Received 19 January 2020; revised 24 February 2020; accepted 11 March 2020.
Date of publication 16 March 2020; date of current version 6 July 2020.

Digital Object Identifier 10.1109/JXCDC.2020.2981048

Energy-Efficient Moderate Precision Time-Domain Mixed-Signal Vector-by-Matrix Multiplier Exploiting 1T-1R Arrays

SHUBHAM SAHAY (Member, IEEE), MOHAMMAD BAVANDPOUR, MOHAMMAD REZA MAHMOODI, and DMITRI STRUKOV (Senior Member, IEEE)

California Nano Systems Institute (CNSI), University of California at Santa Barbara, Santa Barbara, CA 93106 USA
Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106 USA

CORRESPONDING AUTHOR: S. SAHAY (shubhamsahay@ucsb.edu)

This work was supported in part by CRISP, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

ABSTRACT The emerging mobile devices in the era of Internet-of-Things (IoT) require a dedicated processor to enable computationally intensive applications such as neuromorphic computing and signal processing. Vector-by-matrix multiplication is the most prominent operation in these applications. Therefore, there is a critical need for compact and ultralow-power vector-by-matrix multiplier (VMM) blocks to perform resource-intensive low-to-moderate precision computations. To this end, in this article, we propose a time-domain mixed-signal VMM exploiting a modified configuration of 1MOSFET-1RRAM (1T-1R) array. The proposed VMM overcomes the energy inefficiency of the current-mode VMM approaches based on RRAMs. A rigorous analysis of different nonideal factors affecting the computational precision indicates that the nonnegligible minimum cell currents, channel length modulation (CLM), and drain-induced barrier lowering (DIBL) are the dominant mechanisms degrading the precision of the proposed VMM. We also show that there exists a tradeoff between the computational precision, dynamic range, and the area- and energy-efficiency of the proposed VMM approach. Therefore, we provide the necessary design guidelines for optimizing the performance. Our preliminary results indicate that an effective computational precision of 6 bits is achievable owing to the inherent compensation effect in the modified 1T-1R blocks. Furthermore, a 4-bit 200 $\times$ 200 VMM utilizing the proposed approach exhibits a significantly high energy efficiency of $\sim$1.5 Pops/J and a throughput of 2.5 Tops/s including the contribution from the input/output (I/O) circuitry.

INDEX TERMS 1T-1R array, mixed-signal VMM, time-domain encoding, vector-by-matrix multiplication.

I. INTRODUCTION

THE traditional digital processors are extremely energy inefficient while handling high-dimensional data from operations such as object/speech recognition, image processing, and probabilistic inference [1]–[2]. Moreover, the widespread use of computationally intensive applications such as deep neural networks (DNNs)/recurrent neural networks (RNNs), real-time signal processing, and optimization algorithms in this era of Internet-of-Things (IoT) necessitates the development of dedicated processing blocks within the mobile devices. The vector-by-matrix multipliers (VMMs) form the most integral part (and often the bottleneck) of these computationally intensive systems. Therefore, the development of a compact and energy-efficient VMM engine is highly essential [3]–[16].

The analog-domain VMM implementations are more area- and energy-efficient as compared to the digital counterparts for computational tasks such as inference, classification, and recognition that are robust to low-resolution (reduced precision) VMM operations (and can be trained effectively to handle hardware imperfections without compromising accuracy) [3], [6]–[10]. Recently, VMMs based on emerging nonvolatile memories, RRAMs in particular, have attracted considerable attention since the VMM operation is simplified as current accumulation through programmable resistances in the analog domain [5], [6], [10]. However, the current-mode VMM implementations based on RRAM require high current levels [6], [16] and bulky transimpedance amplifiers at each column of the cross-bar [6], resulting in significant area- and energy-overhead. Moreover, the computational precision is also limited in such implementations and may only be improved at the cost of an increased area to accommodate complex peripheral circuitry implementing sophisticated tuning algorithms or nontrivial mapping techniques [6].

Recently, energy-efficient time-domain VMMs [4], [9]–[15] exploiting postsynaptic pulse (PSP) emulators [11],

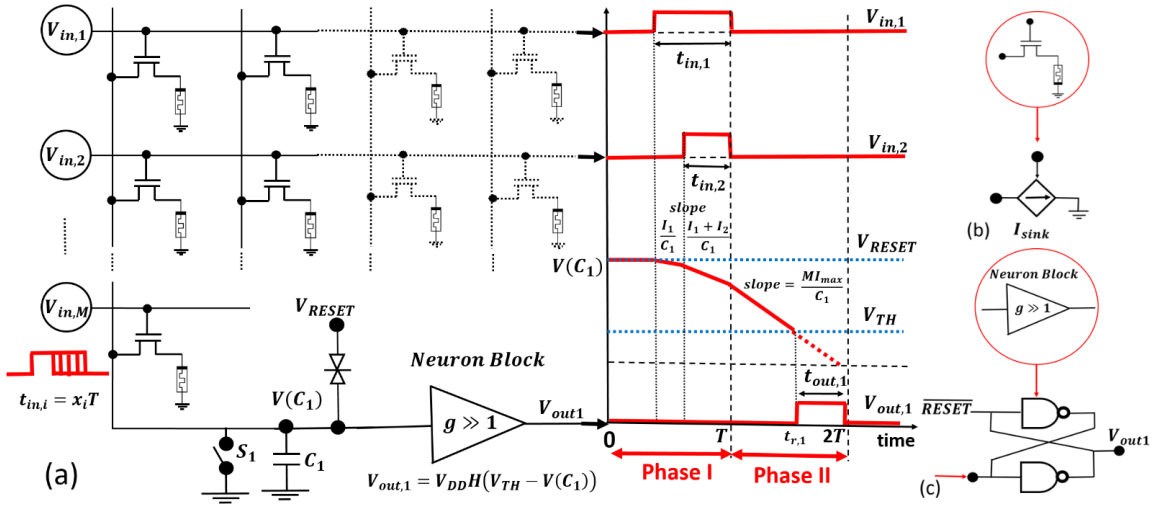

FIGURE 1. Schematic view of (a) VMM circuit utilizing 1T-1R array and the timing diagram of the inputs, outputs, and the voltage across the load capacitor, (b) modified 1T-1R block which acts as a programmable current sink, and (c) peripheral circuit within the neuron block implementing the Heaviside function.

SRAM (binary) outputs [13] as programmable weights have also been proposed. The energy efficiency of even RRAM-based VMM approaches could be significantly improved if such a time-domain switched-capacitor-based approach [8] is followed as opposed to the power-hungry current-mode approach. Although a time-domain VMM exploiting RRAM array was proposed in [10], the utilization of an active neuron circuitry with a current conveyer-based integrator and a ramp-based ADC limits its energy-efficiency. Therefore, to improve the energy-efficiency, time-domain VMMs with passive and digital input/output (I/O) and neuron circuitry exploiting 2-D-NOR flash [15] and 3-D-NAND flash memory [20] were proposed. However, the utilization of a simple digital neuron based on S-R latch without a virtual ground leads to a shift in the voltage at the load capacitor attached to the column of the crossbar and reduces their accuracy. While the high drain-induced barrier lowering (DIBL) in the 2-D-NOR flash memory owing to the poor gate control due to a higher effective oxide thickness (EOT) leads to an increased computational error [15], the significant capacitive-coupling between the bitline (BL) and the bit-select transistor (BSL) in the 3-D-NAND flash memory owing to their 3-D geometry and the consequent charge-disturbance error restricts their computational precision [20]. The utilization of 1Transistor-1RRAM cells consisting of MOSFETs, which exhibit an enhanced electrostatic integrity (and a reduced DIBL) due to a lower EOT and significantly reduced capacitive-coupling due to a low gate-drain capacitance, may overcome these limitations and facilitate realization of VMMs with higher computational precision. To this end, in this article, we propose a time-domain mixed-signal VMM exploiting a modified 1MOSFET-1RRAM (1T-1R) array. Contrary to the conventional 1T-1R blocks, where RRAM is connected to the drain of the MOSFET, the RRAM is attached to the source in this approach. This leads to inherent negative feedback (which we call the "self-compensation effect") that further suppresses the DIBL, and significantly improves the computational precision. In the proposed VMM approach, the weights are realized as programmable current sinks via tuning the conductance state of the RRAM in the modified 1T-1R blocks in the analog domain while the inputs and outputs are encoded as pulse durations in the digital domain. We also performed a rigorous analysis of different factors such as DIBL/channel length modulation (CLM), capacitive coupling, and nonnegligible minimum cell currents, which may degrade the computational precision. Our preliminary results show that an effective computational precision of 6 bits and an energy efficiency of ∼1.5 Pops/J and a throughput of 2.5 Tops/s for a 4-bit 200 × 200 VMM may be achieved utilizing this approach.

This article is organized as follows: the proposed VMM approach is discussed in Section II. The impact of RRAM behavior on the characteristics of the modified 1T-1R cells is discussed in Section A1 (Supplementary Material) and the different physical mechanisms which may affect the performance of the proposed approach are discussed in Section III. The design guidelines for optimizing the performance of the proposed 1T-1R VMM are discussed in Section IV and the area, energy, and throughput estimates are provided in Section V. The detrimental effect of the nonnegligible minimum cell currents and the design methodology to tailor the proposed VMM for lower output currents are discussed in Section VI. The impact of process-induced mismatch effects is discussed in Section A2, a programming scheme for the modified 1T-1R array is proposed in Section A3 in the supplementary material and the conclusions are drawn in Section VII.

II. PROPOSED VMM APPROACH

A generalized $M \times N$ VMM operation may be represented as

$$y_j = \frac{1}{M} \sum_{i=1}^{M} w_{ij} x_i, \quad j = 1, 2, \ldots, N \tag{1}$$

where the inputs x_i, outputs y_j, and weights w_{ij} are normalized such that $(x_i, y_j, w_{ij}) \in [0, 1]$. The proposed time-domain VMM approach exploiting the modified 1T-1R array is shown in Fig. 1. In the time-domain

VMM [9]–[15], [20], [21], the inputs are encoded as the duration of the digital pulses such that

$$t_{\text{in},i} = x_i T \tag{2}$$

where T is the time window for the VMM operation. In the proposed approach, the modified 1T-1R block acts as a programmable current sink as shown in Fig. 1(b) and the digital inputs applied to the gate of the MOSFETs ($V_{\text{in},i}$) enable the ith current sink for a duration $t_{\text{in},i}$. It may be noted that unlike conventional 1T-1R arrays where the RRAMs are connected to the drain of the MOSFETs, in this approach, the RRAMs are connected to the source of the MOSFETs. This modification dissuades the nonidealities such as CLM and DIBL owing to the self-compensation effect as discussed in Section III-C. The weights ($w_{ij} \in [0, 1]$) are mapped to the currents ($I_{ij} \in [I_{\min}, I_{\max}]$) through the programmable current sink as

$$I_{ij} = I_{\min} + w_{ij}\left(I_{\max} - I_{\min}\right). \tag{3}$$

Each column of the programmable current sinks is connected to a load capacitor C_j. A threshold (neuron) circuit proposed in [14] with a transfer function given as

$$V_{\text{out},j} = V_{\text{DD}} H\left(V_{\text{TH}} - V\left(C_j\right)\right) \tag{4}$$

where $H()$ is the Heaviside function that encodes the voltage on the load capacitor C_j into output digital pulse duration.

The entire VMM operation is completed in two cycles (phase-I and phase-II) of duration T each. The load capacitor C_j is initially precharged to a voltage V_{RESET} at the beginning of phase-I ($t = 0$). The inputs are activated only in phase-I (integration phase) and the current sinks start discharging C_j. At the end of phase-I ($t = T$), the voltage across the load capacitor $V(C_j)$ reduces by $\Delta V(C_j)$ where

$$\Delta V\left(C_j\right)_{t=T} = \frac{1}{C_j}\sum_{i=1}^{M} I_{ij} t_{\text{in},i}. \tag{5}$$

Using the expression for I_{ij} from (3) in (5), we get

$$\Delta V\left(C_j\right)_{t=T} = \frac{T\left(I_{\max} - I_{\min}\right)}{C_j}\sum_{i=1}^{M} w_{ij} x_{\text{in},i} + \frac{T I_{\min}}{C_j}\sum_{i=1}^{M} x_{\text{in},i}. \tag{6}$$

As evident from (6), the change in the voltage across the load capacitor at the end of phase-I is mapped to a linear expression of the weighted sum in this scheme. To ensure that this voltage variation across the load capacitor is limited to a targeted operation regime, that is, $V(C_j)_{t=T} \in [V_{\text{RESET}}, V_{\text{TH}}]$, the load capacitor C_j must be designed such that

$$C_j = \frac{M I_{\max} T}{V_{\text{RESET}} - V_{\text{TH}}}. \tag{7}$$

In phase-II (evaluation phase), all the inputs are inactivated and the load capacitor is discharged through a constant current $MI_{\max}$. This discharging current may be generated either via a current mirror or by adding a similar 1T-1R array at the load capacitor with all the inputs activated for the entire duration T during phase-II and all the current sinks programed to $I_{\max}$. In this article, we have followed the latter approach to implement the constant current source during phase-II. The neuron circuit generates an output pulse when the voltage on the load capacitor reaches the threshold voltage, that is, ($V(C_j) = V_{\text{TH}}$). The time instance ($t_{r,j}$) at which $V(C_j) = V_{\text{TH}}$ can be given as

$$t_{r,j} = T - t_{\text{out},j} = T\left[1 - \frac{\sum_{i=1}^{M} I_{ij} t_{\text{in},i}}{M I_{\max} T}\right]. \tag{8}$$

The output pulse duration ($t_{\text{out},j}$) can be simply obtained by using (1) and (3) in (8) as

$$t_{\text{out},j} = a y_j T + b \tag{9}$$

where

$$a = \frac{\left(I_{\max} - I_{\min}\right)}{I_{\max}}, \quad b = \frac{I_{\min}}{M I_{\max}}\sum_{i=1}^{M} x_{\text{in},i}. \tag{10}$$

Equation (9) clearly indicates that the output result obtained using the proposed scheme is different from the targeted ideal output result ($t_{\text{out},j} = y_j T$) due to the nonzero (appreciable) minimum current ($I_{\min}$) of the 1T-1R cells that lead to the undesirable multiplicative coefficient (a) and the input-dependent additive coefficient (b).

However, it may be noted that the input-dependent additive coefficient (b) can be canceled out by utilizing the differential scheme. In the differential implementation, each weight is realized utilizing two subweights w_{ij}^{+} and w_{ij}^{-} such that

$$w_{ij} = w_{ij}^{+} - w_{ij}^{-} \tag{11}$$

and two subneurons are dedicated to calculate the dot product of inputs and each subweight vector as $t_{\text{out},j}^{+}$ and $t_{\text{out},j}^{-}$ as

$$t_{\text{out},j}^{+} = \frac{a}{M}\sum_{i=1}^{M} w_{ij}^{+} x_i + b \tag{12}$$

$$t_{\text{out},j}^{-} = \frac{a}{M}\sum_{i=1}^{M} w_{ij}^{-} x_i + b. \tag{13}$$

Simple logic circuitry is then employed to generate the final differential output pulse as

$$t_{\text{out},j} = t_{\text{out},j}^{+} - t_{\text{out},j}^{-}. \tag{14}$$

On the other hand, the multiplicative coefficient (a) leads to a reduction in the output time window. This shrinkage can be compensated by either lowering the constant current during the evaluation phase (which extends the time window for phase-II) or increasing the output time-to-digital convertor (TDC) counter frequency. Furthermore, the nonnegligible minimum cell currents also lead to a reduction in the portion of the output swing available for performing useful computations as discussed in Section VI.

III. 1T-1R VMM DESIGN GUIDELINES

The performance of the proposed 1T-1R VMM was evaluated at the 55-nm technology node using process design kit (PDK) from Global Foundries in HSPICE (version N-2017.12 [17]). Furthermore, a rather simplistic compact

model was used for RRAM with the static current–voltage relationship expressed as

$$I_{\text{mem}} = \frac{1}{\beta R_0}\sinh(\beta V_{\text{mem}}) \tag{15}$$

where R_0 is the low-voltage resistance and β is the nonlinearity factor (values listed in Table 1) [18]. A maximum ON-state resistance (R_{ON}) of 2.5 KΩ and a minimum OFF-state resistance (R_{OFF}) of 2.5 MΩ were considered for RRAM to explore the entire design space. Furthermore, a maximum permissible read voltage of 0.5 V without disturbing the programed state (parameters R_0 and β) of RRAM was assumed. Under these assumptions, we evaluated the potential of the proposed 1T-1R time-domain VMM under different operating conditions and different parameters for the RRAM. In the subsequent sections, we discuss the operating conditions and provide the necessary design guidelines to extract the optimum performance from the proposed VMM architecture. It may be noted that the optimal conditions also differ with the input constraints such as VMM size, input voltage, time window, dynamic range (DR $= I_{\max}/I_{\min}$), and targeted precision.

A. PRECISION

The effective weight precision (i.e., programmability of the current sinks) depends on the accuracy of tuning the conductance states of RRAM and degrades due to a drift in the analog conductance state with cycling, temperature, and the inherent intrinsic noise such as random telegraph noise (RTN). Previous works have already shown an effective weight precision greater than 7 bits based on a simple tuning algorithm [19]. The weight precision may be further improved by oxide material engineering or by utilizing more efficient tuning algorithms.

As discussed in [15], the computational error (or output error, $e_{\text{out},j}$) may be decoupled from the weight error and defined separately as the maximum difference between the theoretically calculated output time period considering ideal current sinks [$t^{\text{cal}}_{\text{out},j}$ from equation (8)] and the output time period obtained via transient simulation of the entire VMM circuit ($t^{\text{sim}}_{\text{out},j}$) in HSPICE, spanning over the entire sample space of the weights and inputs that is

$$e_{\text{out},j} = \max_{t\text{out},j} \frac{\left|t^{\text{cal}}_{\text{out},j} - t^{\text{sim}}_{\text{out},j}\right|}{T}. \tag{16}$$

For benchmarking against the digital VMM implementations, the effective computational precision ($P_{\text{out},j}$) can then be defined as

$$P_{\text{out},j} = -\log_2 e_{\text{out},j} - 1. \tag{17}$$

Considering the efficacy of the differential scheme in canceling the impact of the input-dependent additive coefficient (b) as discussed in Section II, improving the noise immunity and enhancing the output precision while enabling the inclusion of bipolar weights [8], two adjacent columns of the modified 1T-1R array were tuned for implementing the positive and negative weight components of the bipolar weight matrix. Furthermore, the adjacent neuron circuits were used to calculate the positive ($V_{\text{out},j} = t^{+}_{\text{out},j}$) and negative ($V_{\text{out}(j+1)} = t^{-}_{\text{out},j}$) components of output in this differential implementation. The final output was then obtained as the time difference between the rising edge of the neuron circuits used for obtaining the positive ($V_{\text{out},j}$) and negative ($V_{\text{out}(j+1)}$) components of the output. This rectified linear (ReLU) operation was implemented utilizing a digital gate for $V_{\text{final},j} = V_{\text{out},j} \cdot \overline{V}_{\text{out}(j+1)}$.

B. NONIDEAL FACTORS

The computational precision is degraded by several factors that tend to deviate the current sinks from draining a constant current. While CLM leads to a linear dependence of the MOSFET's drain current on the drain voltage, the DIBL effect induces threshold voltage shift which further increases the variation in the drain current with the drain voltage. Therefore, in addition to the input gate voltage (V_{GS}), the current through the programmable 1T-1R current sink also depends on the drain voltage, that is, the output voltage at the load capacitor.

To minimize this dependence of the cell currents on the output voltage, we modified the conventional 1T-1R array architecture. Although the RRAM is connected to the drain terminal of the MOSFET in the conventional 1T-1R array, one terminal of RRAM is connected to the source of the MOSFET and the other terminal is grounded in this implementation as shown in Fig. 1(b). An increase in the drain voltage in the modified 1T-1R configuration with RRAM connected to the source leads to an enhanced current flowing through the RRAM. This results in a larger voltage drop across the RRAM. The increased voltage drop across the RRAM provides negative feedback and effectively boosts the source potential leading to a reduction in the effective gate-to-source voltage (V_{GS}) which in turn suppresses the increment in the drain current. Therefore, an increase in the drain current due to the application of a larger drain voltage is compensated by a reduction in the effective gate overdrive voltage in the modified 1T-1R array. This inherent self-compensation effect leads to a diminished dependence of the modified 1T-1R cell currents on the output voltage at the load capacitor.

The error due to CLM and DIBL can be defined as

$$e_{\text{CLM/DIBL}} = 1 - \frac{I(V - \Delta V)}{I(V)} \tag{18}$$

where ΔV is chosen as 1 mV to estimate the local error contours with high accuracy (see Fig. 2). As can be observed from the error contour plots in Fig. 2, we performed a rigorous analysis of the CLM and DIBL error for different input gate voltages and nonideality factor (β) of RRAM within the operating regime of the modified 1T-1R configuration. For all the input gate voltages, we found that the cell currents are relatively independent of the drain voltage (i.e., the CLM/DIBL error is low) for higher drain voltages. Therefore, we selected a high reset voltage, $V_{\text{RESET}} = 0.9$ V, and designed the neuron circuit to have a threshold voltage $V_{\text{TH}} = 0.7$ V to ensure a nondisturbing maximum voltage swing of 0.2 V across the RRAM. Furthermore, we also observe from Fig. 2 that the DIBL/CLM error increases as we reduce the input voltage and operate with a smaller maximum current ($I_{\max}$) to limit the load capacitance [see (7)]. Also, as discussed in Section A1 (Supplementary Material), an increase in the nonlinearity

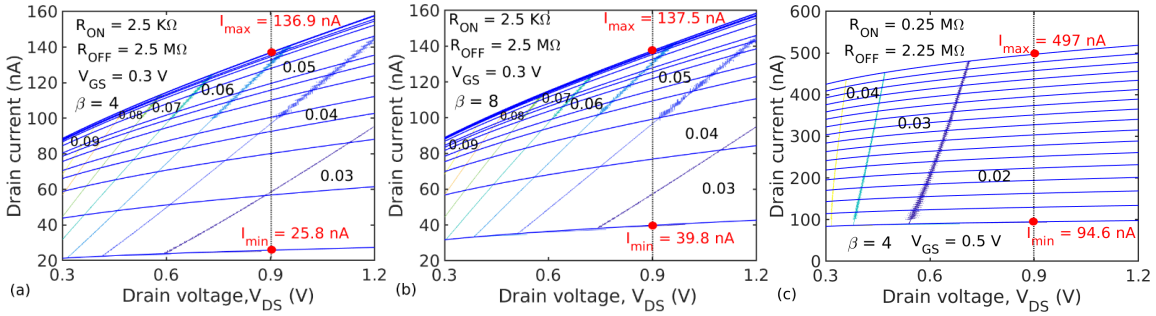

FIGURE 2. **Error contour plot due to the DIBL and CLM effect for different input voltages and nonideality factors: (a) $V_{GS} = 0.3$, $\beta = 4$, (b) $V_{GS} = 0.3$, $\beta = 8$, and (c) $V_{GS} = 0.5$ and $\beta = 4$ for MOSFETs with $L_g = 120$ nm. The values reported within the boundaries of the error contours indicate the maximum CLM/DIBL error for that range of cell currents. The higher cell currents show a lower CLM/DIBL error.**

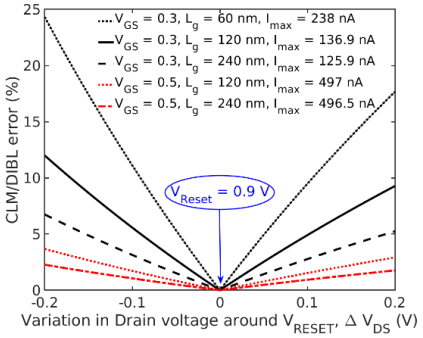

FIGURE 3. **Total error due to the DIBL and CLM effect for different input voltages and gate lengths (L_g) of the MOSFET with nonideality factor, $\beta = 4$.**

factor of the RRAM leads to a reduction in the DR as shown in Fig. 2(b).

Since the DIBL and CLM mechanisms are extremely sensitive to the gate length (L_g) of the MOSFETs, we also performed a thorough investigation of the CLM/DIBL error for 1T-1R cells consisting of MOSFETs with different gate lengths biased at different input voltages as shown in Fig. 3. We found that despite the self-compensation effect, the CLM/DIBL error is significantly high in the MOSFETs with minimum gate length ($L_g = 60$ nm) and reduces drastically by ∼5 times when the gate length is quadrupled to $L_g = 240$ nm for the same input voltage. Moreover, the CLM/DIBL error can be further reduced while obtaining a higher DR by increasing the input voltage V_{GS} to 0.5 V at the cost of an increased capacitor area and energy owing to larger I_{max} (see Fig. 3). Therefore, there exists a tradeoff between area- and energy-efficiency, the DR and the computational error in the proposed approach.

It may also be noted that the DIBL error is lower in the 1T-1R cells as compared to the 2-D-NOR flash memory cells. This is attributed to the smaller EOT and the consequent enhanced electrostatic integrity in the MOSFETs as compared to 2-D-NOR flash cells. Therefore, the computational precision is higher in the proposed approach as compared to 2-D-NOR flash-based time-domain VMM implementation [15].

Apart from the error induced due to CLM and DIBL, the capacitive coupling between the load capacitor and the gate–drain capacitance of the MOSFET could be another possible source of charge disturbance. However, in the proposed architecture, the load capacitor is large as compared to the gate–drain capacitance of the MOSFETs which diminishes the charge disturbance error due to capacitive-coupling. Therefore, the proposed VMM also exhibits a higher precision as compared to the 3-D-NAND flash memory-based time-domain VMM where the significant coupling between the BL and the BSL leads to a considerable charge disturbance error [20].

Furthermore, the intrinsic thermal noise of the MOSFET and the RTN in the RRAM may also affect the computational precision. Also, the process variation in the RRAM may lead to a different retention behavior of the cells within the array even during the read mode and may degrade their computational precision. However, such an analysis is limited due to the lack of a comprehensive model for RRAM covering these variability aspects. Therefore, performance analysis of the proposed VMM under different RRAM noise sources and process variation is important future work.

IV. DESIGN SPACE EXPLORATION

We performed a rigorous analysis to explore the design space for optimizing the performance of the proposed VMM architecture. The input gate voltage (V_{GS}) and time window (T) are the most crucial design parameters for tuning the performance of the proposed VMM for a particular gate length of the MOSFET utilized in the 1T-1R block. The performance-metrics for the proposed VMM with different input voltages (V_{GS}), time window (T), gate lengths (L_g), VMM sizes (M in $M \times M$ VMM), and nonlinearity factor (β) of RRAM are listed in Table 1.

The output (worst case) error (e_{out}) was found by simulating multiple runs of VMM operation in HSPICE with a different combination of random inputs and random weights in each run in an attempt to span the entire sample space of possible input and weight combinations. The line parasitics such as line resistances and capacitances and the corresponding process variations pertinent to the 55-nm technology node were also considered in the HSPICE simulations. The total energy dissipated in the load capacitor, E_{Cl} (which is the dominant energy dissipation mechanism as discussed later in Section V) for the VMM operation has also been included in Table 1. As can be observed from Table 1, the output error reduces with increasing VMM size till $M < 100$. However, as the VMM size, M, increases above 100, the line parasitics and their process variations lead to a nonnegligible increase

TABLE 1. Design space exploration.

V_{GS} (V)	0.3 (L_g = 120 nm, R_{ON} = 2.5KΩ, R_{OFF}=2.5MΩ)						0.3 (L_g = 240 nm, R_{ON} = 2.5 KΩ, R_{OFF}=10MΩ)						0.5 (R_{ON} = 0.25 MΩ)					
β	4			8			4			8			4 (L_g = 120 nm, R_{OFF} = 2.25 MΩ)			4 (L_g = 240 nm, R_{OFF} = 2.375 MΩ)		
I_{max}; I_{min}	136.9 nA ; 25.8 nA			137.5 nA ; 39.8 nA			125.9 nA ; 25.2 nA			126.3 nA ; 38.7 nA			497 nA ; 94.6 nA			496.5 nA ; 94.1 nA		
T (ns)	**16**	**32**	**64**	**16**	**32**	**64**	**16**	**32**	**64**	**16**	**32**	**64**	**16**	**32**	**64**	**16**	**32**	**64**
VMM size, M = 10																		
E_{Cl}(pJ)	0.09	0.19	0.39	0.09	0.19	0.39	0.09	0.18	0.36	0.09	0.18	0.36	0.3	0.7	1.4	0.3	0.7	1.4
e_{out},%	4.5	2.8	2.8	4.4	2.8	2.8	2.6	2.5	2.4	2.6	2.5	2.4	1.44	1.0	0.74	0.88	0.74	0.55
P_{out}	3	4	4	3	4	4	4	4	4	4	4	4	5	5	6	5	6	6
VMM size, M = 50																		
E_{Cl}(pJ)	2.45	4.92	9.85	2.47	4.95	9.9	2.25	4.53	9.06	2.27	4.5	9.09	8.93	17.8	36	8.95	17.8	36
e_{out},%	4.3	2.6	2.7	4.2	2.7	2.6	2.4	2.4	2.3	2.5	2.4	2.2	1.3	0.94	0.72	0.77	0.68	0.48
P_{out}	3	4	4	3	4	4	4	4	4	4	4	4	5	5	6	5	6	6
VMM size, M = 100																		
E_{Cl}(pJ)	9.81	19.7	39.4	9.9	19.8	39.6	9.0	18.1	36.2	9.09	18.2	36.3	35.7	71.5	144	35.6	71.4	144
e_{out},%	4.2	2.6	2.6	4.2	2.6	2.5	2.3	2.3	2.3	2.3	2.4	2.2	1.2	0.92	0.66	0.75	0.64	0.46
P_{out}	3	4	4	3	4	4	4	4	4	4	4	4	5	5	6	6	6	6
VMM size, M = 200																		
E_{Cl}(pJ)	39.2	78.4	157	39.6	79.2	158	36	72.5	145	36.3	72.7	145	142	286	576	142	285	576
e_{out},%	4.3	2.7	2.7	4.4	2.7	2.7	2.4	2.4	2.3	2.4	2.3	2.3	1.3	0.93	0.67	0.77	0.64	0.46
P_{out}	3	4	4	3	4	4	4	4	4	4	4	4	5	5	6	6	6	6

in the computational error. While the line resistances lead to a drop in the effective input (gate) voltage of the MOSFETs on the far end of the 1T-1R array leading to a reduced drain current, the line capacitances add to the latency. Although the differential configuration is effective in mitigating the impact of the fixed line parasitics, the process variations cannot be compensated even exploiting a differential configuration and escalate the computational error. From Table 1, it can also be observed that there is a tradeoff between the computational precision, DR, and the energy dissipated in the load capacitor. For instance, to achieve high computational precision of ∼6 bits for large-sized VMMs ($M > 100$), a higher value of input voltage ($V_{GS} = 0.5$ V) should be used. A higher input voltage results in a higher maximum current (I_{max}) leading to a larger load capacitance. Although the DR is also high for such operating conditions, the area- and energy-efficiency is limited by the load capacitor which dominates the area and energy landscape (as discussed in Section V).

Moreover, to achieve a higher area- and energy-efficiency by limiting the size of the load capacitor, a lower input voltage may be used to reduce the maximum current (I_{max}). However, the computational precision and the DR reduces significantly at such operating conditions. The weight precision may also limit the computational precision in such cases.

Still, the preliminary results indicate that an effective computational precision of 6 bits is achievable for a VMM size, $M > 100$, using the proposed approach. In addition, depending on the targeted precision, input time window, VMM size, area, energy efficiency, voltage swing across RRAM, etc. we may optimize the design parameters to achieve optimum performance of the proposed VMM architecture.

Since the conductance state of the RRAM is sensitive to the voltage drop across them, we have also analyzed the performance of the proposed VMM approach for neuron circuit with different threshold voltages ($V_{TH} > 0.5$ V) to limit the maximum voltage swing across RRAM ($V_{RESET} - V_{TH}$). As can be observed from Fig. 4, a reduction in the maximum voltage swing across RRAM leads to a higher computational precision owing to the lower CLM/DIBL error. Although a reduction in the voltage drop across RRAM increases the load capacitor size according to (7), the energy dissipated in the load capacitor, E_{Cl} decreases owing to the reduced voltage swing as shown in Fig. 4.

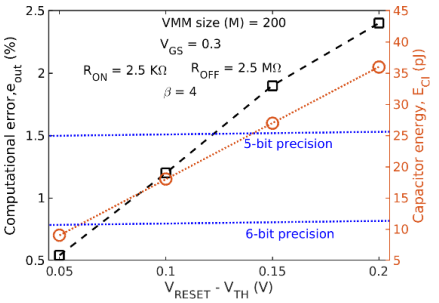

FIGURE 4. Impact of variation in the threshold voltage of the neuron circuit (V_{TH}) to limit the maximum voltage swing across RRAM ($V_{RESET} - V_{TH}$) on the computational error and the capacitor energy of the proposed VMM approach.

V. PERFORMANCE ESTIMATION

It can be observed from Table 1 that the proposed VMM approach yields a computational precision of 3 to 6 bits depending on the design parameters. Targeting a VMM engine with precision of 4 bits, which is sufficient for several applications including neuromorphic computing [8], [10], we select an input voltage $V_{GS} = 0.3$ V, a time window $T = 16$ ns and a gate length $L_g = 240$ nm for estimating the area- and energy-efficiency of the proposed approach. Fig. 5 shows the area and energy breakdown of the proposed VMM considering the I/O peripheral circuitry as well as the neuron circuitry for different VMM sizes.

The basic components of the VMM I/O circuitry are digital input to time-domain pulse converters (DTCs) which consist of a 4-bit shared counter and a 4-bit digital comparator followed by an S-R latch for each input, and the time-domain pulse to digital output converters (TDC) which consist of a 4-bit accumulator for each neuron output [20]. The 4-bit accumulator is realized using a 4-bit full adder and a 4-bit register based on D-flip-flops. A shared clock enables conversion of the pulse duration of the neuron output to digital outputs. The neuron circuit consists of an S-R latch realized

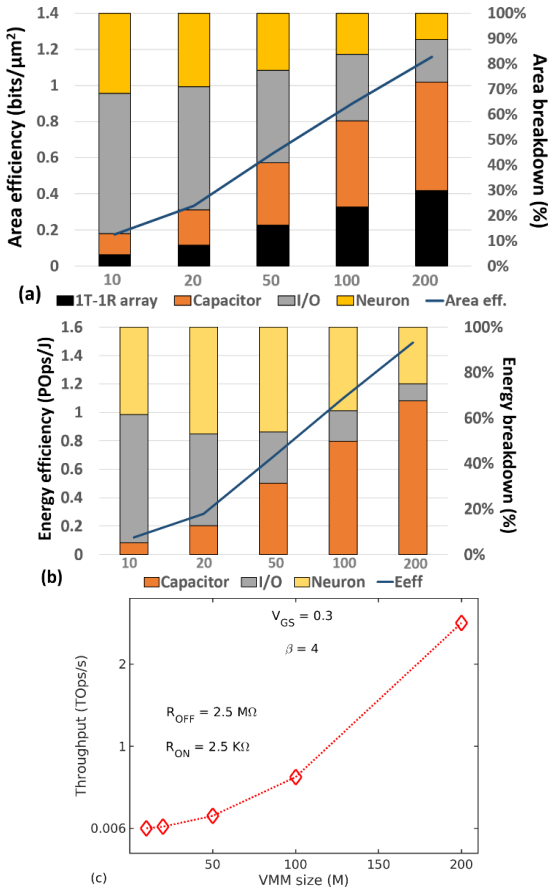

FIGURE 5. Variation of (a) area efficiency, (b) energy efficiency, and their breakdown, and (c) throughput of the proposed VMM with VMM size (*M*) for a ReLU neuron.

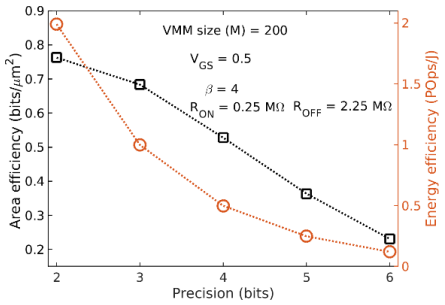

FIGURE 6. Variation of area efficiency and energy efficiency of the proposed VMM with VMM size, $M = 200$ for different targeted precisions.

using a pair of NAND gates followed by an AND gate and NOT gate for implementing the differential scheme [20]. The load capacitors are realized using MOSCAPs from the 55-nm technology node.

It can be observed from Fig. 5 that the I/O circuitry consumes a significant portion of the energy and area landscape of the proposed VMM architecture when the VMM size is small. However, the load capacitor (C_j) tends to dominate the area and energy landscape as the VMM size increases. The preliminary results indicate an energy-efficiency of ~1.5 Pops/J and a throughput of 2.5 Tops/s for a 4-bit 200 × 200 VMM engine utilizing the proposed approach.

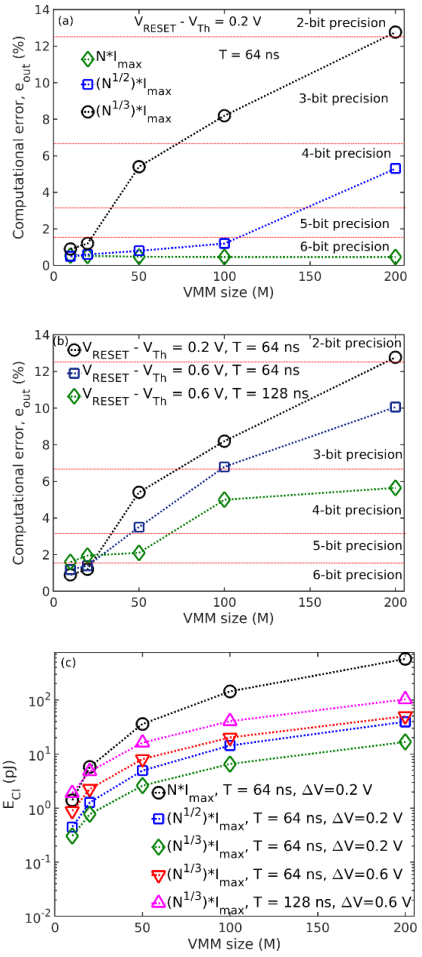

FIGURE 7. Computational error of the proposed VMM for different maximum output currents with (a) fixed output swing ($V_{RESET} - V_{TH} = 0.2$ V), (b) computational error when the maximum current is $M^{1/3} \times I_{max}$ for different output swings ($\Delta V = V_{RESET} - V_{TH}$) and time windows ($T$), and (c) dominant energy dissipated in the load capacitor for all the cases presented in parts (a) and (b).

Although applications such as inference, classification, recognition, etc. may be performed utilizing even low precision (~4 bits) VMM engines [8], [22] without significantly degrading the accuracy, we also analyze the efficacy of the proposed approach for binary-to-moderate (2 to 6 bits) target precisions under a different operating condition ($V_{GS} = 0.5$ V) as shown in Fig. 6. An increase in the targeted precision effectively translates into a larger time window (T) to encode the inputs while operating at the same frequency. Therefore, the capacitor and I/O circuit area and energy consumption increase significantly with an increased target precision. This leads to a considerable degradation in the area and energy efficiency when targeting moderate precision (>4 bits) VMM operations as shown in Fig. 6. However, the proposed VMM still exhibits a significantly high energy-efficiency of ~123.1 Tops/J and a throughput of 0.63 Tops/s for 6-bit 200 × 200 VMM operation.

Moreover, we may utilize RRAMs with higher ON-state resistance to reduce I_{max} further and decrease the load capacitance C_j for enhancing the area- and energy-efficiency. Similarly, a lower reset voltage (V_{RESET}) may further enhance

TABLE 2. Performance benchmarking.

Reference	[3]	[6]	[7]	[8]	[10]	[11]	[14]	[20]	[21]	This work
Approach	CM	CM	CM	SC	TD	TD	TD	TD	TD	TD
Process(nm)	180	22	180	40	14	250	55	55	55	55
Precision (bits)	3	~4	~5	3	<8	~1/5*	~6	~5	~4	6
EE(Tops/J)	6.4	60	5.7	8	18	<290	85	91	1305	1496
I/O included	Yes	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes
Results	Sim	Sim	Exp	Exp	Sim	Sim	Sim	Sim	Sim	Sim

CM: current-mode SC: switch-capacitor TD: time-domain *Binary weights/analog output

the VMM performance metrics. The capacitor area may also be reduced by using a different input encoding scheme whereby the individual input bits are encoded as discrete binary pulses and employing successive integration and rescaling technique to reduce the charge integrated on the load capacitor at the cost of reduced computational precision [21].

VI. DESIGN MODIFICATION FOR LOWER OUTPUT CURRENTS

In the proposed VMM design, for a VMM size $= M$, the maximum current which can be integrated at the load capacitor is $M \times I_{\max}$, when all the inputs and weights are maximum (all current sinks are programed to $I_{\max}$). Moreover, from a digital circuit perspective, such a scenario is equivalent to rounding the full precision, that is, $2P+\log_2 M$ bit-long VMM output (obtained by multiplying M-numbers with P-bit precision in the digital domain) to the most significant P bits where P is the precision of the proposed VMM. However, in some neural networks and several other applications, all the inputs and weights do not attain their maximum value during the operation and the maximum VMM output current is significantly lower than $M \times I_{\max}$. Therefore, the VMM design may be further modified according to the expected maximum dot-product value (which translates to maximum allowable output current in the proposed VMM). Such a modification would not only lead to a reduction in the rounding/quantization error but also facilitate the utilization of a smaller load capacitor to integrate the relatively lower output current.

However, a reduction in the maximum output current and the load capacitor leads to an increase in the computational error and significantly degrades the precision of the VMM as shown in Fig. 7(a). The computational precision degrades to 4 bits when the maximum VMM output current is limited to $M^{1/2} \times I_{\max}$ (which is equivalent to extracting P-bit VMM output from center of the $2P+\log_2 M$ bit-long digital output) and to 2 bits when the maximum VMM output current is reduced to $M^{1/3} \times I_{\max}$.

Apart from introducing a multiplicative coefficient (a) and the input-dependent additive coefficient (b) in the output, the appreciable minimum current of the 1T-1R cells also consumes a significant portion of the output voltage swing. Therefore, the part of the total output swing available for performing useful computation is also low for the proposed VMM due to the nonnegligible $I_{\min}$. Moreover, a reduction in the maximum output current simply implies that the part of the output voltage swing dominated by the minimum cell currents would be even larger. Furthermore, a reduction in the load capacitor may lead to a charge disturbance error [20] owing to the increased coupling (and hence, charge sharing) between the gate–drain capacitance of the MOSFET and the load capacitor. As a result, the useful portion of the available output swing is further degraded by the charge disturbance error. This leads to a significant reduction in the VMM output precision observed in Fig. 7(a).

Although increasing the output voltage swing may appear as a straightforward technique to reduce the output error and improve the precision, it may lead to a degradation in the computational precision when the DIBL/CLM error is high as shown in Fig. 4 (for input voltage $V_{GS} = 0.3$ V). However, when an input voltage, $V_{GS} = 0.5$ V is utilized, DIBL/CLM error is minimized (see Fig. 3). In this case, an increase in the output swing enhances the useful portion available for computation leading to an improved computational precision. Moreover, an increase in the output voltage swing further minimizes the load capacitor [refer (7)]. This leads to a reduction in the energy dissipated in the load capacitor despite an increase in the voltage swing as shown in Fig. 7(c). However, the computational precision is only 3 bits for the maximum output current of $M^{1/3} \times I_{\max}$, even when the output swing is increased to the maximum permissible read voltage, which is not useful for applications such as DNNs. To increase the computational precision further (to 4 bits which is appropriate for applications such as DNNs), a larger time window (T) should be utilized as shown in Fig. 7(b). However, a larger time window not only reduces the operating frequency but also increases the size of the load capacitor [refer (7)] and its energy dissipation [see Fig. 7(c)]. Therefore, there is an inherent tradeoff between the computational precision and the energy dissipated in the load capacitor even when the maximum output current is limited to $M^{1/3} \times I_{\max}$.

VII. CONCLUSION

An energy-efficient time-domain VMM exploiting a modified configuration of 1T-1R array has been proposed in this article. The dominant mechanisms, such as CLM and DIBL, and the nonnegligible minimum cell currents which degrade the performance of the proposed architecture are discussed in detail. Furthermore, we show that there exists a tradeoff between the computational precision, DR and the area- and energy-efficiency of the proposed VMM approach. Therefore, we also provide necessary design guidelines to further optimize the performance of the 1T-1R VMM. The preliminary results indicate that an effective computational precision of 6 bits and a significantly high energy efficiency of ~1.5 Pops/J and a throughput of 2.5 Tops/s as compared to the other VMM approaches (see Table 2) is achievable for a 4-bit 200 × 200 VMM using the proposed approach. Our results may provide an incentive for the experimental realization of the VMM approach based on the 1T-1R array.

REFERENCES

[1] Y. LeCun *et al.*, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for IoT big data and streaming analytics: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2923–2960, 4th Quart., 2018.

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

[4] T. Tohara *et al.*, "Silicon nanodisk array with a fin field-effect transistor for time-domain weighted sum calculation toward massively parallel spiking neural networks," *Appl. Phys. Express*, vol. 9, no. 3, Mar. 2016, Art. no. 034201.

[5] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element," in *IEDM Tech. Dig.*, Dec. 2014, pp. 29.5.1–29.5.4.

[6] M. Hu *et al.*, "Dot-product engine for Neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication," in *Proc. 53rd ACM/EDAC/IEEE Design Automat. Conf. (DAC)*, Austin, TX, USA, Jun. 2016, pp. 1–6.

[7] X. Guo *et al.*, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in *IEDM Tech. Dig.*, Dec. 2017, pp. 6–11.

[8] E. H. Lee and S. S. Wong, "Analysis and design of a passive switched-capacitor matrix multiplier for approximate computing," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 261–271, Jan. 2017.

[9] D. Miyashita, S. Kousai, T. Suzuki, and J. Deguchi, "A Neuromorphic chip optimized for deep learning and CMOS technology with time-domain analog and digital mixed-signal processing," *IEEE J. Solid-State Circuits*, vol. 52, no. 10, pp. 2679–2689, Oct. 2017.

[10] M. J. Marinella *et al.*, "Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 1, pp. 86–101, Mar. 2018.

[11] Q. Wang, H. Tamukoh, and T. Morie, "A time-domain analog weighted-sum calculation model for extremely low power VLSI implementation of multi-layer neural networks," 2018, *arXiv:1810.06819*. [Online]. Available: http://arxiv.org/abs/1810.06819

[12] T. Morie *et al.*, "Spike-based time-domain weighted-sum calculation using nanodevices for low power operation," in *Proc. IEEE 16th Int. Conf. Nanotechnol. (IEEE-NANO)*, Aug. 2016, pp. 390–392.

[13] 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, *arXiv:1902.07707*. [Online]. Available: http://arxiv.org/abs/1902.07707

[14] M. Bavandpour *et al.*, "Mixed-signal Neuromorphic inference accelerators: Recent results and future prospects," in *IEDM Tech. Dig.*, Dec. 2018, pp. 20–24.

[15] M. Bavandpour, M. Reza Mahmoodi, and D. B. Strukov, "Energy-efficient time-domain Vector-by-Matrix multiplier for neurocomputing and beyond," *IEEE Trans. Circuits Syst. II: Exp. Briefs*, vol. 66, no. 9, pp. 1512–1516, Sep. 2019, doi: 10.1109/TCSII.2019.2891688.

[16] B. Chakrabarti *et al.*, "A multiply-add engine with monolithically integrated 3D memristor crossbar/CMOS hybrid circuit," *Sci. Rep.*, vol. 7, no. 1, p. 42429, Mar. 2017.

[17] *HSPICE User Guide: Basic Simulation and Analysis*, Synopsys, Mountain View, CA, USA, 2018.

[18] B. Li, P. Gu, Y. Shan, Y. Wang, Y. Chen, and H. Yang, "RRAM-based analog approximate computing," *IEEE Trans. Comput.-Aided Design Integr.*, vol. 34, no. 12, pp. 1905–1917, Dec. 2015.

[19] F. Alibart, L. Gao, B. D. Hoskins, and D. B. Strukov, "High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm," *Nanotechnology*, vol. 23, no. 7, Feb. 2012, Art. no. 075201.

[20] M. Bavandpour, S. Sahay, M. Reza Mahmoodi, and D. B. Strukov, "3D-aCortex: An ultra-compact energy-efficient neurocomputing platform based on commercial 3D-NAND flash memories," 2019, *arXiv:1908.02472*. [Online]. Available: http://arxiv.org/abs/1908.02472

[21] M. Bavandpour, S. Sahay, M. R. Mahmoodi, and D. Strukov, "Efficient mixed-signal neurocomputing via successive integration and rescaling," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 3, pp. 823–827, Mar. 2020, doi: 10.1109/TVLSI.2019.2946516.

[22] T.-J. Yang and V. Sze, "Design considerations for efficient deep neural networks on Processing-in-Memory accelerators," in *IEDM Tech. Dig.*, Dec. 2019, pp. 1–4.

SHUBHAM SAHAY (Member, IEEE) received the B.Tech. degree (Hons.) in electronics engineering from IIT (BHU) Varanasi, Varanasi, India, in 2014, and the Ph.D. degree in electrical engineering from IIT Delhi, New Delhi, India, in 2018.

He worked as a Post-Doctoral Research Scholar at the University of California, Santa Barbara, CA, USA. He is currently an Assistant Professor with the Department of Electrical Engineering, IIT Kanpur, Kanpur, India. He has authored a book *Junctionless Field-Effect Transistors: Design, Modeling and Simulation* (IEEE-Wiley press) and several peer-reviewed articles on topics, including semiconductor device design and modeling, neuromorphic computing, and hardware security primitives utilizing emerging non-volatile memories.

Dr. Sahay was awarded four gold medals for obtaining the highest marks and securing the first position in B.Tech. at IIT (BHU) Varanasi.

MOHAMMAD BAVANDPOUR received the M.Sc. degree in electrical engineering (digital electronics) from the Sharif University of Technology, Tehran, Iran, in 2012, and the Ph.D. degree in computer engineering from the University of California, Santa Barbara (UCSB), Santa Barbara, CA, USA, in 2020.

During his Ph.D. program, he worked on the efficient mixed-signal implementation of neural circuits/accelerators using non-volatile memory devices. He is currently a Staff Research Associate with ECE Department, UCSB. His main research interests are digital/mixed-signal computing circuits, neural accelerator circuits/architectures, non-volatile memory devices, parallel processing, and machine learning.

MOHAMMAD REZA MAHMOODI , photograph and biography not available at the time of publication.

DMITRI STRUKOV (Senior Member, IEEE) received the M.S. degree in applied physics and mathematics from the Moscow Institute of Physics and Technology, Dolgoprudny, Russia, in 1999, and the Ph.D. degree in electrical engineering from Stony Brook University, Stony Brook, NY, USA, in 2006.

He worked as a Post-Doctoral Associate, first at Stony Brook University from August 2006 to December 2007, and then at Hewlett Packard Laboratories from January 2007 to June 2009, on various aspects of nanoelectronic devices and systems. He is currently a Professor of electrical and computer engineering at the University of California, Santa Barbara (UCSB), Santa Barbara, CA, USA. His research broadly concerns different aspects of computation, in particular addressing questions on how to efficiently perform computation on various levels of abstraction. The research spans across different disciplines, including material science, device physics, circuit design, high-level computer architecture, and algorithms, with an emphasis on emerging device technologies. Over the past decade, his major focus was on neuromorphic computing, and more recently on hardware security implementations with resistive switching devices ("memristors") and floating gate memories.

Dr. Strukov is a member of the ACM and IEEE societies.