Breaking POps/J Barrier with Analog Multiplier Circuits Based on Nonvolatile Memories

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

ABSTRACT

Low-to-medium resolution analog vector-by-matrix multipliers (VMMs) offer a remarkable energy/area efficiency as compared to their digital counterparts. Still, the maximum attainable performance in analog VMMs is often bounded by the overhead of the peripheral circuits. The main contribution of this paper is the design of novel sensing circuitry which improves energy-efficiency and density of analog multipliers. The proposed circuit is based on translinear Gilbert cell, which is topologically combined with a floating nonlinear resistor and a low-gain amplifier. Several compensation techniques are employed to ensure reliability with respect to process, temperature, and supply voltage variations. As a case study, we consider implementation of couple-gate current-mode VMM with embedded split-gate NOR flash memory. Our simulation results show that a 4-bit 100×100 VMM circuit designed in 55 nm CMOS technology achieves the record-breaking performance of 3.63 POps/J.

Keywords

Analog Computing; Sensing Circuit; Floating-Gate Memory; Current Processing; Vector-Matrix Multiplier; Artificial Neural Networks

1. Introduction

Numerous experimental results [1-3] as well as theoretical studies [4, 5] show that analog computing could be extremely energy efficient at low to medium precision of operation. Recent work showed that even accounting for input/output data conversion, mixed-signal computing can be very energy efficient for at least 6-bit operation precision [6]. This creates opportunity for seamlessly integrating analog accelerators into conventional digital computing circuits to improve system's energy efficiency. Analog computing is also enabling some new types of in-memory computing, and hence can address grand challenges of today's digital computers.

Naturally, analog computing is less robust to various nonidealities such as process variations, noise, and nonlinearities and, hence, cannot compete with digital computing at higher (> ~8 bit) precision. There are, however, plenty of important applications, e.g., in machine learning, signal processing, and scientific computing, relying on low-to-medium precision arithmetic, that are serving now as a motivation behind development of efficient purely analog and mixed-signal computing circuits.

Vector-by-matrix multiplication is typically the most frequent operation in many algorithms and computational tasks, most importantly including various types of artificial neural networks. Analog VMMs have been designed in various flavors and topologies utilizing both CMOS and post-CMOS technologies [7]. The most prospective analog VMM circuits are perhaps based on current-mode designs employing metal-oxide memristors [3] due to the excellent scalability, analog properties, and non-volatility of such devices. Yet, memristor fabrication technology is not advanced enough for very large-scale integration. Therefore, some of the research is now focused on more mature, but less dense nonvolatile memories (NVMs), such as floating gate memories [8-11]. For example, an experimentally tested analog neuromorphic chip [2] performed high-fidelity classification with record-breaking speed, density, and energy efficiency, and featured negligible chip-to-chip variations.

Interestingly, numerous papers have been published on the analysis of crossbar array circuits and devices, but little work has been done on optimizing the peripheral circuits. Yet, some prior works claimed that peripheral (sensing) circuitry is the most energy and area demanding component of current-mode VMMs. For example, the power consumption of the peripheral circuitry exceeded 90% in [2] and 83% in [3] of the total budget. The reported area overhead was crudely 95% and more than 55%, correspondingly, for these two studies. This is why the major goal of this work is the development of a high-performance peripheral circuitry for current-mode NVM-based VMM. Our specific focus is on sensing circuit, which is the most important VMM peripheral component. Other peripheral circuits can be typically shared among multiple VMM blocks and have rather negligible overhead. In the context of artificial neural networks, the periphery (neurons) may include activation function circuits, whose overhead is also typically negligible as compared to sensing circuits.

2. Previous Work

2.1 Mixed-Signal VMM Circuits

A number of different VMM topologies, some implemented with unique peripheral circuits, have been proposed in analog and mixed-signal domains [12-15]. For example, time-based [12,13,15] and switch-capacitor [14] multipliers 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 its challenges is process-voltage-temperature (PVT) variations that may limit the smallest integration delay and hence the circuit performance. In addition, a large capacitor (e.g., 25 pF in [12]) might be needed to minimize charge injection issues and increase the signal-to-noise ratio.

In the second approach, precisely-fabricated fringe capacitors are employed to implement active multipliers with a moderate precision (> 4-bit) [14]. Its main issue is large and power hungry active amplifiers. Passive switched-capacitor circuits could in principle address this problem though at the expense of having more leakage, capacitive coupling, and charge injection issues, which in turn limits computing precision to < 4-bit [1].

ISLPED '18, July 23–25, 2018, Seattle, WA, USA

ACM ISBN 978-1-4503-5704-3/18/07. https://doi.org/10.1145/3218603.3218613

In current-mode approach, multiplication and addition are performed with fundamental Ohm and Kirchhoff's laws (Fig. 1). Here, the j-th output (current in Fig. 1), is given by

$$I_j = \sum_{i=1}^{N} W_{ji} V_i,$$

where W_{ji}, are the matrix weights (crosspoint conductances) and V_i is the i^{th} element of the input (voltage) vector. In general, depending on the choice of utilized crosspoint devices and peripheral circuits, both input and output vectors can be presented in terms of either voltage or current. Weights are encoded as conductances in resistive crossbar memories [3], subthreshold currents in floating gate memories [2], or transistor widths in pure CMOS designs [16].

For example, cascode current mirror structure was used in [16] to implement a fully current-mode VMM, i.e. with both input and output vectors encoded via currents. Weights are realized by a set of transistors whose widths are scaled according to the predetermined values. The main caveat of such design is an area (and hence energy) overhead for weight implementation, which exponentially increases with weight precision. A more promising solution is to implement matrix weights with NVMs, such as programmable conductance crosspoint devices. Especially encouraging is a recent work on VMMs based on metal-oxide memristors [3] and floating-gate memories [8-10].

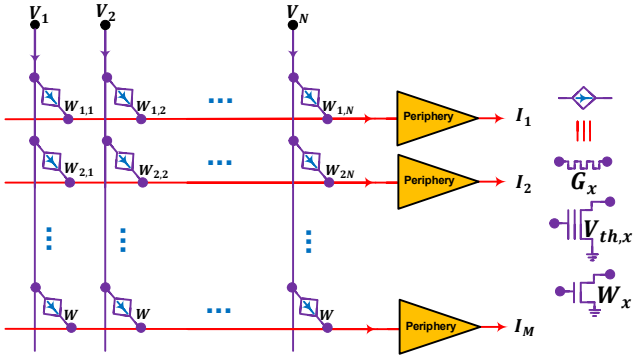

Fig. 1. A general idea of $M \times N$ current-mode all-analog VMM circuit. The inset shows several options for crosspoint device implementation.

2.2 Flash-based VMM design

Implementations of neuromorphic circuits with floating gate memories have a long history [6]. The most prominent examples are circuits based on so-called "synaptic transistor", which is a type of floating gate memory implemented with standard CMOS process. Even though many efficient systems have been built using synaptic transistors, the main caveat of that approach is bulky memory cells, with ~10^3 F^2 footprint per memory device, where F is the process minimum feature size [6]. In another, more recent work, industrial-grade memory cells, that have ~25 F^2 footprint per cell, were modified to for analog circuit applications [10]. The redesign allowed for precise tuning of the individual cells, which is a necessary functionality for analog-mode VMMs. The effective area of redesigned cell was tripled though it was still an order of magnitude denser than that of synaptic devices.

Some of the results presented in this paper are based on such redesigned 55-nm ESF3 NOR flash memory (Fig. 2a-c). Due to its split-gate structure, ESF3 devices offer very high output impedance. For example, the experimentally measured output resistance is about 100 GΩ in subthreshold regime for the targeted current range, which is useful for the considered analog applications. Also, the cell's compact structure results in a very low capacitance, of the order of ~75 aF/cell on average, during subthreshold operation. (More details on the various aspects of this technology, including I-V characteristics, erasure and programming operation, cycling endurance, retention, noise are discussed in [10].)

Figure 2d shows the most common design for floating gate memory VMMs based the gate-couple topology. In such design, the input current vector is applied to an array of diode-connected floating gate memory cells. The two-quadrant multiplication is implemented by dedicating two rows per output and using the conventional differential weight scheme.

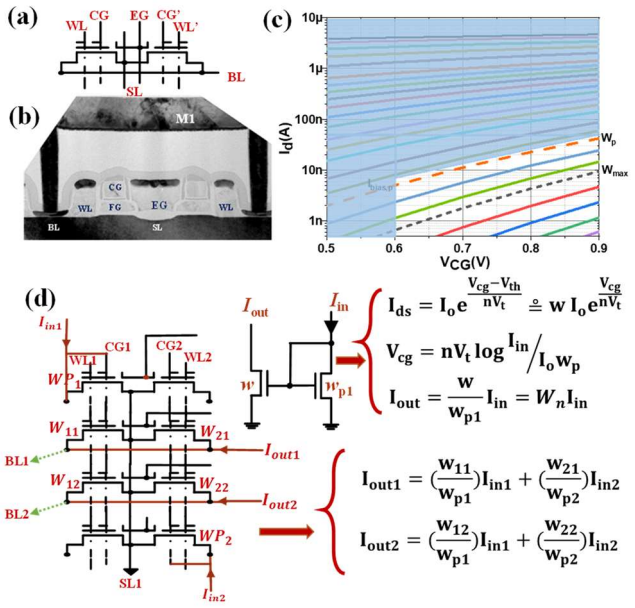

Fig. 2. Current-mode VMM implementation with split-gate NOR flash memory: (a) Schematics and (b) TEM image of SST's ESF3 supercell; (c) Drain-source current as a function of control-gate voltage under typical read conditions (V_{BL} = 1 V, V_{WL} = 1.2 V, V_{SL} = 0 V) for various programmed states. The unshaded area shows typical low-voltage operating region; (d) Example of a 2×2 VMM circuit, inluding two rows of peripheral cells, and the key equations governing its operation.

2.3 Sensing Circuits for Current-Mode VMMs

The peripheral sensing circuit is typically designed to provide low input impedance on a shared bitline, i.e. the horizontal lines in Fig. 1, and sink/source the current flowing in it. The simplest approach for sensing the current is to use a low-voltage cascode current mirror. Its main challenges are nonlinearities in the transfer function and voltage variation on the virtual bias, which can significantly deteriorate the precision. In addition, current mirrors are susceptible to process variations, which mandates large devices. The upshot is low-speed and high-power consumption.

Conventional transimpedance amplifiers (TIAs) have been used in both nanodevice computing engines [3] and flash-based dot-product circuits [10] to pin the virtual bias needed for linear operation and for I-V conversion. The area overhead of operational amplifiers has been disregarded in favor of excellent linearity. In addition, the amplifiers are often designed to work in a certain "operating point" rather than dealing with a large-signal input. This requires a huge overdesign cost in terms of power and area for proper functionality. There are other drawbacks including the

requirement of high gain amplifier in a TIA, the dependence of bandwidth on feedback resistor, the need for compensation, and the circuit slews for a significant period of time.

All-analog current-mode designs could potentially allow for a much better performance/cost. Indeed, another implementation approach is to use a second-generation current conveyor (CCII). The idea was originally introduced in [17], where CCII has been used to build current summers featuring low input impedance. Since then, various CMOS implementations were proposed [18], utilizing either open-loop and close-loop structures, with the former preferred for a better speed and dynamic behavior. It is also worth mentioning that CCII designs based on the topology introduced in [19] are not limited by slew-limited transient response and the gain-bandwidth product tradeoff and hence, in principle, achieve higher speed compared to TIAs. However, their overall energy consumption and circuit area are still very high (see, e.g., [20]). Also, the designs based on operational amplifiers are not appealing for obvious reasons.

In light of the aforementioned shortcomings, we have designed a compact current-mode peripheral circuitry based on Gilbert translinear loop, which provides a relatively low-input impedance and a wide range of gain control and temperature insensitivity. The design unique features ultimately enable excellent linearity and high-speed and low-energy operation.

3. Proposed Sensing Circuit

The proposed circuit is shown in Fig. 3. The array bitline is connected to node "Q", while array current is supplied by M_{3a}. Due to the local feedback loop, the increase in the input current leads to decrease in I_{2a} and, as a result, differential voltage between X and Y nodes, which is then converted to current by the following low-gain amplifier.

M_1, M_2, and M_4 pairs are designed in weak inversion and M_3 pair is velocity saturated. The rest of the devices are operated in the saturation regime. When biased in weak inversion, $M_{1,4}$ pairs form

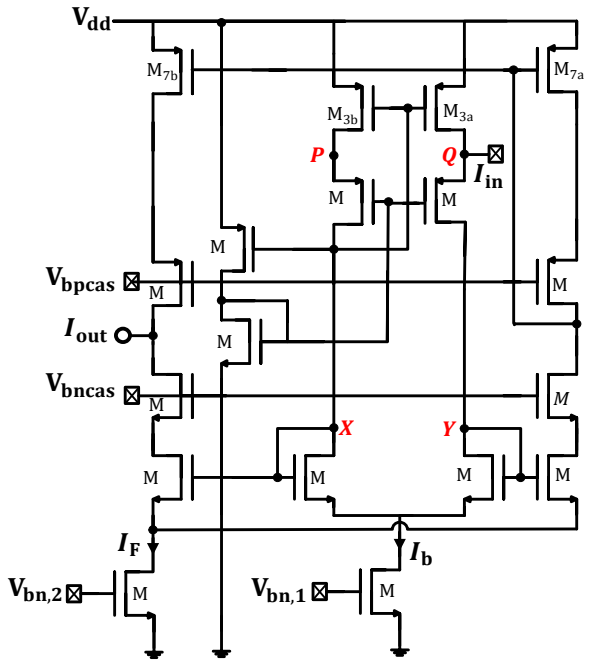

Fig. 3. The proposed sensing circuit for current-mode VMM (not including biasing circuitry).

a translinear loop, which has an excellent wideband current-following behavior.

When the input current is zero, i.e. $I_{in} = 0$, $I_{3a}=I_{3b}$, the symmetrical structure of the circuit imposes $I_{3a} = I_b/2$, where I_b is the bias current provided by M_8. Since I_{in} is supplied by M_{3a}, the circuit analysis yields

$$I_{1a} = (I_b - I_{in})/2, \;\; I_{1b} = (I_b + I_{in})/2 \; .$$

Since $M_{1,4}$ pairs are biased in subthreshold, V_{XY} is expressed as

$$V_{XY} = nV_T \ln((I_b + I_{in})/(I_b - I_{in})) \; ,$$

where V_T and n are thermal voltage and subthreshold slope factor, respectively. Furthermore, a simple analysis shows that

$$\frac{I_{4a}}{I_{4b}} = \frac{I_b+I_{in}}{I_b-I_{in}}.$$

Assuming that $I_F = I_{4a}+I_{4b}$ is the bias provided by M_9, the output current, i.e. the sensing circuit transfer characteristic, is given by

$$I_{out} = \left(\frac{I_F}{I_b}\right) I_{in}. \qquad (1)$$

To improve the performance, one can use low-V_{th} devices for $M_{1,3,4}$ pairs (though this is not mandatory for proper functionality). As we show later, in 55 nm process, this allows reaching the same nonlinearity performance with crudely 15% less power consumption.

4. Circuit Analysis

4.1 Nonlinearity

Closed-loop high-gain amplifiers provide excellent linearity as long as the gain requirements are met. For current processing circuits, nonlinearity becomes challenging in part due to the short channel effects in sub-deca-nm technologies. Both deterministic and random factors result in deviation from the ideal behavior given by Eq. 1.

Specifically, the main intrinsic nonlinearity originates from unequal source drain voltages across M_{3a} and M_{3b}. The maximum relative error, defined as $(\delta_r)_{max} = (|I_{out}-I_{out}^{ideal}|/ I_{out}^{ideal})_{max}$ due to only this factor is shown in Figure 4a. Reducing $(\delta_r)_{max}$ is related to minimizing $\delta = I_{3a}/I_{3b}$, which in turn, is a function of I_{in} and I_b, and is achieved by designing M_3 in the deep velocity saturated region. For example, $(\delta_r)_{max}$ could be made as low as 0.1% by properly adjusting the bias current.

The second issue is process induced variations. For example, mismatch between I_{3a} and I_{3b} creates an offset in the transfer characteristics. One straightforward solution is to adjust accordingly memory cells' conductances. Indeed, I_{3a} - I_{3b} offset can be compensated by properly tuning conductances in two additional auxiliary columns of memory cells, i.e. with two extra devices per each bitline. After measuring the input-referred offset, one of the devices of a pair, based on the sign of the offset current, is set to either sink or source the desired current, while the other one is fully turned off. This approach allows avoiding scaling transistors in the sensing circuit, with minimal power/area overhead. Process-induced variations also impact $(\delta_r)_{max}$, since δ depends on the matching of the M_3 pair. Additionally, a mismatch in the voltage threshold of $M_{1,4}$ pairs could result in deviations from ideal output current. The solution here again is to compensate total resultant offset by fine-tuning crossbar devices.

To evaluate the impact of process variations, we use statistical simulations over all corners to find the worst-case nonlinearity. As

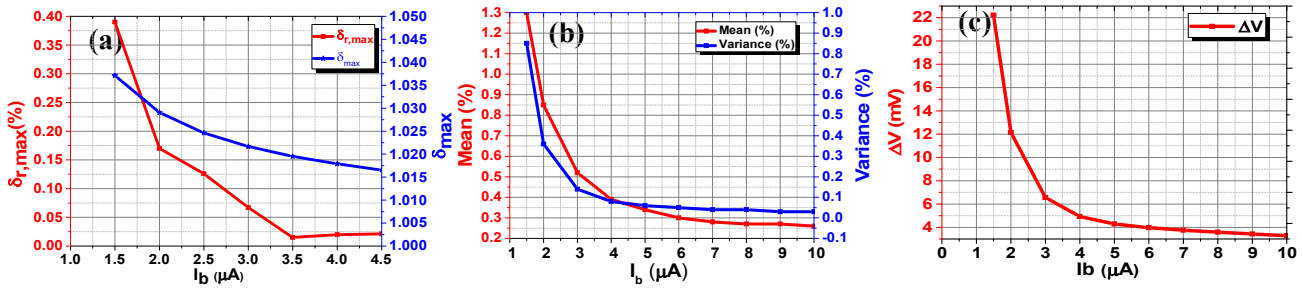

Fig. 4. The impact of nonideal transistor behaviour, process variations, and finite condutance on the circuit linearity as a function of bias current: (a) Error due to source voltage variations across M_{3a} and M_{3b} assuming $I_{in} = (I_{in})_{max}$ in TT corner, (b) total realtive error at the output due to device mismatches, and (c) virtual bias variations. For all panels, $(I_{in})_{max} = 1$ μA and $V_{DD} = 1.2$ V.

shown in Fig. 4b, both mean and variance of the total nonlinearity error could be as low as 0.26%. This can be improved even further by increasing the area of the circuit (discussed below). It is worth mentioning that the discussed techniques raise energy consumption, naturally yielding a precision-energy trade-off. Also, in practice, the nonlinearity error is expected to be less, by a factor of ~5 according to our estimates, when accounting for symmetric layout mismatch reduction techniques, which were not considered in this work.

Finite input conductance of the sensing circuit contributes to the nonlinearity of VMM operation rather than sensing. Intuitively, when the input current increases, I_{2a} decreases and so does the M_{2a} source voltage. The maximum change in M_{2a}'s source-gate voltage is given by $-nV_T \ln[1-(I_{in}/I_b)_{max}]$. However, the negative local feedback, formed by M_{10} and M_{11}, decreases M_{2a}'s gate voltage, and therefore compensates for source-gate voltage change. Additonally, proper sizing of M_{11} and controlling the bias current allow controling virtual bias swing for a given maximum input current (Fig. 4c). The impact of this swing on computing precision depends on the type of memory cell used in the array and will be discussed in Sect. 5.2 for the case of floating gate memory.

Finally, it is noteworthy that all nonlinearity terms reduce simultaneously with respect to the bias current (Fig. 4). Therefore, in a typical design, the minimum bias current could be determined by the precision requirements.

4.2 Noise

The proposed circuit has a relatively low input impedance so that input-referred current noise scales linearly with bias current (Fig. 5a). It should be noted however, that for the case of sub-deca-nanometer memory technologies and, in particular, floating-gate memories, low-frequency noise of these devices would dominate the noise power [10] – see Sect. 5.2 for more discussion.

4.3 Settling Time

In general, transfer function linearity requirement determines determine the transistor sizing, and, in particular, the smallest I_b, and capacitances C_X, and C_Y. With these values fixed, the settling time, and, as a result, energy consumption, can be further optimized by finding the optimal output pole location. The output pole can be relocated by adjusting the output current, e.g., by changing I_F. For a certain translinear loop size, initially increasing I_F improves the settling time (Fig. 5b). However, at some point, the overshoot in time response becomes excessive and deteriorates the settling time. Increasing the output current is no longer helpful since the dominant pole is no longer attributed to the output pole. To summarize, the optimum settling time is obtained by adjusting I_F based on given I_b, C_X, and C_Y, i.e. the location of the first pole, and C_L, i.e. corresponding dimensions of the load array.

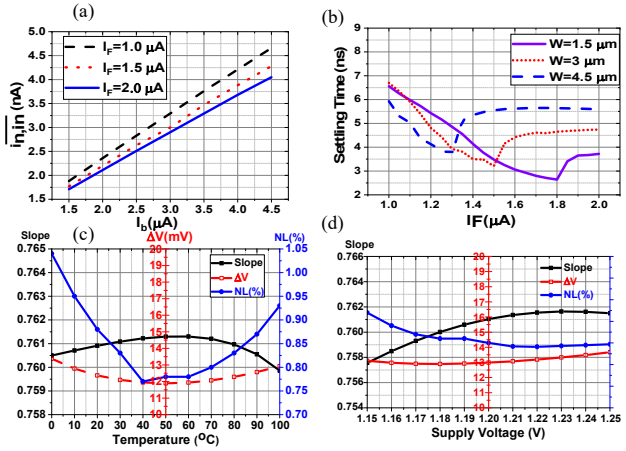

Fig. 5. Analysis of noise, settling time, and PVT variations: (a) Total integrated input-referred current noise for several I_F as a function of bias currents at $I_{in} = 0$ and 100 MHz bandwidth; (b) Settling time as a function of translinear loop size and I_F current at $I_b = 1.75$ μA; (c) Temperature dependence of virtual bias variation (ΔV), slope, and total worst-case nonlinearity error; (d) Impact of supply voltage variations on ΔV, transfer function slope, and total worst-case nonlinearity error. For all panels, $(I_{in})_{max} = 1$ μA, $C_L = 7.5$ fF.

4.4 Temperature and Supply Variations

Figure 5c shows the temperature dependence of the considered nonlinearities. In general, virtual bias is sensitive to temperature variations because V_t is a function of temperature. To bound the worst-case ΔV below the desired value, across all temperatures, I_b is supplied from a PTAT (proportional to absolute temperature) current source. Fig. 5c shows that this compensation scheme allows to limit virtual bias variation within wide range of temperatures. Additionally, to keep the slope of the transfer function temperature invariant (within < 0.2%), I_F is also supplied by the same PTAT source. The temperature sensitivity of both ΔV and slope can be further improved by designing a more complex compensation circuitry.

Finally, Figure 5d shows that reasonable ±4% fluctuations in supply voltage result in <0.5% change of the transfer function slope. This is because the slope depends only on bias currents, so that as long as the current reference, which supplies these bias currents, is voltage insensitive and critical devices remain in their targeted operating region, the linearity remains acceptable.

5. Case Studies

5.1 Metric-Optimal Sensing Circuits

We designed four different styles of the proposed sensing circuit, in each case optimizing bias current and size of the devices according to the specific metric. In particular, we consider power-optimal (referred as S_1), the area-optimal (S_2), the precision-optimal (S_3), and the energy-optimal (S_4) designs. In addition, each style is implemented based on a targeted 6-bit for $S_{1,2,4}$ and 8-bit for S_3 precision requirement. All designs are based on 1.2 V devices in Global Foundries 55 nm process technology.

Fig. 6a summarizes various characteristics of the implemented designs, while Fig. 6b shows the impact of input current for S_4. (The maximum input current is naturally linearly proportional to parameter N of VMM circuit.) Figure 6b shows that for small input currents, the critical devices must be kept large to counteract the process variation effects, resulting in slower operation. On the other hand, I_b and width of $M_{1,2,3}$ can be scaled up accordingly for larger maximum input currents to keep the precision/speed constant. The power, area, and energy naturally increase with respect to the maximum input current.

The impact of process variations on the circuit's linearity is also studied for all designs (Fig. 6a). Statistical simulations across all corners show that the sensing circuitry can effectively operate with up to 8-bit precision. In particular, the process-induced precision errors are controlled by the proper sizing of the translinear devices and the bias current. Dispersion in transfer function slope, which follows a normal distribution, is addressed by adjusting the weights.

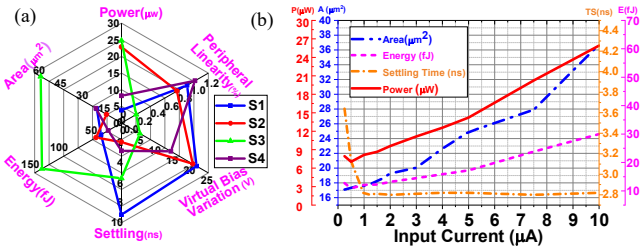

Fig. 6. Sensing circuit results: (a) Various performance characteristics for 4 different implemented designs assuming $(I_{in})_{max}$ = 1 μA and C_L = 7.5 fF, which crudely corresponds to a 4-bit 100×100 1Q VMM based on floating gate memory, driving the same size circuit; (b) Impact of the input current on S_4 sensing circuit chracteristics. Nonlinearity and virtual bias distortion are kept relatively constant at 1.05% and 14.5 mV, respectively.

5.2 All-Analog Current-Mode NOR Flash VMM

Energy-optimal design S_4 is further utilized to investigate performance of a current-input current-output fully analog VMM based on split-gate embedded NOR flash memory technology.

One of the most important characteristics for the analog-mode VMM is its effective operating precision. Even though S_4 design is suitable for 6-bit operation, for simplicity, we here consider rather conservative assumption that VMM's input, weight, and computing precisions are all effectively 4 bit. To justify it, let us first note that the weight precision might be limited by each of the following factors: tuning accuracy, drift, bitline bias variations, and subthreshold slope nonlinearities.

The redesigned layout of the memory array allowed to demonstrate experimentally >8-bit tuning accuracy for a single cell when sufficient number of pulses are applied during tuning procedure [21]. The tuning precision is expected to be somewhat lower for the current mirror structure, especially considering half-select disturbance in the memory array, but still much better than the targeted 4-bit precision [10]. The virtual bias variation ΔV can be limited to less than 15 mV, which corresponds to < 1% overall bitline distortion for the targeted current range. The accelerated retention tests have shown less than 1% drift in memory state after 7 months for the vast majority of devices [2], thus also providing evidence for implementation of 4-bit weights.

In general, the effective weight error due to the subthreshold slope nonlinearities depends on the choice of peripheral device state and the selected range of states used for the array devices. In addition, there is a tradeoff between power consumption and weight precision. Indeed, using the memory states corresponding to the lower operating voltages (Fig. 2c) helps reducing power consumption. The downside is that at these voltages the subthreshold slopes are more nonlinear. In light of this tradeoff, the state of the peripheral cell and the maximum current via array device are assumed to be 30 nA and 10 nA, respectively, under V_{CG} = 0.9 V, V_{WL} = 1.2 V, and V_{BL} = 1 V biasing conditions, shown in the unshaded region of Fig. 2c.

Assuming negligible input-referred noise of the sensing circuit, the main limiting factors for the computing (output) precision are the sensing circuit's nonlinearities and the low-frequency noise of the memory devices. Following the analysis presented in Sect. 4.1, the total relative nonlinearity error of the sensing circuit based on S_4 style is 1.1%. The subthreshold current fluctuations are mainly due random telegraph noise (RTN) in as-fabricated cells, and, more generally, $1/f$ noise after repeated switching. For example, discernible transition between RTN and $1/f$ noise was experimentally observed in 65 nm NOR flash memories within 100 switching cycles [22]. Our own measurements for ESF3 cells show that only few cells (out of 140 total) had severe subthreshold current fluctuations, even after cycling each device 1000 times.

Assuming the targeted maximum current, and the reported spectrum with flat region below 1 KHz and the corner frequency of ~500 KHz [22], the root mean square of the current noise via single device is ~575 pA at 300 MHz operating bandwidth. The total resultant output signal-to-noise ratio is ~44.8 dB for 100 element dot-product operation. (Due to similar physics of operation, $1/f$ noise can be also crudely quantified by considering much more numerous reported noise data for standard 55 nm MOSFETs with the same width and length.)

The above analysis takes into account all important nonideality factors and shows that achieving 4-bit computing and weight precision should be relatively straightforward for the considered VMM design. The estimates are rather conservative and, e.g., even higher weight precision is possible when using larger operating voltages.

To evaluate and optimize operation speed, we assumed that several VMMs are chained in a cascade structure, with output of one VMM sensing circuit feeding directly the input of the next VMM stage. This assumption is representative of all-analog multilayer neural network implementation (though neglects additional circuitry, which might be required for neuron implementation). The propagation delay through such cascade can be minimized by adjusting VMMs' output pole locations. Note that since the input pole of a particular stage VMM is effectively the output pole for the preceding stage multiplier, only output pole can be considered. Also, out of the two output poles, the first one is always fixed based on the sensing circuitry's targeted maximum input current and linearity requirement. Therefore, the goal is to only find optimal

location of the second pole, at which the settling time is the smallest.

More specifically, as discussed in Sect. 4.3, for a desired, fixed sensing circuit linearity and given capacitive load, the smallest delay is achieved at the specific sensing circuit output current. The optimal current value, however, is typically higher than the nominal subthreshold current of the minimum size floating gate transistor. Forcing such optimal current via single peripheral floating gate cell would lead to significant errors in the multiplier operation.

To overcome this issue, we assume that M_p peripheral cells are connected in parallel for each input, which effectively increases the width of the peripheral floating gate transistors. In particular, let us first note that the optimal output current is proportional to the load capacitance, which is $(M+M_p)C_{cell}$, where C_{cell} is memory cell's unit capacitance. Therefore, for the most interesting cases of large M, increasing M_p and, simultaneously, output current for optimal pole location result in both lowering individual currents via peripheral cells and decreasing settling time. More generally, the settling time in this case is proportional to $(1+M/M_p)C_{cell}/(I_p)_{max}$, where $(I_p)_{max}$ is the desired maximum current via peripheral cell.

Figure 7 summarizes various performance characteristics of the considered VMM as a function of its size. As expected, the simulation results show that the average energy consumption for the dot-product operation (one channel) is growing superlinearly with N, mostly due to the increasing maximum input current. The number of operations per channel grows linearly, and with constant settling time, the energy-efficiency saturates. The relative peripheral area overhead is always below 11%.

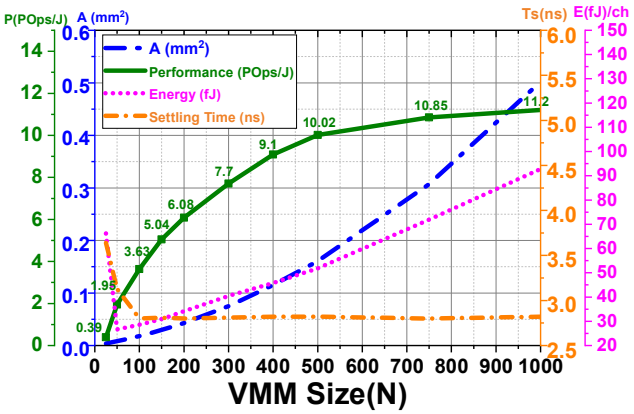

Fig. 7. Performace, total area, and energy per channel for $N \times N$ VMM based on 55 nm ESF3 NOR flash memory. POp/J operation is achieved for $N > 50$, which are practical kernel sizes for many applications.

6. Summary

A very efficient sensing circuitry, which utilizes the translinear principle of Gilbert cell, is proposed to boost the performance of NVM-based analog-mode VMMs. In prior work, the area, energy, and density potentials of current-domain circuits were typically counterbalanced by the overhead of PVT compensation. In this study, offset calibration is performed by considering two auxiliary columns of programmable NVMs in the crossbar array so that robustness against PVT variations is achieved with minimal overhead. As a case study, we investigated several sensing circuits, each optimized for a specific metric. Our simulation results show that 100×100 4-bit VMM designed in 55 nm CMOS technology with embedded NOR flash and employing energy-optimal sensing circuit achieves 3.63 POps/J.

7. References

[1] E. H. Lee, and S. S. Wong, "Analysis and design of a passive switched-capacitor matrix multiplier for approximate computing", *IEEE JSSC*, vol. 52, pp. 261-271, 2017.

[2] X. Guo *et al.*, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology", in: *Proc. IEDM'17*, San Francisco, CA, Dec. 2017, pp. 6-5.

[3] M. Hu *et al.*, "Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication", in: *Proc. DAC'16*, Austin, TX, June 2016, pp. 19-25.

[4] C. Mead, *Analog VLSI and Neural Systems*, Addison-Wesley, 1989.

[5] B. Degnan, B. Marr, and J. Hasler, "Assessing trends in performance per watt for signal processing applications", *IEEE Trans. VLSI*, vol. 24, pp. 58-66, 2016.

[6] J. Hasler and B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems", *Frontiers in Neuroscience,* vol. 7, art.118, 2013.

[7] C. Li *et al.*, "Analogue signal and image processing with large memristor crossbars", *Nature Electronics*, vol. 52, pp. 52-59, 2018.

[8] S. Ramakrishnan and J. Hasler, "Vector-matrix multiply and winner-take-all as an analog classifier", *IEEE Trans. VLSI*, vol. 22, pp. 353-361, 2014.

[9] M. R. Mahmoodi and D.B. Strukov, "An ultra-low energy internally analog, externally digital vector-matrix multiplier circuit based on NOR flash memory technology", in: *Proc. DAC'18*, San Francisco, CA, June 2018 (accepted).

[10] X. Guo *et al.*, "Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells", in: *Proc. CICC'17*, Austin, TX, Apr.-May 2017, pp. 1-4.

[11] L. Fick *et al.*, "Analog in-memory subthreshold deep neural network accelerator", in: *Proc. CICC'17*, Austin, TX, Apr.-May 2017, pp. 1-4.

[12] R. D'Angelo *et al.*, "A time-mode translinear principle for nonlinear analog computation", *IEEE TCAS-I*, vol. 62, pp. 2187-2195, 2015.

[13] D. Miyashita *et al.*, "Time-domain neural network: A 48.5 TSOp/s/W neuromorphic chip optimized for deep learning and CMOS technology", in: *Proc. A-SSCC'16*, Toyama, Japan, Nov. 2016, pp. 25-28.

[14] S. Joshi *et al.*, "21.7 2pJ/MAC 14b 8×8 linear transform mixed-signal spatial filter in 65nm CMOS with 84dB interference suppression", in: *Proc. ISSCC'17*, San Francisco, CA, Feb. 2017, pp. 364-365.

[15] M. Bavandpour, M. R. Mahmoodi, and D. B. Strukov, "Energy-efficient time-domain vector-by-matrix multiplier for neurocomputing and beyond", *ArXiv:1711.10673*, 2018.

[16] J. Binas *et al.*, "Precise deep neural network computation on imprecise low-power analog hardware", *ArXiv:1606.07786*, 2016.

[17] A. S. Sedra, and K. C. Smith, "A second-generation current conveyor and its applications", *IEEE Trans. Circ. Theory,* vol. 17, pp. 132-134, 1970.

[18] K. N. Salama, and A. M. Soliman, "CMOS operational transresistance amplifier for analog signal processing", *Microelectronics Journal,* vol. 30, pp.235-245, 1990.

[19] E. Brunn, "CMOS high speed, high precision current conveyor and current feedback amplifier structures", *Int. Journal of Electronics,* vol. 74, pp. 93-100, 1993.

[20] B. Calvo *et al.*, "High-speed high-precision CMOS current conveyor", *Analog Integrated Circuits and Signal Processing,* vol. 34, pp. 265-269, 2003.

[21] F. Merrikh Bayat *et al.*, "Model-based high-precision tuning of NOR flash memory cells for analog computing applications", in: *Proc. DRC'16*, Newark, DE, June 2016, pp. 1-2.

[22] X. Yang *et al.*, "Impact of P/E cycling on read current fluctuation of NOR Flash memory cell: A microscopic perspective based on low frequency noise analysis", in: *Proc. IRPS'15*, Monterey, CA, Apr. 2015, pp. 5B-7.