IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

Received 12 August 2021; revised 6 September 2021 and 6 October 2021; accepted 15 October 2021.
Date of publication 19 October 2021; date of current version 18 November 2021.

Digital Object Identifier 10.1109/JXCDC.2021.3121534

Assessment of Two-Dimensional Materials-Based Technology for Analog Neural Networks

MAKSYM PALIY[1], SEBASTIANO STRANGIO[1], (Member, IEEE), PIERO RUIU[1], AND GIUSEPPE IANNACCONE[1,2], (Fellow, IEEE)

[1]Dipartimento di Ingegneria dell'Informazione, Università di Pisa, 56126 Pisa, Italy
[2]Quantavis s.r.l., 56126 Pisa, Italy

CORRESPONDING AUTHOR: M. PALIY (maksym.paliy@ing.unipi.it)

This work was supported in part by the H2020 QUEFORMAL under Grant 829035 and in part by the Italian Ministry of University and Research through the CrossLab Departments of Excellence Grant.

This article has supplementary downloadable material available at https://doi.org/10.1109/JXCDC.2021.3121534, provided by the authors.

ABSTRACT Embedding advanced cognitive capabilities in battery-constrained edge devices requires specialized hardware with new circuit architecture and—in the medium/long term—new device technology. We evaluate the potential of recently investigated devices based on 2-D materials for the realization of analog deep neural networks (DNNs), by comparing the performance of neural networks based on the same circuit architecture using three different device technologies for transistors and analog memories. As a reference result, it is included in the comparison also an implementation on a standard 0.18 μm CMOS technology. Our architecture of choice makes use of current-mode analog vector-matrix multipliers (VMMs) based on programmable current mirrors (CMs) consisting of transistors and floating-gate non-volatile memories. We consider experimentally demonstrated transistors and memories based on a monolayer molybdenum disulfide channel and ideal devices based on heterostructures of multilayer–monolayer $PtSe_2$. Following a consistent methodology for device-circuit co-design and optimization, we estimate the layout area, energy efficiency, and throughput as a function of the equivalent number of bits (ENOB), which is strictly correlated with classification accuracy. System-level tradeoffs are apparent: for a small ENOB experimental MoS_2 floating-gate devices are already very promising; in our comparison, a larger ENOB (7 bits) is only achieved with CMOS, signaling the necessity to improve linearity and electrostatics of devices with 2-D materials.

INDEX TERMS 2-D materials, analog neural networks, floating-gate memories, vector-matrix multipliers (VMMs).

I. INTRODUCTION

THE pervasive success of deep learning in artificial intelligence applications [1] is accelerating research efforts toward specialized hardware with optimized computer architecture, circuit design, and even device technology. The main effect is a shift from the general-purpose Von Neumann paradigm to specialized hardware that leverages the properties of deep neural network (DNN) algorithms [2]. Logic in-memory architectures [3] are extremely interesting from this point of view: they consist of many modularized processing elements distributed in space and operating in parallel, implementing the simultaneous operations performed by neurons. In addition, each processing element contains both the logic and part of the memory required for the task, reducing the energy consumption and the delay associated with access to a cache or an external memory [2], [4]. In fact, the main principles of "neuromorphic" silicon circuits date back to the pioneering work at Caltech in late 1990s [5]–[7], but the new wave of interest in integrated circuits for neural networks is much stronger now that there is a potent market pull and is mainly driven by the needs of maximizing data center computing capability at constant power envelope and of bringing cognitive capability to embedded systems [8].

Vector-matrix multipliers (VMMs) are the ubiquitous logic blocks in a DNN, performing the multiplication of a vector of inputs by a matrix of trained features, that is, weights,

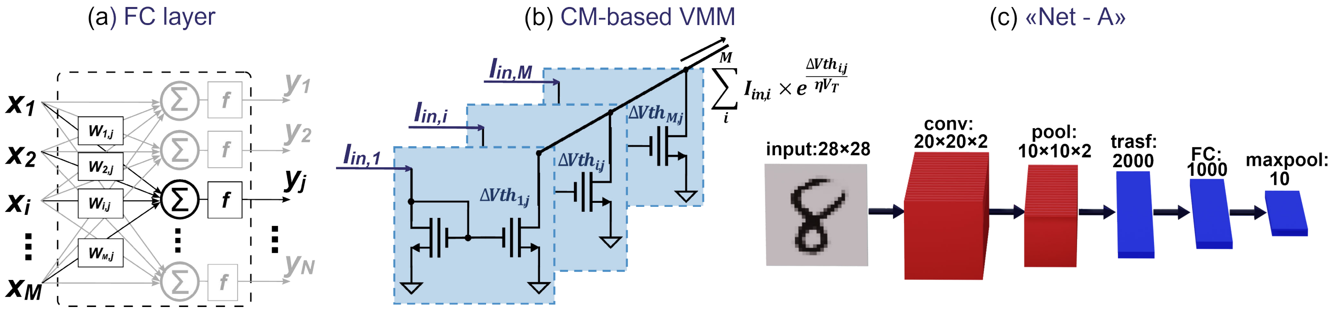

FIGURE 1. (a) Block diagram of a fully connected layer. (b) Simplified circuit of a single VMM node. (c) Architecture of the considered DNN.

as sketched in Fig. 1(a). The VMMs play a crucial role in establishing the performance of a full DNN system, such as the classification accuracy of the network [9], the energy efficiency, and the throughput [10]. It has been demonstrated that the inference operations with a reduced multi-bit precision can reach a comparable classification accuracy to floating-point arithmetics due to the resilience to disturb the learning algorithms [9], [11].

This opens up the possibility to perform computation in the analog domain by exploiting the device physics and circuit properties (e.g., Kirchhoff laws) [9], [10], [12], [13]. It is well known that analog processing blocks are usually affected by circuit non-idealities such as noise and nonlinearity, which limit the effective number of bits of the arithmetic operations. However, there are tradeoffs among area, throughput, energy efficiency, and computation accuracy that must be investigated and that heuristically show that area and energy efficiency can be traded for computation accuracy and throughput.

An in-memory analog VMM has been designed and realized in a commercial 0.18 μm CMOS platform [9] by using programmable current mirrors (CMs) similar to the one in Fig. 1(b). We consider this very same circuit architecture for a comparison of different 2-D device technologies, so that we can use as reference a case for which we have a full range of experimental results.

Transition metal dichalcogenides (TMDCs) are well suited for post-silicon CMOS or for an integration with CMOS technology because: 1) they provide much higher mobility than silicon in the case of ultrathin channel layers required by aggressive scaling and 2) the weak Van der Waals interaction between stacked layers is useful for 3-D integration of transistors [14]–[16]. Some recent experimental results are very promising in view of their use in analog neural networks.

1) A floating gate FET (FGFET) with a monolayer MoS_2 channel has been demonstrated and used to implement reconfigurable logic in the digital domain [17].
2) Monolayer MoS_2 MOSFETs with planar geometry have been used to fabricate fundamental analog circuits such as a two-stage operational amplifier [18].

We also include in the comparison an ideal FET with a channel consisting of a lateral heterostructure (LH) of monolayer and multilayer $PtSe_2$ [19]. Such heterostructure has perfect lattice match, with a monolayer $PtSe_2$ that is a semiconductor with a gap of 1.36 eV and a multilayer $PtSe_2$ that is a metal providing a low contact resistance. The device has been simulated in [19] without defects, with ideal contacts, and with an aggressive 12.8-nm channel length, exhibiting a very high $I_{\mathrm{ON}}/I_{\mathrm{OFF}}$ ratio and an almost ideal subthreshold slope (SS). We use this device as the upper limit of what it could be achievable as fabrication technology improves.

As anticipated above, we assess the potential of the use of these 2-D device technologies in analog neural networks, using as a reference the experimental results obtained with a commercial CMOS technology (single-poly UMC 0.18 μm). Let us stress here that the spirit of the comparison is not to choose among different technologies, because they do not have the same degree of maturity and cost, but rather to understand the different tradeoffs at play and the aspects of each device technology that need to improve the most.

CM-based VMMs with the MoS_2 and the $PtSe_2$-based FETs have been designed and optimized according to the design approach we have proposed in [9] for a standard CMOS. The performance of the designed multipliers has then been tested at the DNN level on a purposely designed network ["Net A" depicted in Fig. 1(c)], trained with the gray-scale MNIST handwritten digit dataset [19]. All the considered device technology options have been evaluated in terms of area occupation, throughput, energy efficiency, and classification accuracy.

The remainder of this article is organized as follows. In Section II, we present the basic operation of the current-mode analog VMM used in our neural network architecture. In Section III, we describe the reference analog neural network used in the evaluation. In Section IV, the main figures of merit are introduced. In Section V, we extract the compact models of MoS_2 and $PtSe_2$ transistors from measurements and simulations, respectively. In Section VI, we explain the design techniques used to optimize the CM-based VMM for the minimum area requirement. The comparison of the considered device technology options is reported in Section VII. Final conclusions are drawn in Section VIII.

II. BASIC OPERATION OF AN IN-MEMORY ANALOG VMM

A fully connected (FC) layer of a DNN is sketched in Fig. 1(a). Each output $y_j\,(j = 1, 2, \ldots, N)$ of a layer is obtained by applying an activation function f to the cumulative sum of the products of each generic input x_i $(i = 1, 2, \ldots, M)$ of the layer times a weight $w_{i,j}$

$$y_j = f\left(\sum_i^M x_i \times w_{i,j}\right). \qquad (1)$$

The sum in (1) can be physically implemented by means of a current-mode VMM, realized with programmable CMs such as the one depicted in Fig. 1(b) [9]. In this case, the $x_i's$ are encoded in the currents $I_{\text{in},i}$ and each $w_{i,j}$ is the current magnification factor of the programmable CMs. The sum is provided as the total output current $I_{\text{out},j}$, obtained by connecting to the same node the output branches of all CMs corresponding to the same j-th output element [9]. Programmable mirrors are realized with three terminal floating-gate (FG) non-volatile memories, and the magnification factor depends on the threshold voltage difference $\Delta V\text{th}_{i,j}$ between the two FG non-volatile memories of the same mirror, programmed by injecting a charge in the floating gate of the CM output device.

All transistors operate in the sub-threshold region allowing to reduce the power consumption and to achieve a range of the weights variation larger than two orders of magnitude [9], [13] considering that the weight can be expressed as

$$w_{i,j} = e^{\frac{\Delta Vth_{i,j}}{\eta V_T}} \qquad (2)$$

where η is the subthreshold ideality factor (typically between 1 and 2) and $V_T = K_B T/q$ is the thermal voltage, in which K_B is Boltzmann's constant, T is the temperature, and q is the elementary charge. We therefore have

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

Weights are determined in the "training" phase of the neural network using a so-called "training dataset," consisting of labeled sample images [21]. The obtained set of weights are then used in the classification phase, also known as "inference." Mismatch between devices can be mitigated during the programming phase, given that this threshold voltage variation can be fully compensated with appropriate tuning of the charge injected in the FGs. This work is mainly focused on the inference phase: the experimental demonstration of programming analog weights is provided in a few papers, either based on CMOS technology [9], [12], [13], [22] or on 2-D-materials [17], using different injection mechanisms.

III. SYSTEM-LEVEL TESTBENCH

The VMM is the most recurrent building block of DNNs, instantiated in all FC and convolutional layers. Therefore, it is reasonable to benchmark different technologies for analog neural networks by comparing the behavior of VMMs.

However, in the case of an analog VMM, the statistical distribution of weights, which depend on the architecture and the training of a particular DNN, can have a quantitative impact on power dissipation and on the speed. For this reason, in the assessment exercise presented here, we will consider a testbench convolutional neural network, "Net-A," shown in Fig. 1(c), where only the analog VMMs are simulated at the circuit level in Cadence Virtuoso, while the full network behavior (both training and inference) are simulated in MATLAB. Net-A consists of the following layers: an input layer, which receives a 28 × 28 pixels grayscale image from the MNIST database [19]; a convolutional layer [23] with 20 9 × 9 filters for feature extraction from images; a pooling layer [23] with 2 × 2 kernels, which halves the overall number of coefficients preventing overfitting; a transform level, which rearranges the 2-D data into a 1-D vector with 2000 elements; a 100-node FC layer and an output layer, composed of ten nodes with softmax activation function for the final 10-digit classification. The rectified linear unit (ReLU) [23] is used as nonlinear activation function for both the convolutional and the FC layers.

The image dataset was used for training and inference in a 8:2 ratio. The training process was performed with a supervised mini-batch method, with each batch composed of 100 images, for a total of 60 epochs.

IV. FIGURES OF MERIT FOR NEURAL NETWORK BENCHMARKING

In this section, we introduce the key figures of merit (FOMs) used for the VMM benchmarking: effective number of bits (ENOB), number of operations, latency time, throughput, energy efficiency, and area occupation.

1) *ENOB:* The ENOB is a measure of the computing accuracy [9]: it depends on the signal-to-noise and distortion ratio (SINAD), according to

$$\text{ENOB} = \frac{\text{SINAD}_{\text{dB}} - 1.76}{6.02}. \qquad (4)$$

The SINAD value depends on the nonlinearity, estimated through the total harmonic distortion (THD), and on the signal-to-noise ratio (SNR), according to

$$10^{-\frac{\text{SINAD}_{\text{dB}}}{10}} = 10^{-\frac{\text{SNR}_{\text{dB}}}{10}} + 10^{\frac{\text{THD}_{\text{dB}}}{10}}. \qquad (5)$$

In order to provide a reference value for the ENOB, one should note that we have previously proved that a VMM with an ENOB of 6-bit can guarantee a 99.7% classification accuracy of the network Net-A, trained with the MNIST database [9], [19]. The required ENOB for a given classification accuracy typically depends on the DNN architecture. For example, the well-known AlexNet [28] requires an ENOB = 7 for a classification accuracy of 92% [9].

2) *Number of operations:* Each multiplication and addition is considered as an elementary arithmetic operation. For an $M \times N$ VMM, which includes M multiply-and-accumulate (MAC) units and N columns, there are a total of M multiplications and $M - 1$ additions per each of the N columns, corresponding to a total number of $(2M - 1) \times N$ elementary operations.

3) *Latency time* (τ_{LAT}) represents the time needed by each processing unit to perform a single arithmetic operation, and it is estimated through the settling time of the output in response to an input step. Considering that all processing elements are arranged in parallel, the worst case latency time of a single multiplier corresponds to the latency time of the whole VMM implementing a layer of the network.
4) *Throughput:* The throughput is the ratio of the number of arithmetic operations performed in parallel to the VMM latency time and is a measure of performance of the VMM expressed in term of operations per second (OPs/s).
5) *Energy efficiency* (EE): It is calculated as the ratio of the number of operations to the total energy consumed by the VMM to perform a vector-matrix multiplication (i.e., the integral of the consumed power over τ_{LAT}), typically measured in tera operations per joule (TOPs/J). The energy is determined using the weights obtained from the training phase for Net-A and it is averaged over many operations, each corresponding to an input vector associated with an image of the MNIST database.
6) *VMM area:* In a multilayer DNN, large VMMs are the dominant elements in terms of area occupation. We have compared both gate and layout area of VMMs realized with the same $M \times N$ size.

V. DEVICES BASED ON 2-D MATERIALS FOR ANALOG DNNs

In this section, we present the considered devices and the calibrated compact models (Berkeley short-channel IGFET model-silicon-on-insulator (BSIM-SOI) [24]) used to perform VMM circuit simulations. Two MoS_2 experimental devices presented in the literature have been considered, as representative of the state of the art for MoS_2 FETs: the FG memory cell (FGFET) presented in [17] and sketched in Fig. 2(a) and (b), and the planar MoS_2 FET presented in [18] and sketched in Fig. 2(c) and (d). In addition, the $PtSe_2$ double-gate LH field-effect transistor (LH-FET) proposed in [19] and represented in Fig. 2(e) and (f), designed and analyzed through multi-scale simulations, has been considered to estimate the maximum achievable performance of ultra-scaled and optimized 2-D material transistors. We have compared VMM implemented by means of these devices with a VMM implementation based on a standard 0.18 μm CMOS technology, typical for the analog circuit design. In order to exploit these devices for analog neural network blocks, we have first calibrated the compact model with the available I–V characteristics by considering the physical properties of each reference device; then, device geometry optimization has been performed to adapt each device to the specific application and circuit.

Concerning 2-D-FET devices, there are no standard compact models for circuit simulations including both static and dynamic behavior. This drawback makes the design and predictive simulation of complex circuits ineffective. Some discrete models based on lookup table (LUTs) [25] and some semi-analytical models [26], [27] have been proposed for DC simulations but without accounting for dynamic signal

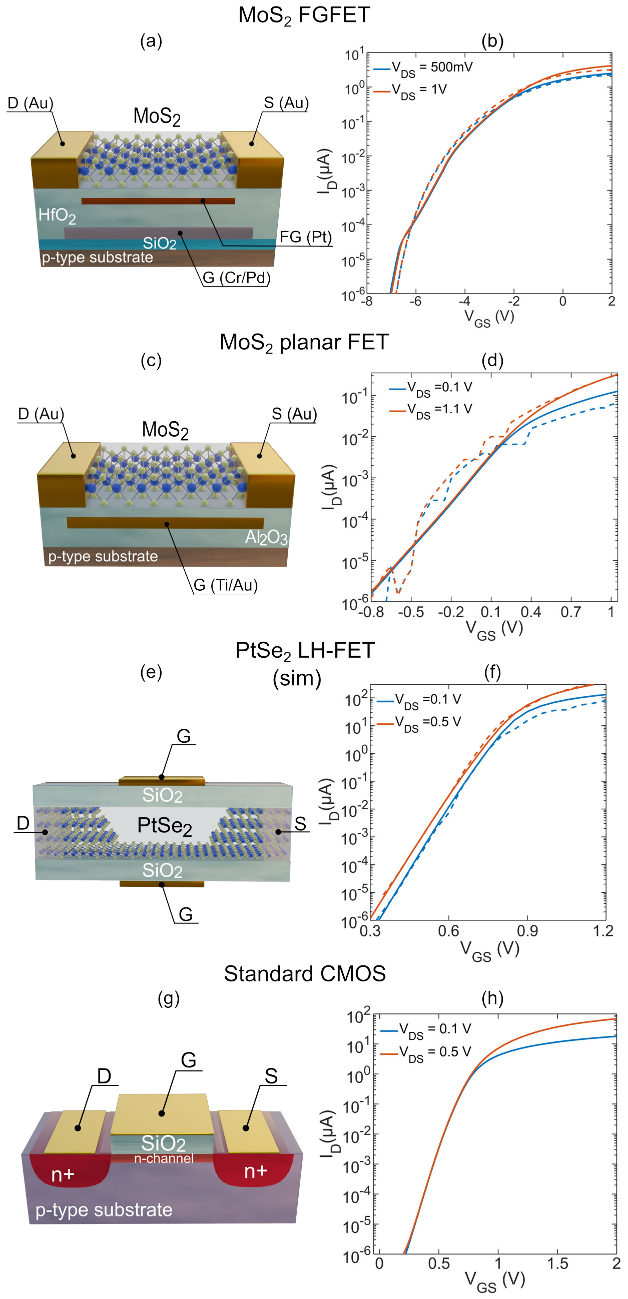

FIGURE 2. Representation of the benchmarked technology platforms and the related drain currents as a function of the applied V_{GS} and V_{DS}. (a) and (b) Structure of the MoS_2-based floating-gate FET (FGFET) and the corresponding drain current (with $W = L = 1$ μm), respectively; (c) and (d) 3-D representation of the planar transistor with the MoS_2 channel and the I_D–V_{GS} and I_D–V_{GS}($W = 20$ μm and $L = 5$ μm), respectively. (e) LH-FET with a monolayer $PtSe_2$ channel. (f) Correspondent output and trans-characteristics for $L = 12.5$ nm and $W = 1$ μm. (g) Reference structure of a standard CMOS technology (UMC 0.18 μm). (h) I_D for an nMOS with a sizing: $W = L = 1$ μm. Dashed lines: measurements (MoS_2-based devices) or device simulations ($PtSe_2$ transistor); solid lines: compact model simulations.

parameters (e.g., device internal capacitances) and noise models. For this reason, we have selected the BSIM-SOI [24] transistor model as a reference template to reproduce the subthreshold operation of all the device options investigated in this work. This model is physically based and can be tailored to reproduce the operation of many FETs by tuning few parameters. Although BSIM-SOI has been developed for silicon-on-insulator FETs, it includes the relevant physics to simulate the ultimate ultrathin-body devices that transistors based on 2-D materials represent. The static and dynamic simulations have been performed within reasonable limits on the geometry and by considering the same operating region as the one of the available experiments, in order to neglect effects which may compromise the model validity. Both physics-based device simulations (for the $\mathrm{PtSe_2}$ LH-FET) and experimental data (for the $\mathrm{MoS_2}$ device options) have been exploited to carefully calibrate the BSIM-SOI model.

A. STANDARD CMOS BENCHMARK

The assessment of novel transistors is given in relation to a mature technology. The reference VMM is realized in a standard 0.18 μm single-poly CMOS technology (UMC 0.18 μm), with a sketch of a transistor represented in Fig. 2(g). This silicon technology includes devices operating at 3.3 V nominal voltage, whose I_D–V_{GS} characteristics are shown in Fig. 2(h). The oxide thickness of 3.3 V transistors is $t_{\mathrm{ox}} = 7$ nm, which provides an adequate retention time for a neural network operation. Standard CMOS reference circuits have been simulated by means of Cadence Virtuoso models available within the dedicated process-design-kit. The realized prototype features competitive performance in terms of both the EE and area occupation [9].

B. MoS$_2$ FGFET

FGFETs are FG non-volatile memories that represent promising candidates for programmable logic applications [5]. Monolayer $\mathrm{MoS_2}$ FGFETs have been successfully fabricated and used for a reconfigurable logic [17]. The experimental device [17] has a bottom-gate configuration with a platinum FG isolated from a control gate (CG) by a 30-nm thick layer of $\mathrm{HfO_2}$ and separated from the channel by a 7-nm-thick $\mathrm{HfO_2}$ layer, as shown in Fig. 2(a). It has a channel length $L = 1$ μm. Additional details can be found in [17]. The BSIM-SOI model has been fit to experimental data: the I_D–V_{GS} transfer characteristics are shown in Fig. 2(b), for a device with $W = 1$ μm. The FGFET has been modeled as a standard MOS transistor with a capacitor in series to the gate. This device is externally accessible through the drain and source of the transistor and from the external terminal of the capacitor (i.e., the CG). The equivalent model of the intrinsic MOSFET has the same oxide thickness of the original device, that is, $t_{ox} = 7$ nm, which guarantees an acceptable retention time. Due to the missing experimental data for a complementary pMOS, a virtual pMOS device has been conceived using similar model parameters as the ones obtained for the nMOS, leading to a device with almost mirrored characteristics. Finally, the I_{off} of the FGFET and of the virtual pMOS were aligned by tuning the threshold voltages, which can be done through a proper choice of the work function of the CG.

C. PLANAR MoS$_2$ TRANSISTOR

As in the single-poly CMOS option, a planar $\mathrm{MoS_2}$ MOSFET can be used to realize a three-terminal FG cell by adding a capacitor in series to the gate. The experimental $\mathrm{MoS_2}$ device presented in [18] and sketched in Fig. 2(c) has already been used to implement an operational amplifier (OPAMP) [18]. The device has been realized as follows: an $\mathrm{MoS_2}$ film for the channel, grown by chemical vapor deposition on a silicon substrate, separated by a 30-nm-thick $\mathrm{Al_2O_3}$ from the Ti/Au back-gate. The fabricated $\mathrm{MoS_2}$ transistor features a large on/off current ratio of eight orders of magnitude [18], as shown in Fig. 2(d). In the same figure, the I_D–V_{GS} of the BSIM-SOI model, calibrated to reproduce experimental I–V characteristics in the subthreshold region, are shown. Also in this case, a virtual pMOS was derived from the calibrated device, by changing the channel type from "n" to "p." After the calibration, some geometrical parameter has been modified in order to get a device suitable for the specific application. For instance, an oxide thickness of 30 nm does not allow to program or erase the floating-gate memory, thus the t_{ox} was lowered down to 10 nm. The reduced $t_{\mathrm{ox}} = 10$ nm provides a steeper SS with respect to the original device. Finally, I_{off} was also tuned by increasing the threshold voltage.

D. PtSe$_2$ LH-FET

As an ideal 2-D material device, we have also considered the $\mathrm{PtSe_2}$ double-gate LH-FET proposed in [19]. This is only a concept based on simulations, and no experimental data have been provided up to date.

The device is based on a channel obtained with an LH consisting of a monolayer $\mathrm{PtSe_2}$ region under the gate and by a multilayer $\mathrm{PtSe_2}$ for the source and drain regions, as shown in Fig. 2(e). The multilayer $\mathrm{PtSe_2}$ is a metal, enabling low contact resistance, whereas the monolayer $\mathrm{PtSe_2}$ is a semiconductor with an energy gap of 1.36 eV. The I–V characteristics for an LH-FET with channel length in the range of few nanometers have been simulated by considering only ballistic transport and ideal contact resistance [19]. A model with a so detailed physical description is not adequate to perform circuit simulations with several devices. Thus, also in this case, we have calibrated a BSIM-SOI model. The device is ambipolar, that is, it can be operated in both nMOS and pMOS modes based on the bias condition. However, we have realized two separate calibrations for each operation mode to independently optimize the threshold voltages: the resulting I_D–V_{GS} transfer characteristics of the nMOS for $V_{\mathrm{DS}} = 1$ and 0.5 V are shown in Fig. 2(f) (the pMOS characteristics are symmetrical). To obtain an acceptable retention time, the silicon oxide thickness has been increased to 7 nm (the original equivalent oxide thickness of the transistor for logic was 0.5 nm), while keeping the other parameters unchanged. The increased t_{ox} of course leads to an electrostatics degradation, resulting in very low on current of 10 nA at $V_{GS} = V_{DS} = 0.5$ V.

VI. PROGRAMMABLE CM DESIGN

The inputs of a CM-based VMM are encoded as a vector of currents with a full-scale value $I_{\mathrm{in,MAX}}$, which is properly chosen in order to ensure the required ENOB. In addition, the selected supply voltage is the minimum V_{DD} which guarantees constant THD and SNR [9]. From the architectural prospective, there are few possible topologies for programmable CMs. Here we consider two options: the symmetric simple CM [SSCM, Fig. 3(a)] and the symmetric cascode CMs [SCCM, Fig. 3(b)]. The FG basic cell is used in both the input and output cells of the mirror, to maximize the symmetry, so that the current magnification factor only depends on the threshold voltage different of the two FG devices of the mirror. We have verified that the use of a "dummy" input FG ensures a better linearity of the multiplier in the whole input current range, compared to the case where the FG cell is employed only in the multiplying stage [9].

For any considered technology, a symmetric simple and/or cascode CM have been carefully designed with proper transistor sizing and operating current values in order to obtain a given ENOB accuracy, by relying on the design methodology reported in [9].

We assume that an $M \times N$ layer is followed by an $N \times K$ layer. In our implementation, the input current signal is provided to a diode-connected transistor, connected to N (or K, for the second VMM) columns with N (K) parallel C_{GS} parasitic capacitances.

In order to address the fan-out issue, we assume that:

1) the output signals are processed by an activation function (not implemented in circuit in our work), before being transferred to the next layer;
2) the activation function of the first layer is designed in a way to provide a magnified output current (by a factor of K) to the next VMM, realized by p-type CMs with a magnification factor of K. Then, the input-cell transistor width of the second VMM is sized K times the width of the multiplying-cell transistors; for consistency, we also assume that the M input cells of the first $M \times N$ VMM are sized accordingly (with a width sized N times the width of the multiplying-cell transistors);
3) for the N output currents of the first layer (and the K output currents of the second layer), they are mirrored by N (K) p-type CMs, similar to the single-cell case represented in Fig. 3. The p-type current-mirror transistors are sized with a width equal to $4 \times M$ ($4 \times N$) times the n-type ones, in order to work in a similar operating region. These outputs are sent to the activation function.

In this configuration, each multiplying cell operates at the nominal current, and the latency becomes independent of the VMM size.

Important parameters for the sizing of the CM are also C_{MULT} and C_{IN}, the capacitances between the CG and the FG of the multiplying and input FG devices of the mirror, respectively. We call "coupling ratio" the ratio $C_{\mathrm{MULT}}/C_{\mathrm{nMOS}}$, where $C_{\mathrm{nMOS}} = C_{\mathrm{ox}} \times$ LW and C_{ox} is the FET oxide capacitance per unit area, and "capacitance ratio" the ratio $C_{\mathrm{IN}}/C_{\mathrm{MULT}}$.

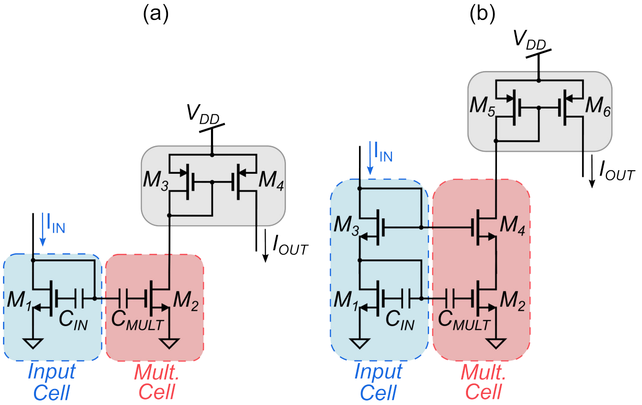

FIGURE 3. Schematic of (a) symmetric simple CM (SSCM) and (b) symmetric cascode CM (SCCM).

The finite coupling ratio $C_{\mathrm{MULT}}/C_{\mathrm{nMOS}}$ introduces non-idealities due to a finite electrostatic coupling, since $\eta = (1 + C_{\mathrm{nMOS}}/C_{\mathrm{MULT}})$, which degrade the mirror linearity. On the other hand, a large coupling ratio would imply a large capacitance area and high area occupation. Therefore, a different optimization approach is proposed in this article: $C_{\mathrm{MULT}}/C_{\mathrm{nMOS}}$ and $C_{\mathrm{IN}}/C_{\mathrm{nMOS}}$ have been independently optimized to compensate for the input-cell/multiplying-cell asymmetry rising from different drain bias. This approach enables quite good linearity characteristics for the designed CMs, with no need to use extremely large area for the capacitors.

Only for the MoS_2 FGFET, an FG structure is intrinsic to the cell, and therefore the coupling ratio is always smaller than 1, since the dielectric between CG and FG has to be thicker than the tunnel dielectric, and the capacitors have the same area. This is the solution that minimizes the CM area. For the standard CMOS, C_{IN} and C_{MULT} were realized with pCAPs, similar to the VMM reported in [9]. For the planar MoS_2 and $PtSe_2$, an ideal capacitor in series to the gate of a transistor was used to reproduce the memory cell, while the layout area was calculated by using the same rules as for the pCAP (see Supplementary Material). If a separate capacitor in series to the FET is used, the coupling ratio can also be larger than 1.

Beyond nonlinearity, also noise can degrade the ENOB of a VMM. In this regard, we have considered the default model available in the BSIM-SOI template, without calibration due to the lack of experimental data for noise in the 2-D devices. This is consistent with the assumption that with technology optimization, 2-D devices can exhibit a level of noise comparable to CMOS devices.

By considering the design of the $PtSe_2$ CM, first we have optimized the THD by varying both $C_{\mathrm{IN}}/C_{\mathrm{nMOS}}$ and $C_{\mathrm{MULT}}/C_{\mathrm{nMOS}}$ ratios (with $C_{\mathrm{IN}} = C_{\mathrm{MULT}}$). Once the optimal $C_{\mathrm{MULT}}/C_{\mathrm{nMOS}}$ ratio has been found, we have fixed the nMOS size and the $C_{\mathrm{IN}}/C_{\mathrm{MULT}}$ ratio has been varied to compensate for the asymmetry between the input and the multiplication. The simulated SSCM shows poor THD values, that is, lower

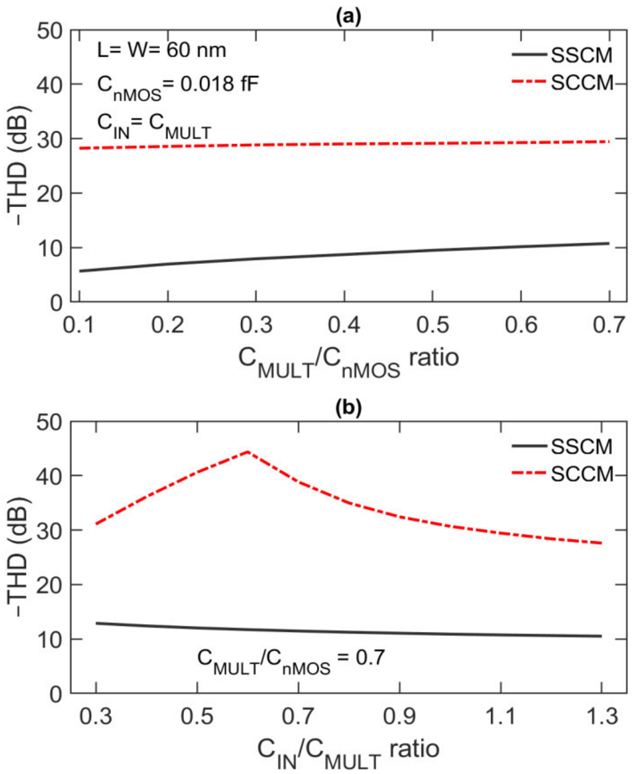

FIGURE 4. Linearity (THD) curves of the $PtSe_2$-based VMMs reported for symmetric simple (solid lines) and symmetric cascode (dashed lines) topologies (a) as a function of the coupling ratio C_{MULT}/C_{nMOS} and (b) for the C_{IN}/C_{MULT} coupling ratio variation for a fixed C_{MULT}/C_{nMOS}.

TABLE 1. VMM Sizing for the Maximum ENOB.

Tech	*UMC 0.18* µm		*$PtSe_2$* (sim)	*MoS_2*		*FGFET*
Topology	*SSCM*	*SCCM*	*SCCM*	*SSCM*	*SCCM*	*SSCM*
Max. ENOB	*7*	*7*	*6*	*5*	*5*	*5*
W (µm)	*6*	*6*	*0.6*	*1*	*1*	*1*
L (µm)	*1.5*	*1.25*	*0.06*	*2*	*1.5*	*1*
C_{MULT}/C_{nMOS}	*25*	*9*	*0.7*	*0.7*	*0.1*	*0.7*
C_{IN}/C_{MULT}	*1*	*1*	*0.6*	*0.5*	*1*	*1*

than 20 dB [see Fig. 4(a)] for any C_{MULT}/C_{nMOS}, and further trimming of C_{IN}/C_{MULT} ratio leads only to a slight improvement [see Fig. 4(b)]. On the other hand, the cascode topology reaches a THD of 30 dB for $C_{MULT}/C_{nMOS} = 0.7$ [see Fig. 4(a)]. Therefore, only cascode mirrors can reach an ENOB of 5. A similar optimization approach can also be applied to the other devices and for a different ENOB.

A 6 bit precision is achieved for the $PtSe_2$ cascode option by using the following parameters: channel length of 60 nm and width of 0.6 μm, $C_{MULT}/C_{nMOS} = 0.7$, and $C_{IN}/C_{MULT} = 0.6$ (see Table 1). As regard to the $PtSe_2$ SSCM option, the achievable accuracy is too low to be exploited for the target application and then it will not be considered in the following.

The same optimization technique has also been applied for VMM implemented with the other considered devices. In Table 1, we summarized the sizing corresponding to the maximum reachable ENOB for each technology for both the SSCM and SCCM topologies. For instance, the VMM realized with MoS_2 transistors can reach the maximum accuracy of ENOB = 5. The UMC 0.18 μm CMOS platform can reach an ENOB = 7, for a coupling ratio larger than 1. It should be emphasized that the FGFET devices has an incorporated FG cell with fixed width and length which cannot be tuned independent of the nMOS, thus a variation of the FGFET geometry corresponds to a variation of both the internal transistor and of the associated FG capacitor. Instead, it is possible to realize a device with a different oxide thickness between the CG and the FG in order to trim the C_{MULT}/C_{nMOS} ratio, if necessary. In this case, only an SSCM topology was designed and tested.

VII. BENCHMARK AND DISCUSSION

Minimum area VMMs were designed for different ENOB specifications through extensive parametric simulations. Extracted FOMs for the designed VMMs are then used to compare the different device/materials options. Table 2 (see Supplementary Material) summarizes the geometrical and electrical design characteristics and the corresponding FOMs, defined in Section IV. It is important to note that the area occupation was calculated with two different approaches: by considering only the gate area, or by considering a possible physical layout (details of the layout area are in the Supplementary Material).

In general, the area increases for a higher ENOB due to the longer devices required to achieve an improved linearity and to increased width needed to meet the SNR specification [9]. Only the UMC 0.18 μm technology can reach a 7 bit specification with 7.43 and 4.05 mm^2 layout area for the SSCM and SCCM cases, respectively, considering a 100 × 100 multiplier (Table 2 in Supplementary Material). The $PtSe_2$ CM can be sized for a 6 bit ENOB using the cascode mirrors. All the considered technologies can reach an ENOB target of 5 or 4, even though only the SSCM topology is possible for the FGFET devices, and only SCCM can operate with such a precision for the $PtSe_2$ devices. As we mentioned, the $PtSe_2$ option represents an ideal asymptotic reference case, and indeed, for a given ENOB, the $PtSe_2$ VMM occupies the smallest area, being more than ten times smaller than the FGFET case and more than two orders of magnitude smaller than the planar technologies (MoS_2 and standard CMOS). The CMOS VMMs are almost 50% smaller than the planar MoS_2 counterparts featuring the same ENOB, except for the 4-bit SCCM, where the CMOS has only a 23% advantage over the planar MoS_2 counterpart. Finally, due to the vertical integration of the FGFET memory, the layout areas of the FGFET implementations are much smaller than those of the other solutions with similar device size.

In Figs. 5 and 6, the EE (expressed as TOPs/J) is benchmarked against the number of operations per layout area and throughput, respectively. One should note that for the

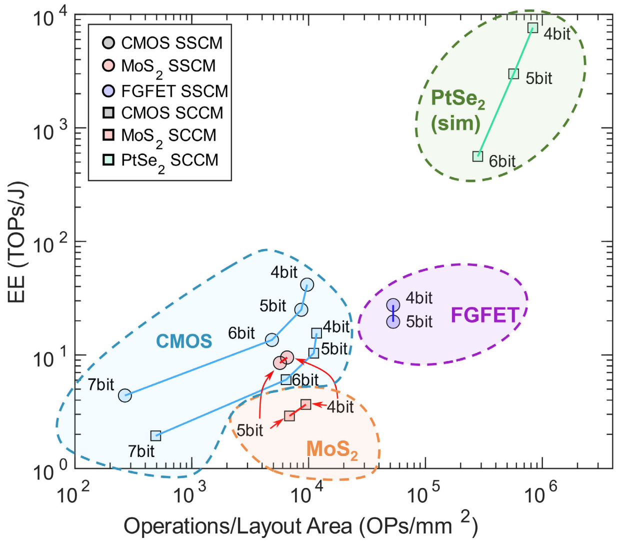

FIGURE 5. **Comparison of area and EE of the designed CM-based VMM implemented with the considered technologies for different computation accuracy. Red symbols and lines: SSCMs; blue symbols and lines: SCCM.**

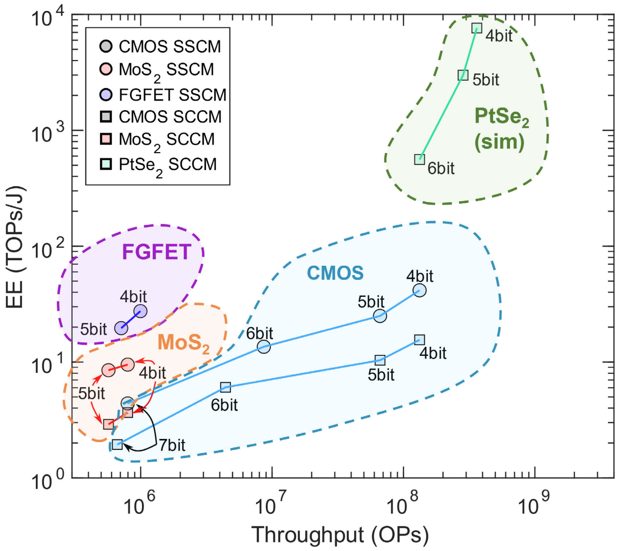

FIGURE 6. **Comparison of throughput and EE of the designed CM-based VMM implemented with the considered technologies for different computation accuracy. Circle symbols: SSCMs; square symbols: SCCM.**

same ENOB and the same device type, the SCCM topologies outperform the SSCM ones in terms of area occupation, even if they require a higher number of devices, while the SSCM case has a better EE than SCCM since it can operate at lower currents for the same linearity conditions (Table 2 in Supplementary Material).

A partially expected but still remarkable observation is that, for a given technology, an increased ENOB can be obtained at the cost of a larger area, lower EE, and throughput. System-level and application-level optimization is therefore key at the design stage of the analog VMM.

Among considered technologies, the performance of the ideal PtSe$_2$ platform represents an asymptotic limit in terms of both energy consumption and throughput. For instance, if we consider the ENOB $= 4$ case, an EE of 7.63E3 TOPs/J can be reached with the PtSe$_2$ VMM. This EE value is $\times 183$, $\times 277$, and almost $\times 2084$ higher than the CMOS (41.6 TOPs/J), MoS$_2$ FGFETs (27.5 TOPs/J), and planar MoS$_2$ (3.66 TOPs/J) platforms for the same ENOB, respectively. Similarly, the simulated PtSe$_2$ has the lowest τ_{LAT} and outperforms other technologies in terms of throughput for the same ENOB case, which is from $\times 3$ to three decades faster than the other technologies.

Considering the two experimental MoS$_2$ devices, the FGFET archives a higher EE due to a lower supply voltage and $I_{\mathrm{in,MAX}}$. On the other hand, the standard CMOS has two orders of magnitude of lower latency than the one achieved by using fabricated MoS$_2$ devices, and therefore reaching a higher EE, despite an almost $\times 40$ of $I_{\mathrm{in,MAX}}$.

VIII. CONCLUSION

In this work, we have investigated the use of 2-D material devices in the design of dedicated hardware for analog deep neutral networks using in-memory computing, considering the same architecture based on current-mode analog VMMs. Layout area, EE, and throughput have been extracted for different target ENOB. The ideal PtSe$_2$ LH-FET transistor simulated in [19] provides asymptotic performance limits with more than one order of magnitude better energy efficiency and area occupation for similar accuracy than standard CMOS. Among two experimental MoS$_2$ options, the FGFET [17] presents an EE comparable to the standard CMOS, despite being slower, while it is advantageous in terms of area occupation. Estimations performed on experimental MoS$_2$ planar MOSFETs, presented in [18], show a comparable area occupation with the standard CMOS. However, the planar MoS$_2$ device presents the lowest EE and throughput compared to the considered technologies. It is also important to note that only the CMOS case achieves ENOB $= 7$, which clearly means that is very important to improve the electrostatics of devices with 2-D materials leading to a better linearity.

As a final remark, let us stress that the scope of this comparison is limited to CM-based VMM and results can be different for other implementations, such as memristor-based ones. We also would like to highlight the fact that for MoS$_2$ options we have considered experimentally demonstrated devices, assuming that deep trap states—often responsible for extremely slow transients [29]—are under control. Very large improvements of the implementations based on MoS$_2$ FETs can be obtained if contact resistances (that are more than 10 kΩ/μm in the considered devices) are reduced, leading to higher currents and therefore lower latency.

REFERENCES

[1] T. J. Sejnowski, "The unreasonable effectiveness of deep learning in artificial intelligence," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 48, pp. 30033–30038, Dec. 2020, doi: 10.1073/pnas.1907373117.

[2] V. Sze, Y.-H. Chen, J. Emer, A. Suleiman, and Z. Zhang, "Hardware for machine learning: Challenges and opportunities," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Austin, TX, USA, Apr. 2017, pp. 1–8, doi: 10.1109/CICC.2017.7993626.

[3] H. S. Stone, "A logic-in-memory computer," *IEEE Trans. Comput.*, vol. C-19, no. 1, pp. 73–78, Jan. 1970, doi: 10.1109/TC.1970.5008902.

[4] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2014, pp. 10–14, doi: 10.1109/ISSCC.2014.6757323.

[5] C. Diorio, P. Hasler, A. Minch, and C. A. Mead, "A single-transistor silicon synapse," *IEEE Trans. Electron Devices*, vol. 43, no. 11, pp. 1972–1980, Nov. 1996, doi: 10.1109/16.543035.

[6] C. Diorio *et al.*, "A complementary pair of four-terminal silicon synapses," *Anal. Integr. Circuits Signal Process.*, vol. 13, pp. 153–166, May 1997, doi: 10.1023/A:1008244314595.

[7] C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A floating-gate MOS learning array with locally computed weight updates," *IEEE Trans. Electron Devices*, vol. 44, no. 12, pp. 2281–2289, Dec. 1997, doi: 10.1109/16.644652.

[8] X. Xu *et al.*, "Scaling for edge inference of deep neural networks," *Nature Electron.*, vol. 1, no. 4, pp. 216–222, 2018, doi: 10.1038/s41928-018-0059-3.

[9] 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, doi: 10.1109/ACCESS.2020.3037017.

[10] J. Binas, D. Neil, G. Indiveri, S.-C. Liu, and M. Pfeiffer, "Precise neural network computation with imprecise analog devices," pp. 1–22, 2016, *arXiv:1606.07786*. [Online]. Available: http://arxiv.org/abs/1606.07786

[11] 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, doi: 10.1109/JPROC.2018.2871057.

[12] 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, doi: 10.1109/JSSC.2014.2356197.

[13] X. Guo *et al.*, "Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Austin, TX, USA, Apr. 2017, pp. 1–4, doi: 10.1109/CICC.2017.7993628.

[14] Q. H. Wang, K. Kalantar-Zadeh, A. Kis, J. N. Coleman, and M. S. Strano, "Electronics and optoelectronics of two-dimensional transition metal dichalcogenides," *Nature Nanotechnol.*, vol. 7, no. 11, pp. 699–712, 2012, doi: 10.1038/nnano.2012.193.

[15] S. Manzeli, D. Ovchinnikov, D. Pasquier, O. V. Yazyev, and A. Kis, "2D transition metal dichalcogenides," *Nature Rev. Mater.*, vol. 2, no. 8, Aug. 2017, Art. no. 17033, doi: 10.1038/natrevmats.2017.33.

[16] G. Iannaccone, F. Bonaccorso, L. Colombo, and G. Fiori, "Quantum engineering of transistors based on 2D materials heterostructures," *Nature Nanotechnol.*, vol. 13, no. 3, pp. 183–191, 2018, doi: 10.1038/s41565-018-0082-6.

[17] G. M. Marega *et al.*, "Logic-in-memory based on an atomically thin semiconductor," *Nature*, vol. 587, no. 7832, pp. 72–77, Nov. 2020, doi: 10.1038/s41586-020-2861-0.

[18] D. K. Polyushkin *et al.*, "Analogue two-dimensional semiconductor electronics," *Nature Electron.*, vol. 3, no. 8, pp. 486–491, Aug. 2020, doi: 10.1038/s41928-020-0460-6.

[19] G. Calogero, D. Marian, E. G. Marin, G. Fiori, and G. Iannaccone, "Physical insights on transistors based on lateral heterostructures of monolayer and multilayer PtSe$_2$ via *Ab initio* modelling of interfaces," *Sci. Rep.*, vol. 11, no. 1, Dec. 2021, Art. no. 18482, doi: 10.1038/s41598-021-98080-y.

[20] *The MNIST Database of Handwritten Digits*. Accessed: Nov. 27, 2021. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[21] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, ch. 5. [Online]. Available: https://www.deeplearningbook.org/

[22] 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, doi: 10.1109/TCSII.2012.2184371.

[23] P. Kim, *MATLAB Deep Learning*. Seoul, South Korea: Apress, 2017, doi: 10.1007/978-1-4842-2845-6.

[24] *BSIM SOI Model—BSIM Group at Berkeley University*. Accessed: Nov. 27, 2021. [Online]. Available: http://bsim.berkeley.edu/models/bsimsoi/

[25] A. Rofougaran and A. A. Abidi, "A table lookup FET model for accurate analog circuit simulation," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 12, no. 2, pp. 324–335, Feb. 1993, doi: 10.1109/43.205011.

[26] M. K. Q. Jooq, A. Mir, S. Mirzakuchaki, and A. Farmani, "Semi-analytical modeling of high performance nano-scale complementary logic gates utilizing ballistic carbon nanotube transistors," *Phys. E, Low-Dimensional Syst. Nanostruct.*, vol. 104, pp. 286–296, Oct. 2018, doi: 10.1016/j.physe.2018.08.008.

[27] L. Wang, Y. Li, X. Gong, A. V.-Y. Thean, and G. Lian, "A physics-based compact model for transition-metal dichalcogenides transistors with the band-tail effect," *IEEE Electron Device Lett.*, vol. 39, no. 5, pp. 761–764, May 2018, doi: 10.1109/LED.2018.2820142.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, 2012, pp. 1097–1105, doi: 10.1145/3065386.

[29] M. Macucci, G. Tambellini, D. Ovchinnikov, A. Kis, G. Iannaccone, and G. Fiori, "On current transients in MoS$_2$ field effect transistors," *Sci. Rep.*, vol. 7, no. 1, Dec. 2017, Art. no. 11575, doi: 10.1038/s41598-017-11930-6.

MAKSYM PALIY received the M.S. degree in electronic engineering from the University of Calabria (UNICAL), Rende, Italy, in 2016, with a thesis titled "Design of 3T CMOS Current Reference for Ultra-Low Voltage Application." He is currently pursuing the Ph.D. degree in electronic engineering with the University of Pisa (UNIPI), Pisa, Italy.

He is developing analog neural networks with CMOS and beyond CMOS technologies. His current interests include the design of low-power analog integrated circuits, analog neural networks, device modeling, and power management circuits.

SEBASTIANO STRANGIO (Member, IEEE) was with the University of Udine, Udine, Italy, as a Temporary Research Associate, from 2013 to 2016, and with 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 with 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 in electronics at the University of Pisa, Pisa, Italy. He has authored and coauthored over 25 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)], as well as CMOS image sensors, power devices, and circuits based on wide-bandgap materials.

PIERO RUIU received the B.S. and M.S. degrees *(cum laude)* in electronic engineering at the University of Pisa, Pisa, Italy, in 2017 and 2020, respectively, with a master thesis on non-volatile memory design for analog computation.

He is currently an Analog Design Engineer at the University of Pisa. He has worked on the design of analog and mixed signal integrated circuits for analog deep neural networks (DNNs) and on the design of analog-based physical unclonable functions (PUFs) for anticounterfeiting chips.

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 at 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 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.

Dr. Iannaccone is a fellow of the American Physical Society.