CMOS based Ultra-low Power High-Precision Analog Vector Matrix Multiplication Circuit with $\pm 0.1\%$ Error for Vision Application

Nikita Mirchandani and Aatmesh Shrivastava

Dept. of Electrical and Computer Engineering
Northeastern University, Boston, USA
mirchandani.n@husky.neu.edu, aatmesh@ece.neu.edu

***Abstract*—This paper presents the design of a low power multiplier cell which multiplies an input voltage with current. The multiplier achieves high precision with $\pm 0.1\%$ error for a wide range of both voltage and current. An approach to increase the range of the current is also presented. The multiplier cell is used to perform 2-D discrete convolution on an input matrix of size 3×3 with a 2×2 kernel. The outputs of the matrix multiplier are available after $7.5\ \mu s$, which is independent of the matrix size.**

I. Introduction

With the end of CMOS scaling in sight, the future road-map for the expansion of computing power remains uncertain. State of the art digital computing systems, which benefited significantly from scaling, actually use area and power inefficiently as they operate transistors at two extreme points. Contrary to digital, analog circuits make use of the entire operating voltage range of the transistor and are highly area and power efficient [1]. Analog computing has long been associated with high variability and linearity issues when implemented on-chip. However, sub-threshold g_m stages have been shown to be very stable against process, voltage, and temperature variations [2]. We can leverage this technique to come up with high precision analog computing circuits. Analog implementation of vector matrix multipliers (VMM) has become quite popular in machine learning vision applications where they are used to implement convolutional neural networks (CNN).

CNNs have been proven to be very accurate in image classifiers [3] where convolutional and pooling layers of a CNN perform feature extraction. The convolution layer outputs a feature map by sliding a kernel over an entire image which consists of element wise multiplication and accumulation/addition(MAC). These successive MAC operations are efficiently performed in hardware. While conventional hardwares such as GPUs and FPGAs use the digital implementation of MAC, the trend in application specific circuits (ASIC) is to use the analog implementation of MACs to achieve higher computation efficiency. However, analog multipliers can suffer from lower precision due to process, voltage, and temperature variations, and inherent nonlinearity in the devices used for VMM. In this paper, we present a highly accurate VMM implementation using analog circuits to achieve a 10-bit digital equivalent mixed precision accuracy.

Conventional analog multipliers are designed as Gilbert cells based on operational transconductance amplifier (OTA) based implementations, and are typically used in amplitude modulation applications [4]. Translinear principle of BJTs can also be exploited with mosfets in the subthreshold region to achieve multiplication in current mode multipliers [5]. These circuits consume high power in the μWs range and are hence not suitable for analog computation applications. Current mode multipliers operating in the subthreshold region are also used to improve linearity of the multiplier [6].

Another design approach to perform VMM is using the Dot Product Engine (DPE) [7]. One unit DPE is placed at each intersection of a crossbar array to speed up the operation. Each DPE consists of a memresistor and series transistor. Memresistors were first introduced in [8] and are used in VMM operations to implement the weight matrix. The conductance of the memresistor is proportional to the weight, and is tuned to update the weights. However, memresistor based designs provide lower resolution for computing as high energy barrier is needed to separate two adjacent resistor values.

Advances in floating gate devices have made weight storage with non-volatile memories possible [6]. Floating gates do not lose charge as they are surrounded by an insulator. To update weights, the charge is modified with Fowler-Neidham tunnelling and hot carrier injection. In [9], floating gate based NOR flash memory cells are used to design a 10×10 multiplier. However, non volatile memory (NVM) arrays suffer from sneak path effect which limits their large scale integration. Also, programming the floating gates is based on hot electron injection and tunneling, which are prone to error and affect overall precision of the multiplier.

In [10], a switched capacitor multiplier for MAC operations is presented. However, it can only perform 3-bit MAC operations. The VMMs presented so far have achieved low to medium precision. This paper presents a time-based high precision multiplier cell with less than $\pm 0.1\%$ error. We further present the use of this circuit as a VMM implementation for machine-learning vision applications.

978-1-7281-2788-0/19/$31.00 ©2019 IEEE

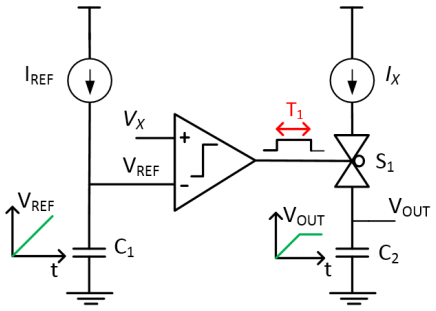

Fig. 1: A 10-bit equivalent analog multiplier circuit using time and current to achieve multiplication output.

II. Analog Multiplier Implementation

Fig 1 shows our proposed analog multiplier circuit which uses voltage to time conversion (VTC) to perform analog multiplication. Final output of the circuit results in an output voltage which is a weighted output of multiplication of V_X and current I_X. The concept behind the multiplication circuit is explained as follows. Capacitors C_1 and C_2 are charged with currents I_{REF} and I_X respectively until the voltage V_{REF} reaches V_X. Capacitor C_1 charges up to V_X in time T_1, given by

$$T_1 = \frac{C_1(V_X + V_{OS})}{I_{REF}} + t_d \tag{1}$$

where V_{OS} is the offset of the comparator, and t_d is the delay. Since C_2 is charged for the same time by I_X,

$$V_{OUT} = (\frac{C_1}{C_2 I_{REF}})V_X I_X + \frac{C_1 I_X V_{OS}}{C_2 I_{REF}} + \frac{I_X t_d}{C_2} \tag{2}$$

V_{OUT} is ideally a scaled multiplication of the input voltage V_X and weight represented by current I_X.

A. High Precision Realization

Equation 2 shows that the output voltage is a weighted multiplication of current and voltage. However, several non-idealities can effect the accuracy of the proposed analog multiplier. The accuracy of the results depends on the matching of C_1 and C_2 and on the accuracy of I_{REF}. Through circuit design, an ideal current source can be realized for I_{REF} making it a constant [11]. The matching for C_1 and C_2 can be achieved through analog matching techniques. The delay of the comparator circuit to output a zero will add additional time to T_1 which is another source of error. We address this by using higher values for the capacitors C_1 and C_2 while keeping the output load of the comparator very small to achieve very low delay compared to T_1. Additionally, the charge injection from the switches in the transmission gate during transition can create non-idealities which we address by switch sizing to cancel charge injection. Finally, the device mismatch can result in an offset voltage at the output of the comparator. In this paper, we do not address this issue but it can be addressed through offset correction technique while incorporating an offset correction phase [12] at the expense of computation efficiency. The offset of the comparator can be brought down from a few mVs to hundreds of μVs using offset compensation techniques [13]. Offset correction technique will involve trading off performance with accuracy.

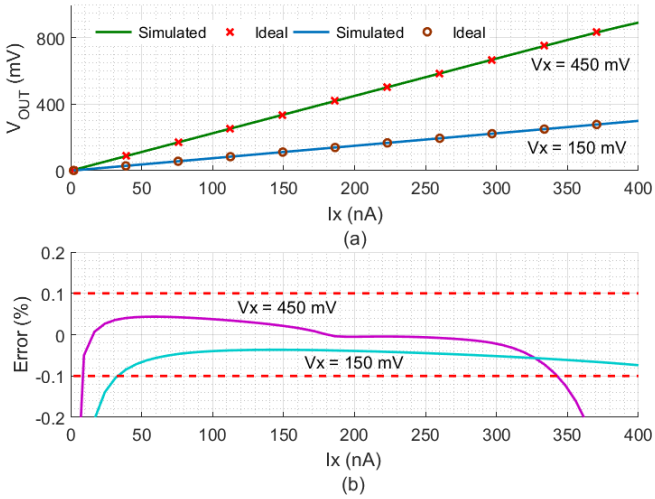

Fig. 2: (a) Simulation results of multiplication of V_X and I_X compared to actual multiplication. (b) Error (%) of the multiplier circuit for $V_X = 450\ mV$ and $150\ mV$ at temperature = 27^oC. The error is within $\pm 0.1\%$

Our range of operation for I_X for which the error is within $\pm 0.1\%$ is 35 nA to 335 nA. This range can be extended by changing the scaling factor of the multiplication. This is easily achievable by trimming capacitor C_2. Hence, area can be traded off for an increased dynamic range. V_X is within the range $150\ mV$ to $450\ mV$. The time to settle for V_{OUT} is equal to T_1 and is input voltage dependent. V_{OUT} is sampled after worst case time T_1 corresponding to $V_X = 450\ mV$, which is equal to 7.5 μs. Since capacitors show small variation with process and temperature, the multiplication is invariant to PVT variations. Fig 2 shows the simulation results of the multiplier with a scaling factor of 200 nA. The simulation results show that the multiplication closely follows the ideal values with less than $\pm 0.1\%$ error achieving an equivalent 10-bit digital computation accuracy. Fig 2 (b) shows the error percentage of the multiplication between variables V_X and I_X at 27^oC. The range of I_X, for which the error is within $\pm 0.1\%$, is 35 nA to 335 nA. The major contributor to this error is the comparator delay. Effect of offset is not included in this simulation. A low power (340 nW) high performance comparator is designed to reduce delay.

Fig 3 shows the percentage error of multiplication for temperature range of 0^oC to 80^oC. For normal operating temperature range from (0°C-27°C), the error is less than $\pm 0.1\%$. The error increases at high temperature because of increased leakage in switch S_1. Since the sampling of the output voltage is done after the worst case time corresponding to $V_X = 450\ mV$, the error is highest for $V_X = 150\ mV$ as it has the fastest settling time. The error is within 0.55% for the given range of current. Additional design techniques such as low-leakage switches can be explored for future designs to

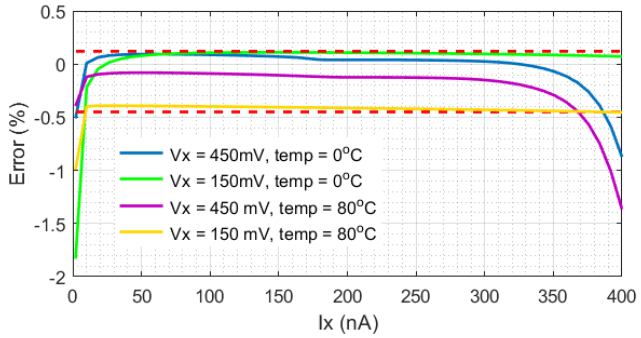

Fig. 3: Variation of error (%) with temperature variation of 0-80°C.

reduce this error caused at higher temperatures.

B. Range Extension

As shown in the previous section, the range of I_X for which error is within $\pm 0.1\%$ is 35 nA to 335 nA. This is achieved with a scaling factor of 200 nA, obtained by sizing C_2 4× the size of C_1. C_1 is chosen to be $817fF$. This range can be extended by changing the scaling factor of the multiplication. This is easily achievable by trimming capacitor C_2. Hence, area can be traded off for an increased dynamic range. Increasing C_2 to 22× of C_1, a higher dynamic range of 100 nA to 1.53 μA is achieved. The area of the multiplier cell increases from $65 \times 65 \mu m^2$ to $135 \times 135 \mu m^2$ when range is extended from 0.98 decades to 1.64 decades in CMOS 130 nm node. Our area is comparable to an 8-bit digital array multiplier implemented in 130 nm CMOS [14]. Fig 4 shows percentage error for extended range of I_X.

III. Vector Matrix Multiplication

The multiplier cell can be expanded to perform 2-dimensional convolution operations on matrices as needed in a CNN. We take an input array of size 3×3, and a kernel of size 2×2 to show the operation of the proposed multiplication unit. The convolution operation results in a 2×2 output matrix.

$$\begin{bmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{bmatrix} \times \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$$

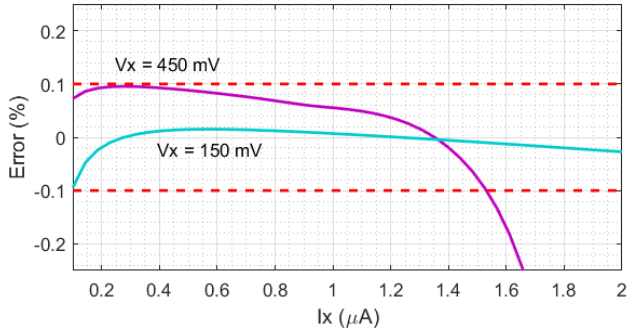

Fig. 4: Error (%) variation with Ix with extended range from 100 nA to 2 μA

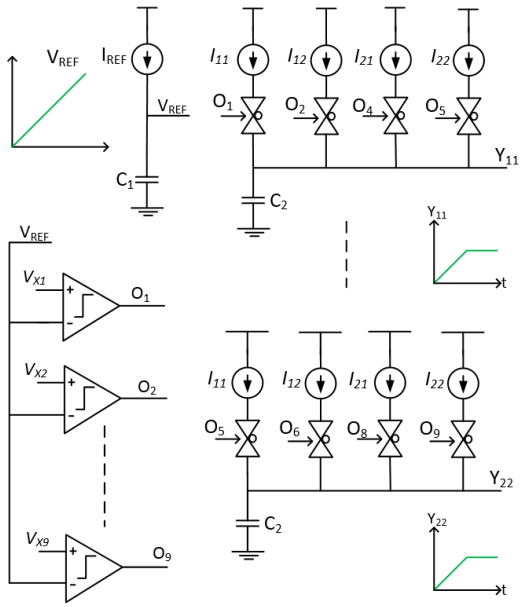

Fig. 5: Matrix multiplication of a 3×3 input matrix with 2×2 kernel. Settling time is 7.5 μs, which is independent of input or kernel size.

where $Y_{ij} = \sum_{m=1}^{N} \sum_{n=1}^{N} I(m,n)V(m+i-1, n+j-1)$

Fig 5 shows the circuit architecture of the proposed VMM. The capacitor C_1 is charged through I_{REF} to generate T_X for all parallel units. Note that all operation is performed in parallel to achieve high computation efficiency. The kernel is placed over each 2×2 block of the input array and it performs the VMM operations on each block simultaneously. The settling time for the entire convolution operation is same as the settling time for a single multiplier cell. Since the currents are represented as weights, they are added together, and charge capacitor C_2. All four outputs are sampled after worst case time corresponding to $V_X = 450\ mV$. Since the kernel is not slid over the entire image, the settling time is 7.5 μs, and is independent of the input array size.

The range of the multiplier is limited by the power supply. Since each output consists of four MAC additions, the range of the VMM is smaller than the multiplier cell range. As discussed in Section III, this range can be extended by trimming the capacitor C_2, and increasing the scaling factor.

Fig 6 shows the percentage error of all four outputs obtained after matrix multiplication. 75 random input matrices of size 3×3 were multiplied with 75 random 2×2 kernels. The comparator delay contributes to the error in the four outputs. Scaling factor is 200 nA without employing range extension, hence current values were limited to $35\ nA - 135\ nA$ range. The 3σ error is found to be 0.11%.

Table I shows comparison of this work with state of the art matrix multipliers. This work achieves much higher precision than previously reported works by using time based analog multiplication. High computation efficiency of 1.72 TFlops/W is achieved.

TABLE I: Performance comparison with state of the art VMM Circuits

Ref	CICC'04 [6]	TCAS'19 [15]	DAC'16 [7]	CICC'17 [9]	ISCAS'18 [16]	DAC'18 [17]	This Work
Type	FG NOR	NOR flash	ReRAM	FG NOR	Charge based	FG NOR	CMOS
Power (μW)	7.2/cell				24 *		2.46
Input Size	128x128	100x100	784x500	10x10	9x1	400x400	3x3
Kernel Size	8x8						2x2
Linearity	$>$ 2 decades	N.R.	N.R.	N.R.	N.R.		1.64 decades **
Error (%) or Precision (bit)	$< 2.5\%$	6 bit	4 bit	2 %	5 bit	5 bit	$\pm$ 0.1 %
EE (POps/J)		0.085	0.06	N.R.	284.4 GOps/W	1.68	1.72 TFlOps/W

* Including ADC ** Multiplier cell, after range extension

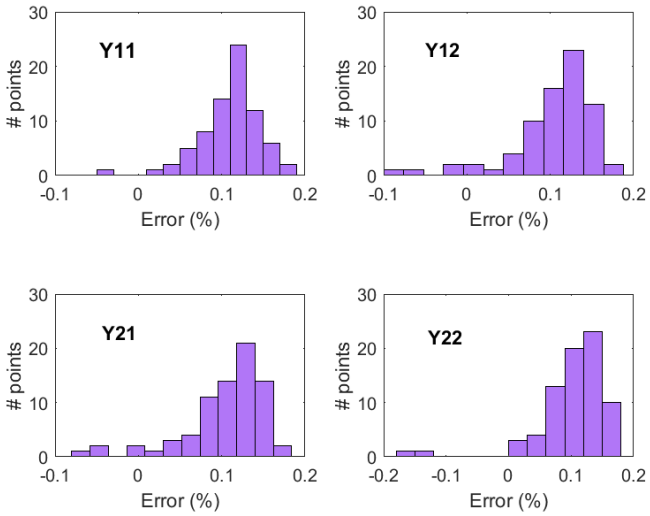

Fig. 6: Error (%) of matrix multiplication for 3×3 input matrix with 2×2 kernel. 75 matrix multiplications were performed with randomized input and kernel matrix values.

IV. Conclusion

In this work, a capacitor based multiplier for performing 2-D convolution operations is presented. The proposed multiplier has small settling time of 7.5 μs which is independent of input matrix size. Error (%) for all outputs is found to be less than $\pm$0.1 % at 27^oC. The error increases at higher temperatures due to switch leakage. Linearity of 0.98 decades is achieved without range extension. Range extension can be obtained by trimming the output capacitor C_2 to achieve range of 1.64 decades.

Acknowledgements

This work was supported in part by the National Science Foundation under award # 1812588.

References

[1] Y. Huang, N. Guo, M. Seok, Y. Tsividis, and S. Sethumadhavan. Evaluation of an analog accelerator for linear algebra. In *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, pages 570–582, June 2016.

[2] N. Mirchandani and A. Shrivastava. High stability gain structure and filter realization with less than 50 ppm/ c temperature variation with ultra-low power consumption using switched-capacitor and sub-threshold biasing. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, May 2018.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[4] A. H. Miremadi, A. Ayatollahi, and A. Abrishamifar. A low voltage low power cmos analog multiplier. In *2011 NORCHIP*, pages 1–4, Nov 2011.

[5] C. Popa. Improved accuracy current-mode multiplier circuits with applications in analog signal processing. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 22(2):443–447, Feb 2014.

[6] R. Chawla, A. Bandyopadhyay, V. Srinivasan, and P. Hasler. A 531 nw/mhz, 128/spl times/32 current-mode programmable analog vector-matrix multiplier with over two decades of linearity. In *Proceedings of the IEEE 2004 Custom Integrated Circuits Conference (IEEE Cat. No.04CH37571)*, pages 651–654, Oct 2004.

[7] M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila, C. Graves, S. Lam, N. Ge, J. J. Yang, and R. S. Williams. Dot-product engine for neuromorphic computing: Programming 1t1m crossbar to accelerate matrix-vector multiplication. In *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pages 1–6, June 2016.

[8] L. Chua. Memristor-the missing circuit element. *IEEE Transactions on Circuit Theory*, 18(5):507–519, Sep. 1971.

[9] X. Guo, F. M. Bayat, M. Prezioso, Y. Chen, B. Nguyen, N. Do, and D. B. Strukov. Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm nor flash memory cells. In *2017 IEEE Custom Integrated Circuits Conference (CICC)*, pages 1–4, April 2017.

[10] E. H. Lee and S. S. Wong. Analysis and design of a passive switched-capacitor matrix multiplier for approximate computing. *IEEE Journal of Solid-State Circuits*, 52(1):261–271, Jan 2017.

[11] Y. Park, H. Kim, J. Oh, Y. Choi, and B. Kong. Compact 0.7-v cmos voltage/current reference with 54/29-ppm/c temperature coefficient. In *2009 International SoC Design Conference (ISOCC)*, pages 496–499, Nov 2009.

[12] A. Shrivastava, N. E. Roberts, O. U. Khan, D. D. Wentzloff, and B. H. Calhoun. A 10 mv-input boost converter with inductor peak current control and zero detection for thermoelectric and solar energy harvesting with 220 mv cold-start and−14.5 dbm, 915 mhz rf kick-start. *IEEE Journal of Solid-State Circuits*, 50(8):1820–1832, Aug 2015.

[13] C. C. Enz and G. C. Temes. Circuit techniques for reducing the effects of op-amp imperfections: autozeroing, correlated double sampling, and chopper stabilization. *Proceedings of the IEEE*, 84(11):1584–1614, Nov 1996.

[14] P. C. H. Meier, R. A. Rutenbar, and L. R. Carley. Exploring multiplier architecture and layout for low power. In *Proceedings of Custom Integrated Circuits Conference*, pages 513–516, May 1996.

[15] M. Bavandpour, M. R. Mahmoodi, and D. B. Strukov. Energy-efficient time-domain vector-by-matrix multiplier for neurocomputing and beyond. *IEEE Transactions on Circuits and Systems II: Express Briefs*, pages 1–1, 2019.

[16] K. Sanni, T. Figliolia, G. Tognetti, P. Pouliquen, and A. Andreou. A charge-based architecture for energy-efficient vector-vector multiplication in 65nm cmos. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, May 2018.

[17] M. R. Mahmoodi and D. Strukov. An ultra-low energy internally analog, externally digital vector-matrix multiplier based on nor flash memory technology. In *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, pages 1–6, June 2018.