed design flow targeting a specific precision. It has been demonstrated that complex current mirrors such as the cascode topology feature a better trade-off between ENOB and area occupancy than the simpler version implemented with a reduced number of transistors. Furthermore, the electrostatic symmetry produced by placing a pCAP in both the input and multiplying cell allows to further reduce the area, allowing the current mirror multiplier to reach the accuracy specifications with much smaller transistor sizes. Both MNIST and ImageNet databases have been used as representative examples to train two DNNs, which are a purposely developed DNN and the well-known AlexNet, respectively. System-level simulations were performed for both cases, and the inference accuracy has been extracted as a function of the assumed ENOB. We have found that a precision of 6 equivalent bits allows an almost maximum accuracy in classifying images from the MNIST database, while ImageNet requires at least 7 bits. Our CM-VMM reach an energy efficiency of 26.4 TOPs/J, that is very promising with respect to the state-of-the art of experimentally tested analog neuromorphic circuits considering the relatively high precision (ENOB = 6) and small area occupation of the proposed VMM.

APPENDIX

APPENDIX A: Definitions of THD, SNR and SINAD

Definitions of THD, SNR and SINAD for current waveforms are given below.

$$THD = -10 \cdot log_{10}(I^2_{SIGNAL}/I^2_{DISTORTION}) \quad (8)$$

$$SNR = 10 \cdot log_{10}(I^2_{SIGNAL}/I^2_{NOISE}) \quad (9)$$

$$SINAD = 10 \cdot log_{10}\left(\frac{I^2_{SIGNAL}}{I^2_{NOISE} + I^2_{DISTORTION}}\right) \quad (10)$$

I_{SIGNAL}, $I_{DISTORTION}$ and I_{NOISE} are the RMS values of signal, distortion and noise contributions, extracted by means of the Fourier Transform of the output waveform in response to a clean sine waveform provided as an input.

APPENDIX B: THD and SNR trends as a function of design parameters

Details of THD and SNR trends as a function of basic design parameters are discussed in this Appendix. In this analysis a simple and idealized CM is considered, where both the input and the multiplying cell are implemented by a nMOS transistor only; the magnification of the current conversion ratio (i.e. the weight) is not accounted with the realistic FG but by relaying on an ideal weight represented by a DC voltage generator simulated in series to the gate of the multiplying transistor. By analyzing simulated trends, we are able to suggest a consistent design flow which can be applicable also to more complicated current mirror topologies.

A current sine waveform with a peak-to-peak amplitude equal to the selected $I_{in,MAX}$ is applied to the input cell. The THD and SNR FOMs are computed by post-processing the waveform of the corresponding output for a variable weight. In this discussion it will be recurrent the normalization of the operating current with respect to the width-to-length ratio (i.e. $I_{NORM.} = I_{in,MAX} \times L/W$), so that the transistor working point is maintained in similar sub-threshold region conditions (and similar linearity) when the transistor aspect ratio is changed.

In Fig.9, simulations were carried out by varying electrical parameters such as the maximum amplitude of the input signal $I_{in,MAX} \times L$, ((a) and (e)), the supply voltage V_{DD} ((b) and (f)), the transistor L ((c) and (g)) and the transistor W ((d) and (h)).

Two different trends can be observed in Fig.9(a) and (e): first, there is a trade-off between THD and SNR in terms of $I_{in,MAX}$. If the current is increased, the THD curves worsen and, at currents higher than ~100 nA, their shape and the related slopes change as transistors are on the edge between subthreshold and inversion regions. According to this trend, it would be recommended to operate the transistors in deep sub-threshold to increase linearity, although in case of short channel devices, e.g. L = 0.5 µm, the benefit of reducing the current is less pronounced considering that the short-channel

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037017, IEEE Access

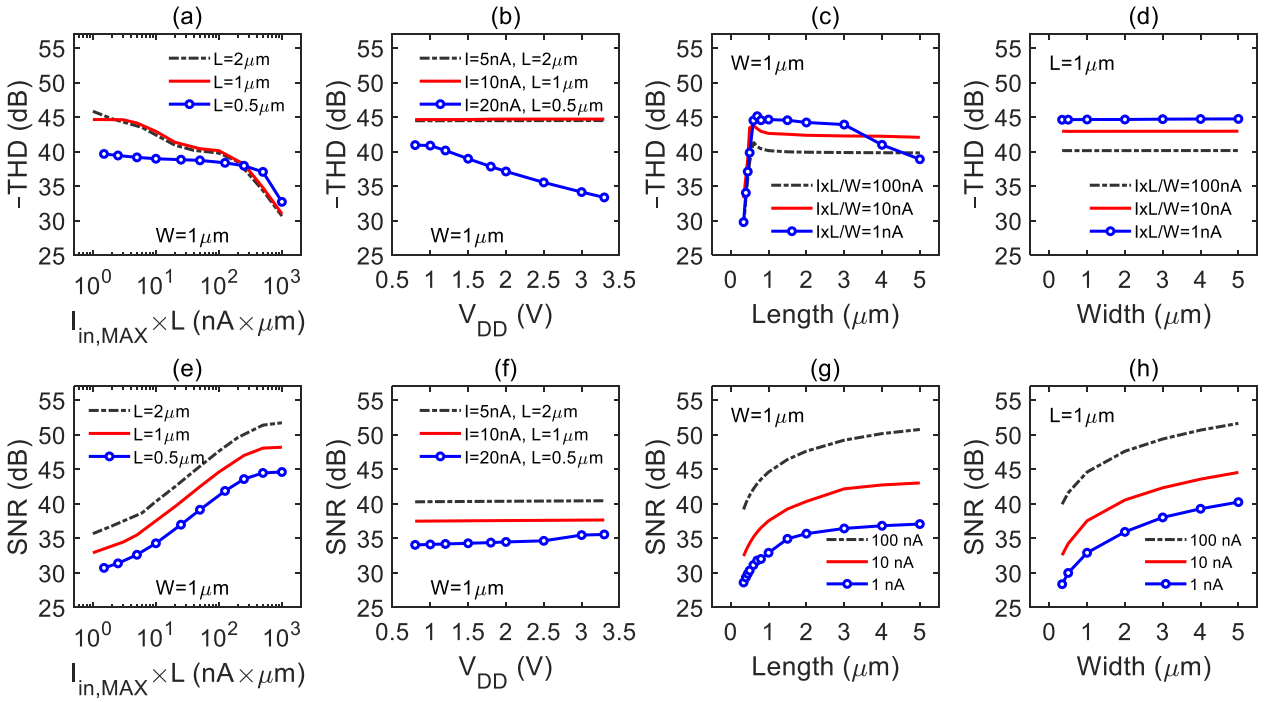

FIGURE 9. **Simple n-type current mirror simulated with an ideal threshold voltage offset to mimic the weight (i.e. FG cell with ideal coupling ratio). THD and SNR as a function of the normalized input current I_{MAX}×L at V_{DD} = 1.5 V ((a) and (e), respectively), of the supply voltage V_{DD} at constant I_{MAX}×L = 10 nA×μm ((b) and (f), respectively), of the nMOS transistor length ((c) and (g), respectively) and of the nMOS transistor width ((d) and (h), respectively). Nominal parameters (unless specified differently in the plots): W_N = 1 μm, W_P = 4×W_N, L_P = L_N = 1 μm. V_{DD} = 1.5V, $I_{in,MAX}$×L/W = 10 nA.**

effects (SCEs) affect the THD. As an opposite trend, increasing $I_{in,MAX}$ is instead beneficial from the SNR point of view, as shown in Fig.9(e). In addition, even at the same biasing condition (constant $I_{in,MAX}$×L), longer devices feature higher SNR, as highlighted by the three different curves.

According to Fig.9(b) and Fig.9(f), once the bias point is set by the operating current, THD and SNR values are typically not affected by the supply voltage variation. A V_{DD} dependence can be observed only for short channel devices featuring a worsening of the THD for an increasing V_{DD}, and in those cases a low supply voltage should be preferred in order to save power consumption. We choose a value of V_{DD} = 1.5 V for the remainder of the analysis.

The analysis based on geometrical parameters, L and W, were still performed for constant I_{MAX}×L/W, where W = 1 μm when L is varied, and L = 1 μm when W is varied. In Fig.9(c) there is a very small length range where the linearity increases by moving toward longer devices because of the reduction of SCEs. Furthermore, the curves taken at 10 and 100 nA×μm/μm show a flat THD region in the longer cases. Here the simulated length is sufficient to screen any impact of SCEs, and the similar operating points in subthreshold (guaranteed by the same I_{MAX}×L) result in similar values for linearity. However, for very low current levels (i.e., 1 nA×μm/μm), after the initial rise, a flat region extends only for few μm, i.e. up to ~3 μm, considering that beyond this value THD starts to decrease for increasing length. This is due to the fact that, for long transistors, a normalized current of

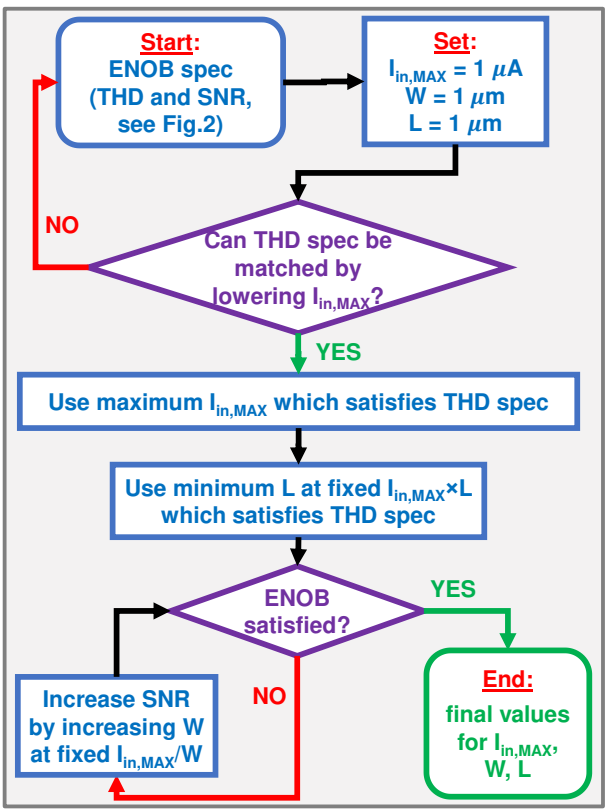

FIGURE 10. **Proposed design flow.**

1 nA×µm/µm corresponds to a very small unnormalized current (e.g. 200 pA for L = 5 µm), and at this current value it corresponds a V_{DS} lower than $4V_T$ which does not guarantee a proper transistor saturation. However, if we focus on the 1 ~ 3 µm range, lower I_{MAX}×L values always corresponds to a better linearity, in agreement with Fig.9(a). SNR in Fig.9(g) has a strong increase with increasing L for short channel devices, although it tends to saturate for longer values. Finally, when varying W for fixed normalized currents, linearity is practically independent (Fig.9(d)), while SNR always increases for an increasing width (Fig.9(h)), with an almost linear dependence on the square root of the width.

By taking into account all the plots depicted in Fig.9, as a conclusion we can assert that there is a certain region in the design space where THD and SNR can be independently set, and a possible design flow with the target to reach a given ENOB can be suggested, as detailed in Fig.10. For instance, by using one of the curves depicted in Fig.9(a) (e.g. with at least L = 1 µm to screen SCEs), one could decrease $I_{in,MAX}$×L down to the value which guarantee the specification on the linearity (i.e. desired THD). Then, by referring to Fig.9(c), the length can be scaled down to the value that does not produce a linearity degradation (still for fixed $I_{in,MAX}$×L). Both design choices aim at linearity by trading off against a SNR degradation (see Fig.9(e) and Fig.9(g)). However, the specification on the SNR can be reached by a final trimming of W (according to Fig.9(h)) which can be modified – for fixed normalized current – without impacting the linearity obtained by previous design choices (see Fig.9(d)).

REFERENCES

[1] R. Sarpeshkar, "Analog Versus Digital: Extrapolating from Electronics to Neurobiology," in Neural Computation, vol. 10, no. 7, pp. 1601-1638, 1 Oct. 1998, doi: 10.1162/089976698300017052..

[2] K. H. Lee and N. Verma, "A Low-Power Processor With Configurable Embedded Machine-Learning Accelerators for High-Order and Adaptive Analysis of Medical-Sensor Signals," in IEEE Journal of Solid-State Circuits, vol. 48, no. 7, pp. 1625-1637, July 2013, doi: 10.1109/JSSC.2013.2253226

[3] V. Sze, Y. Chen, J. Emer, A. Suleiman and Z. Zhang, "Hardware for machine learning: Challenges and opportunities," 2017 IEEE Custom Integrated Circuits Conference (CICC), Austin, TX, 2017, pp. 1-8, doi: 10.1109/CICC.2017.7993626.

[4] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, 2014, pp. 10-14, doi: 10.1109/ISSCC.2014.6757323.

[5] L. Cavigelli, M. Magno and L. Benini, "Accelerating real-time embedded scene labeling with convolutional networks," 2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC), San Francisco, CA, 2015, pp. 1-6, doi: 10.1145/2744769.2744788.

[6] Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2017, doi: 10.1109/JSSC.2016.2616357.

[7] D. Han, J. Lee, J. Lee, and H. J. Yoo, "A Low-Power Deep Neural Network Online Learning Processor for Real-Time Object Tracking Application," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 66, no. 5, pp. 1794–1804, 2019, doi: 10.1109/TCSI.2018.2880363.

[8] 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, 2019, doi: 10.1109/JPROC.2018.2871057.

[9] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *J. Mach. Learn. Res.*, vol. 18, pp. 6869–6898, 2017.

[10] The MNIST database of handwritten digits. [Online] Available: http://yann.lecun.com/exdb/mnist/

[11] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, vol. 2, pp. 1097-1105, 2012.

[12] The ImageNet database. [Online] Available: http://image-net.org/

[13] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML'15) v. 37, pp. 1737–1746, 2015.

[14] R. Andri, L. Cavigelli, D. Rossi, and L. Benini, "YodaNN: An ultra-low power convolutional neural network accelerator based on binary weights," 2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Pittsburgh, PA, 2016, pp. 236-241, doi: 10.1109/ISVLSI.2016.111.

[15] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1," 2016. [Online]. Available: https://arxiv.org/abs/1602.02830.

[16] K. Ando et al., "BRein Memory: A Single-Chip Binary/Ternary Reconfigurable in-Memory Deep Neural Network Accelerator Achieving 1.4 TOPS at 0.6 W," in IEEE Journal of Solid-State Circuits, vol. 53, no. 4, pp. 983-994, April 2018, doi: 10.1109/JSSC.2017.2778702.

[17] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An Energy-Efficient SRAM With In-Memory Dot-Product Computation for Low-Power Convolutional Neural Networks," in IEEE Journal of Solid-State Circuits, vol. 54, no. 1, pp. 217-230, Jan. 2019, doi: 10.1109/JSSC.2018.2880918.

[18] J. Binas, D. Neil, G. Indiveri, S.-C. Liu, and M. Pfeiffer, "Precise neural network computation with imprecise analog devices," pp. 1–22, 2020. . [Online]. Available: https://arxiv.org/abs/1606.07786.

[19] H. -. P. Wong et al., "Metal–Oxide RRAM," in Proceedings of the IEEE, vol. 100, no. 6, pp. 1951-1970, June 2012, doi: 10.1109/JPROC.2012.2190369.

[20] M. Kund et al., "Conductive bridging RAM (CBRAM): an emerging non-volatile memory technology scalable to sub 20nm," IEEE International Electron Devices Meeting, 2005. IEDM Technical Digest., Washington, DC, 2005, pp. 754-757, doi: 10.1109/IEDM.2005.1609463.

[21] D. Apalkov *et al.*, "Spin-transfer torque magnetic random access memory (STT-MRAM)," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 9, no. 2, pp. 1–35, 2013, doi: 10.1145/2463585.2463589.

[22] M. Bavandpour, S. Sahay, M. R. Mahmoodi and D. Strukov, "Efficient Mixed-Signal Neurocomputing Via Successive Integration and Rescaling," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 28, no. 3, pp. 823-827, March 2020, doi: 10.1109/TVLSI.2019.2946516.

[23] S. Duan, X. Hu, Z. Dong, L. Wang and P. Mazumder, "Memristor-Based Cellular Nonlinear/Neural Network: Design, Analysis, and Applications," in IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 6, pp. 1202-1213, June 2015, doi: 10.1109/TNNLS.2014.2334701.

[24] S. Kim et al., "NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning," 2015 IEEE International Electron Devices Meeting (IEDM), Washington, DC, 2015, pp. 17.1.1-17.1.4, doi: 10.1109/IEDM.2015.7409716.

[25] M. Suri and V. Parmar, "Exploiting Intrinsic Variability of Filamentary Resistive Memory for Extreme Learning Machine Architectures," in IEEE Transactions on Nanotechnology, vol. 14, no. 6, pp. 963-968, Nov. 2015, doi: 10.1109/TNANO.2015.2441112.

[26] E. Vianello et al., "Resistive Memories for Spike-Based Neuromorphic Circuits," 2017 IEEE International Memory Workshop (IMW), Monterey, CA, 2017, pp. 1-6, doi: 10.1109/IMW.2017.7939100.

[27] X. Guo et al., "Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells," 2017 IEEE

Custom Integrated Circuits Conference (CICC), Austin, TX, 2017, pp. 1-4, doi: 10.1109/CICC.2017.7993628.

[28] G. Malavena, A. S. Spinelli and C. M. Compagnoni, "Implementing Spike-Timing-Dependent Plasticity and Unsupervised Learning in a Mainstream NOR Flash Memory Array," 2018 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2018, pp. 2.3.1-2.3.4, doi: 10.1109/IEDM.2018.8614561.

[29] K. Parat and A. Goda, "Scaling Trends in NAND Flash," 2018 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2018, pp. 2.1.1-2.1.4, doi: 10.1109/IEDM.2018.8614694.

[30] 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," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 10, pp. 4782-4790, Oct. 2018, doi: 10.1109/TNNLS.2017.2778940.

[31] M. Bavandpour, , M. R. Mahmoodi and D. B. Strukov, "Energy-Efficient Time-Domain Vector-by-Matrix Multiplier for Neurocomputing and Beyond," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 66, no. 9, pp. 1512-1516, Sept. 2019, doi: 10.1109/TCSII.2019.2891688.

[32] 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," in IEEE Journal of Solid-State Circuits, vol. 50, no. 1, pp. 270-281, Jan. 2015, doi: 10.1109/JSSC.2014.2356197.

[33] M. Judy et al., "A Digitally Interfaced Analog Correlation Filter System for Object Tracking Applications," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 65, no. 9, pp. 2764-2773, Sept. 2018, doi: 10.1109/TCSI.2018.2819962.

[34] M. Yamaguchi, G. Iwamoto, H. Tamukoh, and T. Morie, "An Energy-efficient Time-domain Analog VLSI Neural Network Processor Based on a Pulse-width Modulation Approach," 2019. [Online]. Available: https://arxiv.org/abs/1902.07707.

[35] M. M. Hasan and J. Holleman, "Implementation of Linear Discriminant Classifier in 130nm Silicon Process," in 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 2018, vol. 2018-May, no. 3, pp. 1–5, doi: 10.1109/ISCAS.2018.8351829.

[36] X. Guo et al., "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2017, pp. 6.5.1-6.5.4, doi: 10.1109/IEDM.2017.8268341.

[37] Y. Ji, L. Nan and K. Mouthaan, "Analysis of the drain thermal noise for deep submicron MOSFETs," 2009 Asia Pacific Microwave Conference, Singapore, 2009, pp. 1659-1662, doi: 10.1109/APMC.2009.5384327.

[38] W. Kester et al, "Chapter 2: Fundamentals of Sampled Data Systems," in "The Data Conversion Handbook," Newnes, 2005. [Online]. Available: https://www.analog.com/media/en/training-seminars/design-handbooks/Data-Conversion-Handbook/Chapter2.pdf

[39] Y. Da Wu, K. Cheng, C. Lu and H. Chen, "Embedded Analog Nonvolatile Memory With Bidirectional and Linear Programmability," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 59, no. 2, pp. 88-92, Feb. 2012, doi: 10.1109/TCSII.2012.2184371.