An Ultra-Low Energy Internally Analog, Externally Digital Vector-Matrix Multiplier Based on NOR Flash Memory Technology

M. Reza Mahmoodi and Dmitri Strukov
ECE Department, UC Santa Barbara
Santa Barbara, CA, 93106-5630, USA
mrmahmoodi@ucsb.edu, strukov@ucsb.edu

ABSTRACT

Vector-matrix multiplication (VMM) is a core operation in many signal and data processing algorithms. Previous work showed that analog multipliers based on nonvolatile memories have superior energy efficiency as compared to digital counterparts at low-to-medium computing precision. In this paper, we propose extremely energy efficient analog mode VMM circuit with digital input/output interface and configurable precision. Similar to some previous work, the computation is performed by gate-coupled circuit utilizing embedded floating gate (FG) memories. The main novelty of our approach is an ultra-low power sensing circuitry, which is designed based on translinear Gilbert cell in topological combination with a floating resistor and a low-gain amplifier. Additionally, the digital-to-analog input conversion is merged with VMM, while current-mode algorithmic analog-to-digital circuit is employed at the circuit backend. Such implementations of conversion and sensing allow for circuit operation entirely in a current domain, resulting in high performance and energy efficiency. For example, post-layout simulation results for 400×400 5-bit VMM circuit designed in 55 nm process with embedded NOR flash memory, show up to 400 MHz operation, 1.68 POps/J energy efficiency, and 39.45 TOps/mm^2 computing throughput. Moreover, the circuit is robust against process-voltage-temperature variations, in part due to inclusion of additional FG cells that are utilized for offset compensation.

KEYWORDS

Analog Computing; Vector-by-Matrix Multiplier; Floating-Gate Memory

1 INTRODUCTION

Analog computing circuits, in particular those implementing low-to-medium precision VMM [1,2], the most common operation in signal and data processing algorithms [3], have been shown to be extremely energy efficient [4-5]. An internally analog, externally digital VMM circuit offers the best of both worlds: The density and energy efficiency of an analog domain, and the noise-robustness and versatility of a digital communication [6]. Accordingly, mixed-signal VMMs have been realized in variety of applications including neural networks [7,8], support vector machines [9], and IoT systems [10]. Some of the most prospective proposals are based on emerging nonvolatile memory (NVMs) [11,12].

Time-based VMMs [13] and switch-capacitor multipliers [1,14] use charge to encode data. The former approach, designed to operate in very low voltages, is based on charge integration from digitally programmable current sources. One of the challenges is process-voltage-temperature (PVT) variations that may limit the smallest integration delay and hence the circuit performance. For the latter case, metal fringing capacitors have been exploited to build VMM circuits with moderate computing precision. These topologies have been explored for implementing (> 4 bit) multipliers using bulky and power hungry active amplifiers. In the passive version of such circuits, amplifier is eliminated [1], which can lead to potentially more power efficient and faster design. The main challenges, however, are leakage, capacitive coupling and charge injection issues, which confine passive switch-capacitor approaches to 2-3 bit resolutions.

In another approach, current/voltage is employed as a state variable. For example, VMM circuit with digitally controllable single MOS-based current sources, in which width of the transistors were scaled according to the predetermined weights, has demonstrated very high energy efficiency [15]. The main caveat of such design is an area (and hence energy) overhead for

DAC '18, June 24–29, 2018, San Francisco, CA, USA

ACM ISBN 978-1-4503-5700-5/18/06 $15.00
https://doi.org/10.1145/3195970.3195989

weight implementation, which exponentially increases with weight precision.

A more promising solution is to implement matrix weights with NVMs, such as programmable conductance cross-point devices and FG memories. The most prospective VMM circuits are perhaps based on metal-oxide memristors [11,12] due to the excellent scalability, analog properties, and non-volatility of such devices. Yet, memristor fabrication technology is not mature enough for very large scale integration and hence some of the research is now focused on more mature but less dense NVMs, such as FG memory [4,7,9,16]. For example, a number of VMM circuits were recently experimentally demonstrated using commercially available NOR flash memory [8, 17, 18], whose matrix structure was modified to allow for individual tuning of FG cells' conductances [18]. Though the modification tripled the cell area, the memory density was still more than an order of magnitude better as compared to previous FG memories utilized in analog circuits [4].

The general architecture of a digital-input digital-output (DIDO) VMM circuit is shown in Fig. 1. The circuit computes in parallel M P_o-bit dot-products between N-element P_{in}-bit input vector and corresponding N-element vector of P_w-bit weights. Note that, in general, the precision of dot-product computation might be higher compared to that of analog-to-digital (ADC) converter.

The efficiency of similar, previously proposed VMM circuits was greatly limited by the overhead of sensing circuitry and data converters. For example, the power per channel (VMM output) reported in Ref. [12] was nearly 100 µW and peripherals consumed > 90% of power and occupy > 95% of chip area.

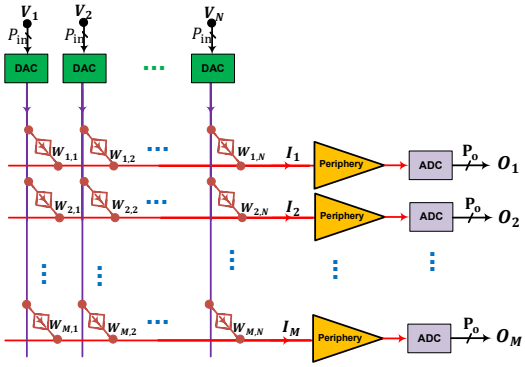

Figure 1: A general idea of $M\times N$ DIDO VMM circuit. In FG memory implementation, the weights are encoded by the cell's subthreshold currents.

2 VMM CIRCUIT

2.1 Top-Level Architecture

In our design, the aforementioned issues in mixed-signal VMM circuits are resolved by utilizing several features, including very efficient peripheral circuitry, merged digital-to-analog (DAC) implementation, algorithmic ADC converter, and additional columns of FG cells in the array to cope with process variations. The combination of these techniques allows implementing all operations in VMM completely in current domain, which greatly increase computational bandwidth and energy efficiency.

Specifically, the top level architecture of the proposed VMM circuit is shown in Fig. 2. In this architecture, data are buffered into a shift register to hold it during the processing, which is triggered by φ_1 control signal. Upon completion of the data transfer, digital voltages are applied to the array to generate currents in each channel proportional to the dot-product of input and weight vectors. To reduce conversion overhead, a merged DAC (MDAC) architecture is employed at the input interface. In this case, each matrix weight W_{ji} in the original scheme (Fig. 1) is implemented with a set of P_{in} FG devices, i.e.

$$W_{ji}^{k} = 2^{k}/(2^{P_{in}} - 1)\, W_{ji}, \quad P_{in} \geq k \geq 1 ,$$

where k is input bit significance. Assuming that i-th input is binary vector $\{b_{P_{in}}, \ldots, b_1\}$, a current injected by the memory cells implementing weight W_{ji} to the j-th output is given by:

$$I_{ji} = \sum_{k=1}^{P_{in}} b_k\, 2^{k}/(2^{P_{in}} - 1)\, W_{ji}$$

Negative weights with FG memory devices are implemented using differential pair of weights $W_{ji} = W_{ji}^{+} - W_{ji}^{-}$, so that for two quadrant VMM implementation, the total current in the j-th differential output is given by

$$I_j = \sum_{i=1}^{N} \left(I_{ji}^{+} - I_{ji}^{-}\right) .$$

Naturally, the proposed MDAC implementation is based on VMM circuit and specific tuning of the weights. Therefore, MDAC's area and energy are simply contributed by the additional $2\times M\times(P_{in}\text{-}1)$ array of FG cells.

In the following sections, we discuss in detail VMM components, i.e. FG memory array, sensing circuit, and ADC.

2.2 Floating-Gate Memory Array

FG memory array was implemented using split-gate ESF3 SuperFlash®, which is commercialized embedded NOR flash technology developed by SST [19]. (Details on this technology and various device characteristics are reported in Refs. [17,19]). The ESF3 flash memory is very desirable for realization of couple-gate arrays [16,17]. For example, due to its split-gate structure, FG memory cells offer very high output impedance, of the order of 100 GΩ in subthreshold regime. The robust subthreshold operation in ESF3 devices are typically in 100 pA - 300 nA range. (In our design, the weights' least significant bit always corresponds to 500 pA.)

Maximum achievable P_w depends on state drift, tuning accuracy and virtual bias variations. In order to have 5-bit effective precision, the signal-to-noise and distortion ratio of the device should be > 40.9 dB, which roughly corresponds to 0.9% weight error, i.e. the normalized difference between the desired and actual subthreshold currents of the cell. It is reasonable to assume that the tuning accuracy (which could be improved with increasing write time) and drift can be bounded within 0.4% [17] and hence the error due to maximum sustainable bias variation distortion should be < 0.5%. The virtual bias variations impact the absolute value of the weight via channel length modulation and

drain induced barrier lowering. For the utilized range, 0.5% crudely translates into ΔV_b = 10 mV.

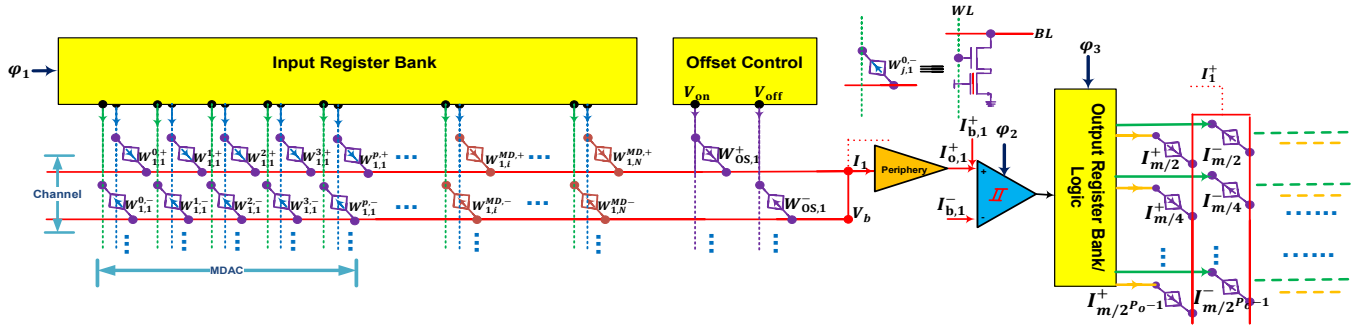

Figure 2: One channel of the proposed two-quadrant VMM circuit with digital inputs and outputs. Here we assume that inputs and outputs are non-negative, while weights can be negative or positive.

2.3 Sensing Circuit

As discussed in previous section, sensing circuit must provide precise virtual voltage V_b on shared bit (i.e. horizontal on Fig. 2) lines. In previous works, this condition was enforced by using transimpedance amplifiers (TIA) and integrators [12,17]. TIAs, however, typically consume large area and are optimized to work at a certain operating point rather than dealing with a large amplitude signals. The very limited settling time of TIAs also mandates large biasing currents. Here we proposed a circuit design in which V_b variations are bounded with minimum overhead. Our design is also very efficient for controlling PVT variations.

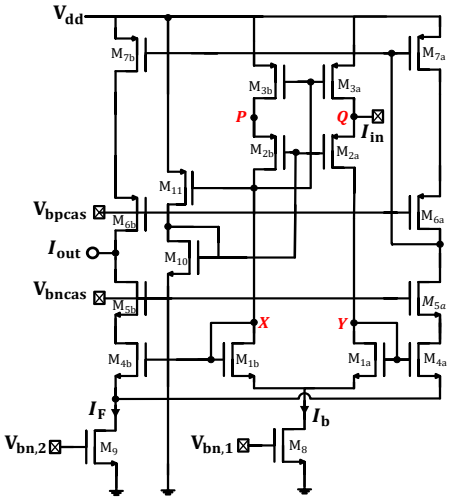

Figure 3: The sensing circuit.

The sensing circuit is shown in Fig. 3. The circuit could be viewed as a floating resistor ($M_{1\text{-}3}$) followed by a low-gain amplifier $M_{4\text{-}7}$. M_1, M_2 and M_4-pairs are designed in weak inversion and M_3-pair is velocity saturated. Rest of the devices are in saturation regime. The translinear loop, constructed by $M_{1,4}$-pairs have excellent wideband current following behavior. The current drawn from node "Q" is supplied by M_{3a}. The larger such current, the smaller is I_{2a}. The resulting differential voltage generated at XY node is then converted to an output current I_{out} by the low gain amplifier.

The transfer characteristics of the circuit is given by

$$I_{out} = \left(\frac{I_F}{I_b}\right) I_{in},$$

while virtual bias swing without the local feedback, formed by M_{11} and M_{10}, is $n V_T \ln\left(1 - \frac{(I_{in})_{max}}{I_b}\right)$. The negative feedback compensates the drop by pulling $V_{G,M2a}$ down and pinning $V_{S,M2a}$. The proper sizing of M_{11} and adjustment of bias current I_b allows reducing ΔV_b to 3 mV (Fig. 4a), which ensures 5-bit weight precision.

Both deterministic and random non-idealities result in offset and distortion. The offset, originated from mismatch in $M_{1,3,4}$-pairs, is compensated by adding two additional columns of FG cell (Fig. 2) and tuning their conductances according to the total input-referred offset of the corresponding channel. Such approach relaxes other design specifications without a considerable power/area overhead. The mismatch between drain current of M_{3a}, M_{3b} and threshold voltage of $M_{1\text{-}4}$-pairs impacts the linearity of sensing circuit. As shown in Fig. 4a, both mean and standard deviation (SD) of relative nonlinearity error are reduced dramatically by slightly increasing the bias current. (Nonlinearity error could be further reduced, by a factor of ~4, when implementing advanced layout techniques.)

To keep ΔV_b below a desired value at all temperatures, I_b is designed using a proportional to absolute temperature current source. Additionally, to keep the slope of transfer function invariant to temperature variations, I_F is also supplied by the same source, allowing to limit slope variations to < 0.2% (Fig. 4b).

The relatively large, ±4% fluctuations in supply voltage result in only < 0.5% variation of the slope (Fig. 4c). This is because slope only depends on bias currents, and as long as critical transistors remain in their targeted operating region, degradation in linearity is negligible. In general, a desired weight and computing precision determines the minimum transistor scaling, and, in particular, the smallest I_b, and capacitances C_x, and C_y. With these values fixed, the settling time, and, as result, energy consumption, can be further optimized by finding optimal output pole location. The

output pole can be relocated by adjusting output current, e.g. by changing I_F. For a certain translinear loop size, initially, increasing I_F improves the settling time (Fig. 4d). However, at some point, the overshoot in time response becomes excessive and deteriorate phase and settling time. Increasing output current is not helpful anymore since the dominant pole is no longer attributed to the output pole.

Finally, let us note that the dominant noise power is due to random telegraph noise (RTN) of FG cells, while the peripheral noise is of much less importance.

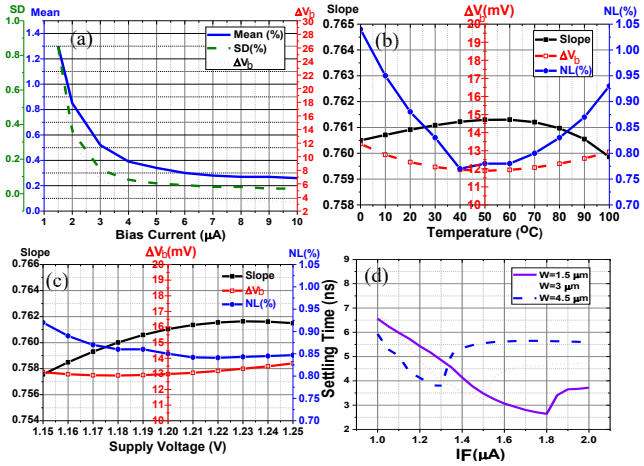

Figure 4: Sensing circuit simulation results for $(I_{in})_{max}$ = 1 μA. (a) Worst-case nonlinearity error, which accounts for process variations, as a function of I_b. The impact of (b) temperature and (c) supply voltage on ΔV_b, slope of transfer function, and total worst-case nonlinearity error. (d) Settling time as a function of loop size and I_b.

2.3 ADC Design

Algorithmic ADCs feature high resolution, throughput, and small area. Among such architectures, conventional current-mode ADCs typically offer the best speed-area performance [21]. In our work, we use current-mode cyclic ADC to minimize the conversion cost and, more importantly, to leverage tunability of FG cells for precise current generation. Specifically, a 1-bit per stage cyclic current-mode ADC was implemented. Note that we have not used the common 1.5-bit per stage design since it has a significant power overhead. Instead, comparator's dynamic offset was compensated by adjusting input bias currents I_{BA} for each channel using FG cells. The bias currents, although contribute to power consumption, are critical to support a bipolar output and keep the mirror devices turned always on, which facilitates faster conversion. The constant current sources are generated by an auxiliary MDAC arrays of FG devices, which share bit lines with the main array (Fig. 5).

The operation is performed in a sequence of P_o steps. In the first step, current comparator determines whether the input current (I_{in}), fed by sensing circuitry, is positive or not, and generates a sign bit. In the next cycle, based on the sign bit, $I_{max}/2$ is either subtracted from or added to input current, where I_{max} is the maximum possible amplitude of ADC input current. At the end of k-th step, residual current is given by:

$$I_{res} = I_{in} + \Sigma_{l=1}^{k-1}(-1)^{D_{Po-l+1}}(I_{max}/2^l),\ k>1$$

where D_l represents l–th output bit. The process is repeated until LSB (D_1) is generated. Then, $D_{<Po:1>}$ is buffered to a parallel register. The operation needs minimum control and is shared between all channels, and hence can have very compact implementation. The controller is essentially a simple logic and a shift register, which is cleared at the end of each conversion and shifts logic "1" at each conversion step.

The comparator design is shown in Fig. 6. The circuit utilizes a cascode current mirror as a preamplifier and a latch stage similar to that of StrongARM. At the beginning of each step, when φ_2 or φ_3 is high, nodes X, Y, P and Q are precharged to ground, which is typically referred as shielding mode. The purpose of shielding is to reset the state of the comparator and avoid storing excess charge on node X right after conversion, which may happen due to peripheral circuit delay. Shielding continues until I_{6b} restores to I_b^+. In the following, both φ_2 and φ_3 go down, while the current I_{6b} + I_{com}, where I_{com} = (I_{M4a} + I_{M4b}) /2, charges node X. The circuit operation is similar for V_Y, with I_b^- + I_{com} charging node Y. (Here, I_{com} is used to inject a dynamic common-mode current and quickly turn $M_{1a,b}$ on.) When V_X reaches the threshold of transistor M_{1a}, φ_3 goes high and regeneration begins.

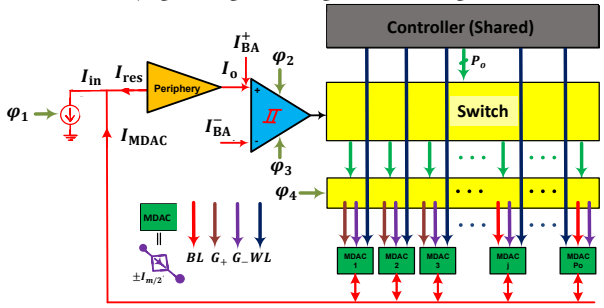

Figure 5: Block diagram of the algorithmic ADC.

Finally, cross-coupled transistors turn on and the differential current, amplified by the positive feedback loop, brings one of the outputs to V_{DD}. In Fig. 7, the input current is 700 nA and $(I_{in})_{max}$ = 1 μA. Since input current is positive at first, Q becomes "1" after comparison is finalized and 500 nA is subtracted in the following step. The process continues until LSB is generated. The proposed circuit, though leverages currents for signal representation, is not impacted by the device matching and charge injection issues. This is due to unique features of FG cells which are exploited in performing multiplication, offset compensation, and generation of constant scaled current sources. The high performance, achieved in proposed ADC, stems from three factors: Low-overhead offset compensation, which relaxes the trade-off between speed and resolution, embedded design of current references with zero power overhead, and low-power design of dynamic current comparator. For example, the comparator settles at 0.65 ns for 30 nA differential input current, while dissipating only 2.07 μW dynamic power on average. It should be noted that at high precisions, clocking scheme of cyclic ADC was redesigned to maintain its energy efficiency [23]. Similarly to sensing circuit,

the input referred current noise of the comparator is much smaller than the RTN noise associated with FG cells, and can be neglected.

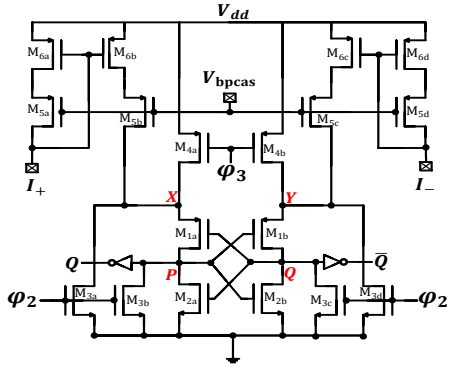

Figure 6: Dynamic current comparator circuit.

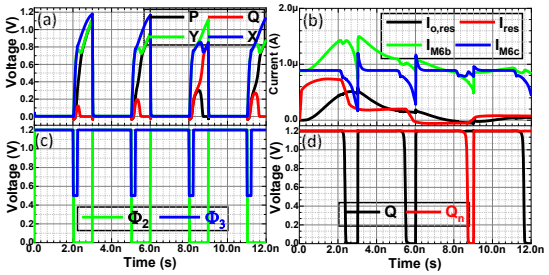

Figure 7: Timing diagram of the ADC (shown for 4 cycles): (a) transient voltage of nodes X, Y, P, and Q, (b) transient residual currents and drain currents of $M_{6c,6b}$, (c) clocking scheme, and (d) corresponding digital outputs.

3 RESULTS AND DISCUSSION

The proposed DIDO VMM circuit was implemented in Global Foundry's 55-nm LPe 2P8M process technology. The design was optimized with respect to energy efficiency. The dynamic power (in comparator and array), and the static power (in peripheral circuitry and array) were both included in power consumption estimates. The same precision for inputs and weights, which is limited to ~ 5-bit as discussed above, were always assumed. From the experimental results [8,17,18], the compute (output) precision is typically limited by RTN in FG cells and hence increases with *N*. For 55 nm technology, 8-bit output precision is achievable for $N > 25$.

Simulation results show that the circuit area grows rapidly as a function of input/weight/output precision (Fig. 8a) because of the merged-DAC overhead. The same trends are observed in the settling time and energy due to the cyclic structure of the ADC. To preserve tolerance to process variations, the sensing circuit cannot be scaled down efficiently at very low input currents (e.g. at < 3 bit precision), which explains the trend for delay and energy consumption. Throughput (TH) decreases as expected because the same number of operations are performed slower. For the same reasons, energy efficiency (EE) and area efficiency (AE) gradually decrease as precision increases. On the other hand, with input/weight precision fixed at 5-bits, the total active area does not change much with output precision since ADC has negligible area overhead (Fig. 8b).

The number of operations grows quadratically as a function of VMM size, and so does the total active area (Fig. 8c). As mentioned before, at very low currents (smaller size VMMs), sensing circuit is slower. Because of that, TH is increasing roughly quadratically with VMM size. Though the total energy consumption is increasing with VMM size, the EE is also increasing because of TH and is saturating at ~1.8 POps/J for $N > 500$.

Fig. 9a shows energy breakdown for several VMM circuit implementations. Peripheral circuitry and ADCs are typically the major source of energy consumption. ADC's power consumption is ~6 μW per channel and almost the same for all designs. The first, relatively small VMM circuit is designed at 4-bit, and hence the array and sensing power are less than power consumed in comparator. For larger precision and large size VMM circuits, i.e. the second and third considered cases, respectively, the contribution of sensing circuitry becomes more prominent.

The area breakdown is provided in Fig. 9b. Note that 10% is added to each block to account for routing among the blocks. FG array dominates the area for large VMM circuits. For smaller ones, the contribution of programing/erasing circuitry is almost equal to that of array size. However, based on our previous experience [8], the overhead would be insignificant when it is shared between multiple blocks (and hence was not neglected in Fig. 8.) Finally, chip prototype of a 4-bit 64×64 DIDO VMM circuit, fabricated in GF's 55 nm process, is provided in Fig. 9c.

The performance metrics of our design compares very favorably with the best reported results. For example, Ref. [12] reports the ReRAM-based dot-product engine with 30 TOps/J maximum energy efficiency for 128×128 crossbar circuit. Switch-capacitor VMM circuits, proposed in [1] and [14], are the state-of-the-art low precision multipliers based on conventional technology. The former reference reports a serial 6b/3b/6b 40 nm VMM circuit, which achieves 7.72 TOps/J in 0.012 mm^2, while the latter uses the same approach on 8b/14b/8b 16-parallel channels at 28 nm and reaches 9.61 TOps/J in ~0.011 mm^2. For comparison, the proposed approach achieves 1.68 POps/J for 400×400 VMM circuit when computation, I/O, and weights are all at 5-bit precision. This number is ~ 100× better than that of state-of-the-art switch-capacitor ASIC designs.

In principle, for both switch-capacitor designs, weights can be programmed quickly, making it suitable for larger range of applications, as compared with the proposed design. However, this advantage often comes with the cost of bandwidth limitations in large scale systems. For example, in Ref. [22], FG memory based circuits were fabricated in 130 nm to realize a deep neural network featuring 7.2 bit weight precision. Though the system is fully analog, it only achieved ~1 TOps/J, while occupying 0.36 mm^2.

4 CONCLUSIONS

Earlier work has shown that mixed-signal VMM circuits based on nonvolatile memories could greatly surpass their digital counterparts in energy and area efficiency. The maximum achievable performance in previously reported mixed-signal

implementations has been limited by the peripheral circuits, including those used for conversion between analog and digital domains. In this paper, we propose novel design of mixed-signal VMM circuit with all its parts implemented in current domain. The very high energy and area efficiency of the proposed design stems from three factors. First, it is due to very compact, optimized, reliable, and low-power ESF3 technology. Second, we propose efficient sensing circuitry and compensation of process variations by fine tuning FG memory, which relax design requirements for high bandwidth current processing. Finally, it is due to the considered algorithmic ADC design in which FG memory is used to generate very precise current sources. Our simulation results show that 400×400 5-bit VMM implemented in a 55 nm CMOS technology with embedded NOR flash memory achieves record-breaking 1.68 POps/J energy efficiency and 39.45 TOps/mm^2 computing throughput.

REFERENCES

[1] Edward Lee and Simon Wong. 2017. Analysis and design of a passive switched-capacitor matrix multiplier for approximate computing. *IEEE Journal of Solid-State Circuits*. 52(1). 261-271.

[2] Lita Yang and Boris Murmann. 2017. Approximate SRAM for energy-efficient, privacy-preserving convolutional neural networks. *IEEE Computer Society Annual Symposium on VLSI (*ISVLSI*)*. Bochum. Germany. 689-694.

[3] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. 2015. BinaryConnect: Training deep neural networks with binary weights during propagations. *Advances in Neural Information Processing Systems*. Montreal. Canada. 3123-3131.

[4] Jennifer Hasler and Bo Marr. 2013. Finding a roadmap to achieve large neuromorphic hardware systems. *Frontiers in Neuroscience*. 10(7). 113.

[5] Brian Degnan, Bo Marr, and Jennifer Hasler. 2016. Assessing trends in performance per watt for signal processing applications. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*. 24(1). 58-66.

[6] Roman Genov and Gert Cauwenberghs. 2001. Charge-mode parallel architecture for vector-matrix multiplication. *IEEE Transactions on Circuits and Systems II*: Analog and Digital Signal Processing. 48(10). 930-936.

[7] Laura Fick, David Blaauw, and Dennis Sylvester. 2017. Analog in-memory subthreshold deep neural network accelerator. *IEEE Custom Integrated Circuits Conference* (CICC). Austin. TX. 1-4.

[8] Xinjie Guo, et al. 2017. Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology. *International Electron Device Meeting* (IEDM). San Francisco. CA. 1-4.

[9] Shantanu Chakrabartty and Gert Cauwenberghs. 2007. Sub-microwatt analog VLSI trainable pattern classifier. *IEEE Journal of Solid-State Circuits*. 42(5). 1169-1179.

[10] Siddharth Joshi, Chul Kim, Sohmyung Ha, and Gert Gauwenberghs. 2017. From algorithms to devices: Enabling machine learning through ultra-low-power VLSI mixed-signal array processing. *IEEE Custom Integrated Circuits Conference* (CICC). Austin. TX. 1-4.

[11] Mirko Prezioso, et al. 2016. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature*. 521. 61-64.

[12] Miao Hu, et al. 2016. Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication. *Design Automation Conference* (DAC). Austin. TX. 1-6.

[13] Robert D'Angelo and Sameer Sonkusale. 2015. A time-mode translinear principle for nonlinear analog computation. IEEE Transactions on Circuits and Systems I: Regular Papers. 62(9). 2187-2195.

[14] Daniel Bankman and Boris Murmann. 2016. An 8-bit, 16 input, 3.2 pJ/Op switched-capacitor dot product circuit in 28-nm FDSOI CMOS. *Solid-State Circuits Conference* (A-SSCC). Toyama. Japan. 21-24.

[15] Jonathan Binas, et al. 2016. Precise deep neural network computation on imprecise low-power analog hardware. arXiv:1606.07786.

[16] Shubha Ramakrishnan and Jennifer Hasler. 2014. Vector-matrix multiply and winner-take-all as an analog classifier. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*. 22(2). 353-361.

[17] Xinjie Guo, et al. 2017. Temperature-insensitive analog vector-by-matrix multiplier based on 55-nm NOR flash memory cells. *IEEE Custom Integrated Circuits Conference* (CICC). Austin. TX. 1-4.

[18] Farnood Merrikh Bayat, et al. 2015. Redesigning commercial floating-gate memory for analog computing applications. *IEEE International Symposium on Circuits and Systems*. Lisbon. Portugal. 1921-1924.

[19] Nhan Do. 2018. Split-gate floating poly SuperFlash® memory technology, design, and reliability. Embedded Flash Memory for Embedded Systems: Technology, Design for Sub-systems, and Innovations. 131-178.

[20] David Narin and Andre Salama. 1990. Current-mode algorithmic analog-to-digital converters. *IEEE Journal of Solid-State Circuits*. 25(4). 997-1004.

[21] Min Kim and Pavan Hanumolu. 2009. A 10 MS/s 11-bit 0.19 mm^2 algorithmic ADC with improved clocking scheme. *IEEE Journal of Solid-State Circuits*. 44(9). 2348-2355.

[22] Junjie Lu, Steven Young, Itamar Arel, and Jeremy Holleman. 2015. A 1 TOps/W analog deep machine-learning engine with floating-gate storage in 0.13 µm CMOS. *IEEE Journal of Solid-State Circuits*. 50(1). 270-281.

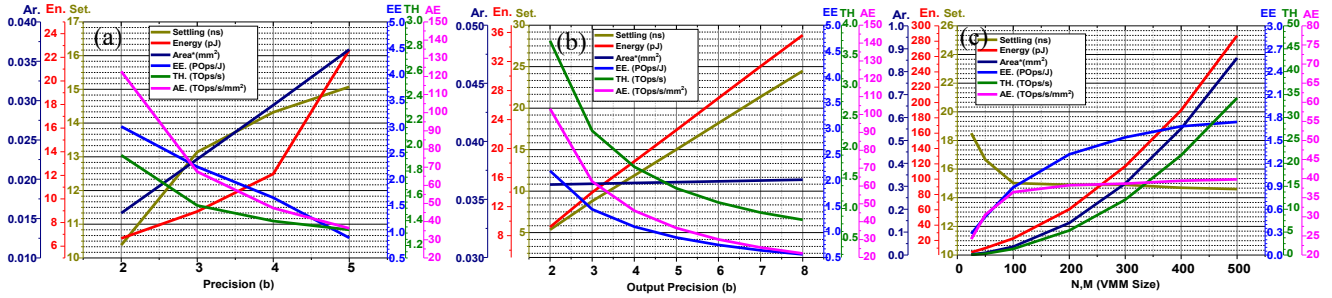

Figure 8: **Various performance metrics of DIDO VMM circuit as a function of (a) precision assuming P_i=P_w=P_o and M=N=100, (b) output precision P_o, assuming P_i=P_w=5 and M=N=100, and (c) VMM dimensions, assuming P_i=P_w=P_o=5.**

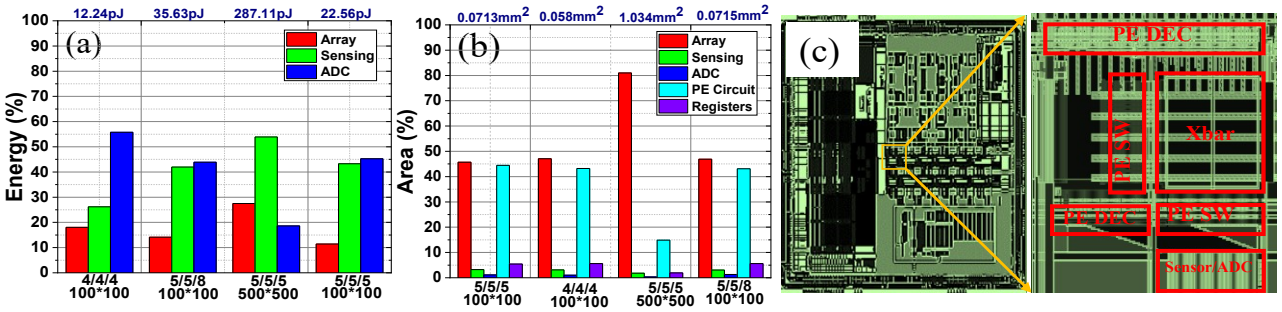

Figure 9: **(a) Energy and (b) area breakdown of 4 different VMM implementations. (c) Chip prototype of a 64×64 DIDO.**