PIC-RAM: Process-Invariant Capacitive Multiplier Based Analog In Memory Computing in 6T SRAM

Kailash Prasad, Aditya Biswas, Arpita Kabra and Joycee Mekie
Department of Electrical Engineering, Indian Institute of Technology Gandhinagar, India
{kailash.prasad, adityab, arpita.kabra, joycee}@iitgn.ac.in

***Abstract*—In-Memory Computing (IMC) is a promising approach to enabling energy-efficient Deep Neural Network-based applications on edge devices. However, analog domain dot product and multiplication suffers accuracy loss due to process variations. Furthermore, wordline degradation limits its minimum pulsewidth, creating additional non-linearity and limiting IMC's dynamic range and precision. This work presents a complete end-to-end process invariant capacitive multiplier based IMC in 6T-SRAM (PIC-RAM). The proposed architecture employs the novel idea of two-step multiplication in *column-major* IMC to support 4-bit multiplication. The PIC-RAM uses an operational amplifier-based capacitive multiplier to reduce bitline discharge allowing good enough WL pulse width. Further, it employs process tracking voltage reference and fuse capacitor to tackle dynamic and post-fabrication process variations, respectively. Our design is compute-disturb free and provides a high dynamic range. To the best of our knowledge, PIC-RAM is the first analog SRAM IMC approach to tackle process variation with a focus on its practical implementation. PIC-RAM has a high energy efficiency of about 25.6 TOPS/W for 4-bit $\times$ 4-bit multiplication and has only 0.5% area overheads due to the use of the capacitance multiplier. We obtain 409 bit-wise TOPS/W, which is about 2$\times$ better than state-of-the-art. PIC-RAM shows the TOP-1 accuracy for ResNet-18 on CIFAR10 and MNIST is 89.54% and 98.80% for $4bit \times 4bit$ multiplication.**

I. Introduction

In-memory computing (IMC), an antidote for the Von Neumann bottleneck, is a promising approach to deal with emerging ML workloads, especially in Deep neural networks (DNNs). To enable MAC operation, which is the primary compute operation in DNNs, the IMC architecture must support the direct multiplication of multi-bit input and filter values. Existing analog IMC techniques based on 6T SRAM cells allows multi-bit multiplication for the row-major approach and have limited parallelism.

To perform analog IMC in 6T SRAM, which is most preferred Cache for density reasons, four main issues need to be addressed: (a) **N**on-linearity and low yield, as the computation outputs are based on bit-line discharge and are subject to process variations (b) The practicality of the minimum **W**ordline pulse supported due to wordline degradation because interconnect and parasitics (c) Limited **D**ynamic range as the access transistors need to be in saturation region during the computation (d) **C**ompute disturb, as multiple wordlines are simultaneously activated for each computation. We refer to these as NWDC-challenges.

To the best of our knowledge, none of the existing analog IMC solutions explicitly simultaneously address these issues in the presence of process variations. To circumvent the problem of compute disturb and to improve linearity, 6T SRAM cells are often replaced with 8T1C, 9T1C, 12T1C, etc. [1]–[4]. These cells with additional transistors and a capacitor incur higher area overhead ranging from $30\% - 100\%$.

The existing works on 6T SRAM-based analog IMC architectures based on column-major are limited to dot product multiplication or single-bit multi-bit multiplication [5], [6] due to the stringent circuit requirements for multi-bit multi-bit multiplication. For instance, Zhang et al. [7] report that the minimum WL pulse is 20ps. This short pulse is neither feasible to generate nor transmit to the far-end column of a wide SRAM array due to parasitics capacitance and resistance [8]. In our experiment, we observe that to perform computation with a significant pulse width of 200ps, we need high capacitance at the bitline, approximately 310fF, to maintain the linearity.

To the best of our knowledge, this is the first process invariant column-major analog IMC for multi-bit multi-bit multiplication on 6T SRAM. Our proposed PIC-RAM IMC architecture (Fig. 1) has the following features: (a) it employs process tracking voltage reference and fuse capacitor to tackle process variation dynamically, and post-fabrication, (b) it allows practical wordline pulses enabled by the capacitive multiplier, (c) it supports multi-bit multiplication, (d) it requires minimal change in peripheral circuits and has low area overhead, (e) it is compute-disturb free, (f) it provides high dynamic range and is resilient to process variations. From the results based on post-layout simulations carried out for 256×128 6T SRAM array, we obtain 409 bit-wise TOPS/W, 1.97$\times$ larger than the state of the art [7] 6T SRAM IMC. We tested the proposed analog IMC on two NNs and obtained the Top-1 accuracy comparable to state-of-the-art [5]–[7].

II. Challenges in implementing multi-bit multiplication in column-major 6T SRAM IMC

In analog IMC, weights are stored in a row-major or column-major fashion. In the column-major approach [5], [6], data-1 (typically, filter values) is stored column-wise, data-2 (typically, input value) is given in the word lines, and the operation output is collected from bit-lines. While in a row-major approach [9], data-1 is stored row-wise, data-2 is given in the word lines or bitlines, and the operation output is collected from bit-lines. The column-major approach provides high parallelism and increases the throughput as N multiplications will be performed in one cycle where N is the number of columns.

The input activations are provided onto the word lines in pulse amplitude modulated or pulse width modulated signal with amplitude or pulse width proportional to activation value. The pulse amplitude modulation (AM) technique is less preferred due to the requirement of an expensive digital to analog converter unit in the periphery. Further, at lower technology

978-3-9819263-7-8/DATE23/© 2023 EDAA

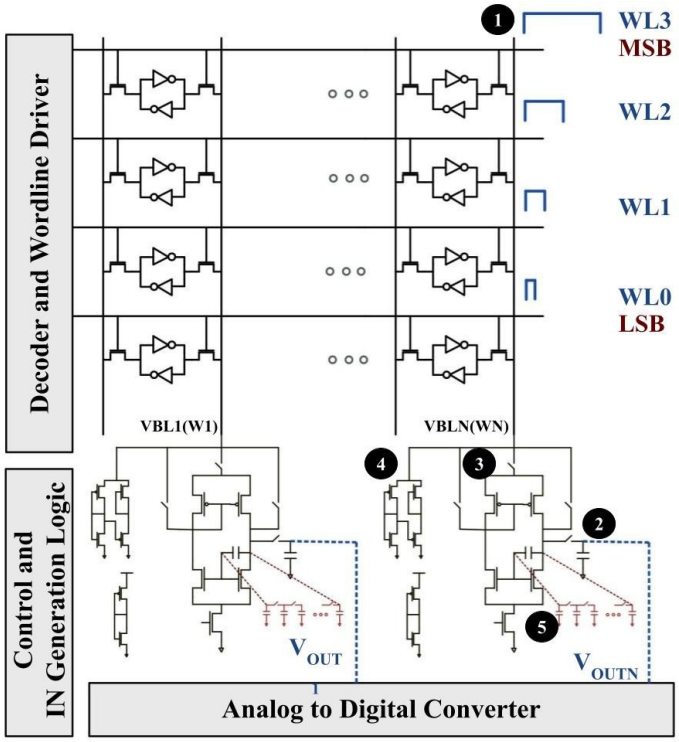


Fig. 1: Proposed Approach (PIC-RAM) for Multi Bit Multiplication : 1) PWM WL for Data-1(Data-1 is stored column-wise in SRAM), 2) Sampling Switch (Data-2), 3) Capacitance Multiplier Logic, 4) Process Variation Tracking Voltage Reference and 5) Fuse Capacitor for handling post fabrication variations

nodes, the dynamic range is reduced, and generating the voltage for a higher bit is challenging because of a lower supply voltage range. Hence, AM technique is not feasible at lower technology nodes, and giving input as the pulse width modulated WL signal is a preferred option.

To perform multi-bit multiplication, we simultaneously turn on multiple WL, which requires a high column current. There is a higher chance of degradation in linearity because the large current flows through the bitline. The bitline discharges quickly, and the access transistor enters the triode region, thus limiting the multiplication product's linearity and accuracy since the product is directly proportional to the bitline discharge. We can control the bitline discharge rate and prevent it from entering the triode region by allowing it to discharge for a shorter time using a shorter pulse width.

Our simulation results of column-major IMC with PWM input show that the bitline discharge is very fast, and to have a bitline of 350mV for accurate multiplication of 15(Data-1)x15(Data-2), the maximum pulse-width allowed would be 31.5 ps, which is not feasible. [7] propose a strategy that decomposes multi-bit multiplication into combinations of binary operations based on bitline shifting. The voltage differences can be stored at corresponding capacitances instead of at one point. In this work also, the required pulsewidth reported is 20ps which is not practical.

Since it is not possible to generate the required pulse width mentioned in [7], [8], we need to add a large capacitance

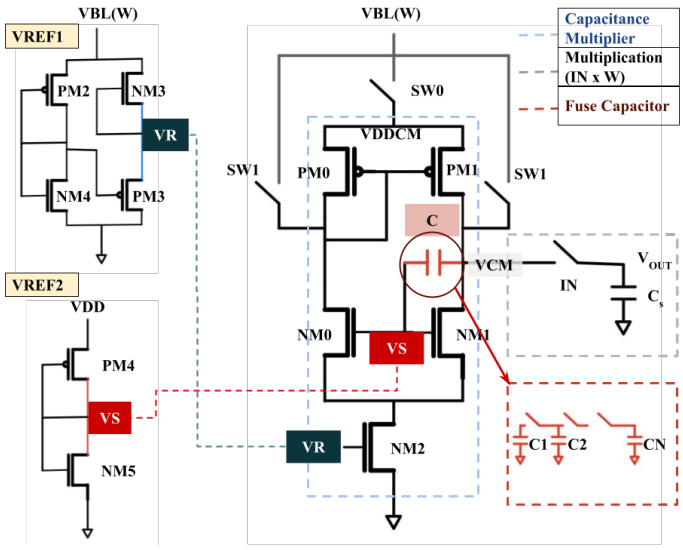


Fig. 2: Proposed Capacitance Multiplier

to support larger WL pulse-widths to slow down the bitline discharge. Based on our post-layout simulations, for 200 ps WL pulse width, a huge bitline capacitance of 310fF is required to maintain the access transistor in the saturation region throughout the operation. Also, this approach makes the decoder design highly complex as it will have to generate word-line with varying pulse widths in the ratio of 8:4:2:1.

Further, analog domain operations have a limited dynamic range due to the nonlinear current(I_{DS}) characteristics of the access transistor. Furthermore, all the analog-based IMC approaches are prone to process variation, degrading the results' linearity and dynamic range. In these implementations, large read current variations due to process variation and mismatch limit the maximum parallelism and throughput. Many SRAM cells with capacitors [1]–[4], [10] have been proposed to increase parallelism in the presence of device mismatch. However, they still fail due to process variation and mismatch due to access transistors, peripherals, and interconnects.

III. Proposed PIC-RAM IMC Architecture

This section discusses the proposed PIC-RAM IMC architecture for handling the NWDC-challenges mentioned in Section 1. Fig. 1 shows our proposed analog IMC architecture. It consists of five sub elements.

A. Column Major Wordline and Wordline Underdrive

The first part of the proposed IMC is to get binary-weighted Data-1(4-bit) stored across the column in the 6T SRAM array to the bitline. Multiple wordlines (4) are activated simultaneously to get Data-1 on bitline. For an m-column array, m Data-1 values are obtained on the respective bitlines. Compute disturb issue in our IMC is eliminated by using word-line underdrive (WLUD). The WL pulses ratio-ed as 8:4:2:1 for a 4-bit data are WLUD pulses ❶ in Fig. 1. Bit-line will discharge based on the combined effect of the data stored in the SRAM cells and the pulse widths of PWM WLs. That is, the LSB shall contribute least to the bit-line discharge. The resultant voltage in the bitline will represent Data-1.

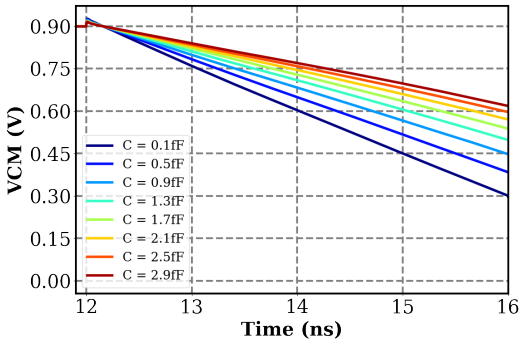


Fig. 3: VCM discharge rate for different fuse capacitor values

B. Sampling Circuit

The most important contribution of this work is the small additional sampling circuit ❷ in Fig. 1 consisting of a switch (IN) and a small capacitance that captures the charge equivalent of the product term. This is how the multiplication takes place. A 4-bit Data-1 is stored bit-wise in the column. When WL pulses are given, as shown in Fig. 1, the bit-line with large capacitance discharges based on the data value stored in the column. After this, the sampling switch (IN) is activated for a duration equivalent to 4-bit Data-2. The sampling capacitor (Cs) stores the charge equal to the product of Data-1 and Data-2. This technique eliminates the need for pulse width/amplitude modulation of the WL pulses based on the data value, and thus the decoder circuit needs minimal modification for PIC-RAM IMC. However, one additional decoder is required to provide the input (Data-2) to the sampling switch.

TABLE I: Area Comparison of Capacitor with equivalent MOM(Metal-Oxide-Metal) capacitor

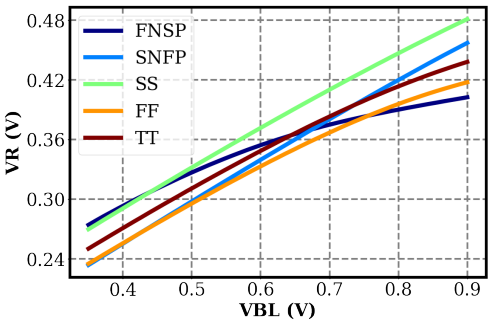
C (fF)	Ceq (fF) (V_{CM})	Area of Ceq without CM (um x um)	Area of Ceq With CM (um x um)
0.1	350	17.6	2
0.5	390	19.6	2.1
0.9	470	23.6	2.2
1.3	570	28.6	2.32
1.7	670	33.7	2.45
2.1	770	38.7	2.56

C. Capacitance Multiplier(CM)

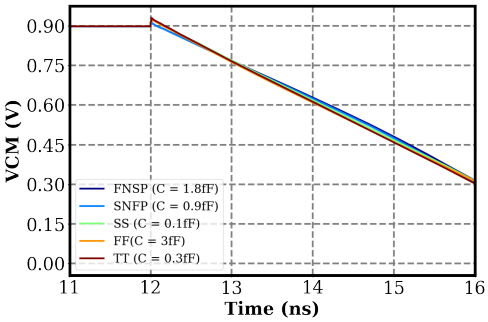
The second important contribution of the proposed architecture is the use of capacitance multiplier circuit ❸ in Fig. 1 as a replacement for the large capacitor on the bit-line, which reduces area. The detailed circuit is shown in Fig. 2. The bit-line voltage is transferred to the capacitive multiplier via three paths. The results are taken from the output node (V_{CM}) of the capacitive multiplier circuit. The proposed capacitive multiplier employs an active operational amplifier to multiply the small capacitance value provided by a passive capacitor. The circuit works in voltage feedback mode, and through the Miller effect, the voltage applied to one terminal of the capacitor is amplified by a gain of (-K) and applied to its other terminal. Hence, a capacitive amplification factor is equal to K.

The equivalent capacitance at the V_{CM} node is given by,

$$Ceq = (1 + A) * C = (1 + gm * ro) * C \quad (1)$$



(a) V_R Vs VBL (Highest V_R at SS corner and Lowest V_R at FF corner



(b) V_{CM} node discharge with Different Capacitor(Same VCM node discharge rate across corners)

Fig. 4: Capacitance Multiplier in various Process Corner

where A = gm * ro is the low-frequency voltage gain, gm is the transconductance, and ro is the output resistance. Fig.3and Table I show the capacitance multiplication and area reduction. Compared to an MOM capacitor the capacitance multiplier can achieve area benefit from $8\times$ *to* $17\times$.

The transistors PM0, PM1, NM0, NM1, and NM2 form a single-stage operational transconductance amplifier. The miller capacitance is the capacitor C connected between the gate and drain of the transistor NM1. The PM0 and PM1 are self-biased through the current mirror at PM0. The NM0 and NM1 are biased with VS voltage generated from voltage reference 2. The NM2 is the tail current source, biased by BL-dependent voltage reference (VREF 1) generating V_R voltage. V_{CM} is the output node of the capacitance multiplier. The NM2 Transistor operates in the subthreshold regions, thereby consuming extremely low power. The capacitance multiplication can be seen from Fig. 3 where a small change in the capacitor is causing a large change in V_{CM} discharge.

D. Process Variation Tracking Voltage Reference

One of the most striking features of our proposed circuit is that of process tracking circuit VREF1 ❹ in Fig. 1. The detailed circuit shown in Fig. 2 makes use of voltage V_R, which is dependent on bitline voltage and changes based on process variations. The gain of the amplifier changes based on V_R, which in turn adjusts the gain of the amplifier and the slope of common node V_{CM} voltage, thus tracking the process variations. Usually, in the FF corner, the V_{CM} discharge is fast,

and in the SS corner, it is slow. In order to reduce the process variation, V_{REF1} generates higher V_R for SS corner for fast discharge and lowers V_R for FF corner for slow discharge, as shown in Fig. 7. The other source V_S is generated by a diode-connected voltage reference circuit, and it is common to all the capacitance multiplier (CM) circuits connected to all the columns.

E. Fuse Capacitor

Finally, to adjust for post-fabrication variations, we use ten fuse caps 0.3 fF each with fuse ❺ in Fig. 1. The ten fuse caps will be employed with the capacitive multiplier per column. The fuses will be programmed post-fabrication with a small ROM unit. Fuses [11] is a technique well-adapted in industry. We have performed simulations in all five corners, and Fig. 4(b) shows the same V_{CM} discharge with capacitor value in each corner. The rate of discharge is constant across corners.

IV. Working of the Proposed IMC Architecture

We have implemented the complete IMC architecture with a 256×128 SRAM cell array with all the required peripheral circuits. Filter weights (W) are first stored bit-wise in the columns of the SRAM array. The working of the entire IMC architecture for multibit multibit multiplication can be divided into three steps.

- Multibit single bit Multiplication
- Precharge
- Multibit Multibit Multiplication

Multibit single-bit Multiplication: Multi-bit to single-bit multiplication is done by turning on four wordlines with pulses in the ratio 8:4:2:1 as shown in Fig. 4. Thus, the bitline discharges correspond to the data stored in the column. SW0 is also activated at T0 to precharge the VDDCM node to VBL.

Precharge: At T1, the SW1 switch is activated to precharge the V_{CM} node of the capacitive multiplier to VBL. At the same time, VBL is provided to VREF 1, and it generates V_R voltage corresponding to VBL. The SW0 switch is turned off in this phase, making VDDCM a floating node. It allows V_{CM} to discharge in the third phase.

Multibit Multibit Multiplication: In the last phase at T2, SW1 is deactivated, allowing V_{CM} to discharge from precharged voltage with a slope depending on V_R voltage. The pulse width modulated input IN (which corresponds to the input activation data) is provided, which charges the output capacitor C_S with V_{CM} voltage. When the input goes low, multiplication is stored at the V_{OUT} node. The final output voltage V_{OUT} is sampled at T3 across the capacitor C_S resulting in the voltage corresponding to the product of multi-bit multi-bit multiplication.

Fig. 6 further explains how the multiplication operation is performed using the proposed approach. The discharge of the V_{CM} node is maximum for W=15(Data-1). There is no discharge in bitline, but at the same time, V_R is maximum, resulting in the highest slope. When W stored is 1, the bitline will discharge a maximum of 350 mV, and V_{CM} will discharge with the lowest slope. The difference in BL voltage varies

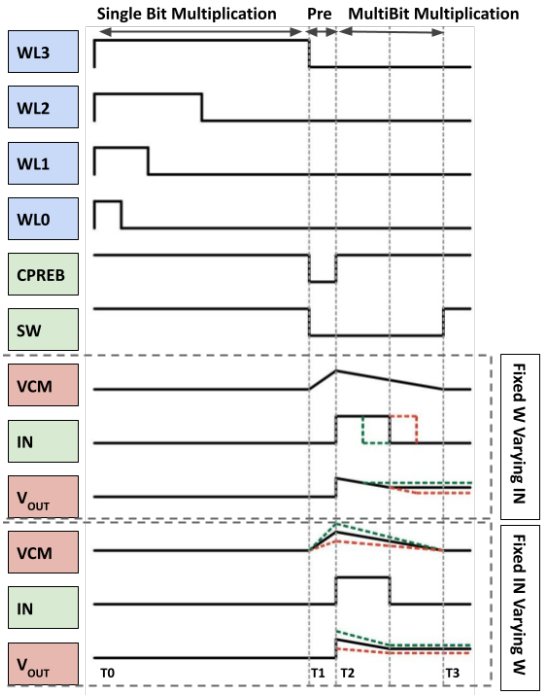


Fig. 5: Timing Diagram for Multibit Multiplication

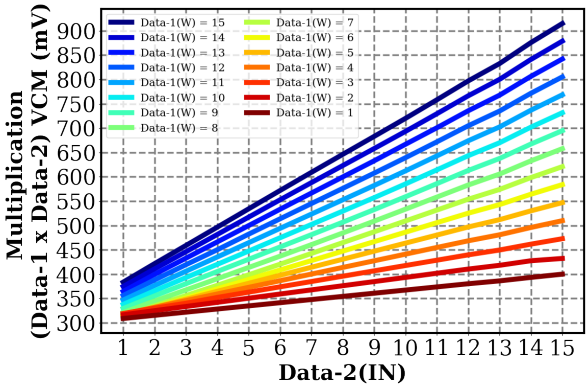


Fig. 6: MultiBit MultiBit ($4b \times 4b$) Multiplication result for PIC-RAM

linearly, with ΔV for W ranging from 1 to 15. The V_{CM} discharges ΔV for W = 1 and 15 ΔV for W=15. This provides the required priority of $8:4:2:1$ to the input given to IN.

V. Analysis of the Proposed PIC-RAM

A. Experimental Setup

We have performed the post-layout simulation of 256×128 Array in CMOS 28nm Technology at 0.9V, 27C. We have used edge driving to eliminate wordline degradation [8]. The capacitance multiplier circuit is connected to the bitline of every column have 0.5% area overhead. The bias voltage is provided with a voltage reference, and Control Logic generates the timing signals. The IN input is generated similar to [6]. The analog-digital converter (ADC) digitizes the analog output voltage at V_{OUT} to the digital output.

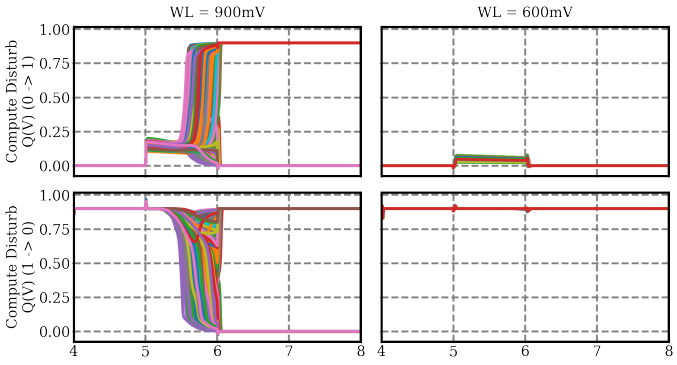


Fig. 7: Compute Disturb (No Compute Disturb at 600mV)

B. Linearity of Multiplication

The proposed circuit is highly linear due to a linear discharge of the V_{CM} node. All the transistors of the capacitance multiplier are in the saturation region for the entire duration of the computation, result enhancing the linearity. The analog multiplication is highly linear with a wide dynamic range(290mV-900m). The tradeoff between dynamic range and linearity has been taken care of by the capacitance multiplier. Fig. 6 shows the linearity in the multiplication of all possible 4b x 4b multiplication between activation and weight.

C. Mitigating Process Variation Effects

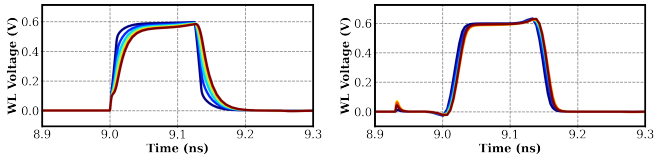
The proposed PIC-RAM explicitly handles the process variation using two techniques. In the first, a process tracking voltage reference is designed. The voltage reference tackles the process variation as it generates higher V_R for the SS corner and lower V_R for the FF corner. The discharge rate of the V_{CM} node depends on the V_R voltage applied to the tail transistor of the capacitive multiplier. Higher V_R at SS corner increases the V_{CM} discharge rate while lower V_R at FF corner reduces the V_{CM} discharge rate. The change in V_R compensates for the V_R discharge change due to process variation.

Second, to further handle process variations, we are using 10 0.3 fF with fuse such that after fabrication, we can get equivalent capacitance for every corner for the same discharge rate. The solution is possible here because the capacitor requirement is minimal. The capacitors are MOM(Metal-Oxide-Metal) capacitors, and the fuses are transmission gates. This is a well-accepted technique in industry and academia [11]. We have performed simulations in five corners, and Fig. 4 shows the same VM discharge with capacitor value in each corner.

D. Wordline Underdrive to Improve Dynamic Range and Remove Compute Disturb

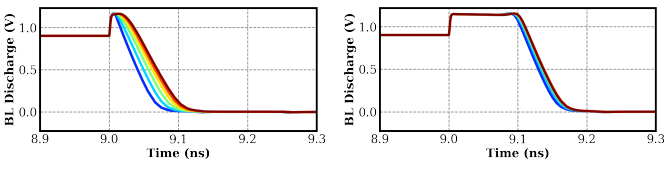
We have also used wordline underdrive further to improve the linearity and energy efficiency of the architecture. High voltage on WL leads to non-linearities in the discharge of BLB, as I_{DS} through access transistors of 6T-SRAM bitcell is directly proportional to the WL voltage. As fast the BLB discharges, the slower its discharging rate becomes due to the parasitic capacitance of bit-lines. Therefore, we use the word-line under-drive technique to increase linearity in the multiplication operation. The write and conventional read operations are performed at 0.9 V, whereas the in-memory computation is at 0.6 V. The dynamic range improves because of the wordline underdrive as the overdrive voltage ($V_{gs} - V_{th}$) of the access transistor goes down. The nominal V_{th} is around 450 mV. So underdriving the wordline from 900mV to 600mV makes overdrive voltage go down from 450mV to 150mV. It means that even if the bitline discharges to 150mV, the access transistor still provides the same current maintaining the linearity.

Further in our approach, we activate multiple wordlines simultaneously, increasing the chances of data flip during computation. To study the compute (read) disturbance for the overlapping condition when all the wordlines are activated and adjacent cells store opposite bits (one storing digital high and the other storing digital low), we perform 2000-point Monte Carlo simulations of the array at wordline voltage of 900mV and 600mV to observe the number of times the bit stored in the internal node flips. To remove the read disturbance, we underdrive the wordline to 600mV. Fig. 7 shows the occurrence of compute disturb at higher gate voltage due to large current flow through this short circuit path. However, at 600mV, no data flip is observed. Further, the maximum change in voltage during functional read operation at node storing 0 is 80mV, and the maximum change in node storing 900mV is 30mV. This analysis shows that the proposed multiplication approach is robust to compute disturbs. We have further tested up to 64-row activation simultaneously for the worst case (1 bitcell storing 0 and others storing 1 and vice versa), and under 2000-point Monte Carlo simulation, there is no compute disturb in any of the bit cells.



(a) WL Distortion (b) WL with WL Shaping

Fig. 8: Wordline Distortion due to parasitic resistance and capacitance



(a) Without WL Shaping (b) With WL Shaping

Fig. 9: Variation in Bitline (BL) Discharge across columns

E. Wordline Pulse Improvement

Wordline degradation is one of the critical sources of non-linearity in Analog IMC. When the same data is stored in all the columns of a row, the expected BL and BLB discharge for all columns should be the same. However, due to WL pulse distortion, there is non-linearity in discharge, as shown in

TABLE II: Comparison with Existing Literature

	PIC-RAM	TCAS'20 [9]	TCAS'21 [12]	JSSC'21 [13]	JSSC'21 [14]	DAC'21 [15]
Bitcell	6T	6T	6T	6T	8T	8T
Technology(nm)	28	65	28	65	40	28
VDD (V)	0.9	1.2	0.9	1.2	-	1.1
Array Size	256 x 128		64 x 64	512 x 128	512 x 64	256 x 64
Activation precision	4	4	4	4	Arbitrary	2
Weight precision	4	4	3	1/2/3/4/5/8	Arbitrary	8
Frequency (MHz)	167	-	282	70		125
Energy Efficiency (TOPS/W)	25.6 (A:4b W:4b)	-	25.9 (A:4b W:2b)	49.4 (A:4b W:2b)	17 (A:4b W:5b)	11.2 (A:2b W:8b)
Bitwise Energy Efficiency (TOPS/W)	**409.6**	-	207.2	319.2	340	179.2
CIFAR 10 Accuracy	89.54%	88.83%	-	89.00%	88.5%	91.4%
MNIST Accuracy	98.80%	99.05%	98.12%	98.80%	98.2%	-
Process Variation Tracking	**Yes**	No	No	No	No	No
Wordline Improvement	**Yes**	No	No	No	No	No

Fig. 9(a). We have used Central Spine architecture for memory subsystem design. This reduces the wordline degradation across the columns. To further improve the WL pulse, we strapped the WL at the end and ran a parallel dummy WL with higher metal with driving the wordline from both the end [16]. The main reason for using higher metal is to reduce the resistance and make dummy WL faster. Driving the WL from the far end improves the BL discharge deviation, as can be seen Fig. 9(b).

VI. Comparison and Results

Table II shows a detailed comparison between the proposed IMC with existing IMC architecture. The proposed architecture achieves the energy efficiency of 25.6 TOPS/W for 4b X 4b multiplication at 167 MHz. We obtained the CNN accuracy for the CIFAR-10 and MNIST datasets by mapping the circuit results into a neural network simulator. The 4bit multiplication with ADC non-linearity is modeled on the simulator. Here, the ADC resolution is set to 5bit to retain accuracy. We show the quantization accuracy results for ResNet-18 pretrained for CIFAR10 and MNIST classification. For uniform symmetric quantization, TOP-1 accuracy is 89.54% and 98.80% for 4bit. The proposed PIC-RAM reports higher accuracy due to improvement in linearity and handling of the process variation. Further, we used the bitwise figure-of-merit (BFOM) concept for a fair comparison, as proposed in [13]. The bitwise BFOM is defined as TOPS/W × activation bit × weight bit. The bitwise energy efficiency in the case of our work is higher than that of other reported studies.

VII. Conclusions

This paper discusses the process-invariant column-major analog IMC for multi-bit multiplication on 6T SRAM(PIC-RAM). The PIC-RAM employs process tracking voltage reference and fuse capacitor to tackle process variation dynamically and post-fabrication. The PIC-RAM proposes a capacitive multiplier circuit to perform 4-bit multiplication using a two-step approach. The capacitive multiplier allows a practical wordline pulse with 0.5% area overhead. The PIC-RAM is compute-disturb free and provides a high dynamic range. From the results based on post-layout simulations carried out for 128x256 6T array, we obtain 409 bit-wise TOPS/W, 1.97× larger than the state of the art [7] 6T SRAM IMC.

VIII. Acknowledgement

This work is supported through grants received from Prime Minister Research Fellowship (PMRF), SMDP-C2SD and YFRF Visvesvaraya Ph.D. scheme from the Ministry of Electronics and Information Technology (MEITY), and through SERB grants CRG/2018/005013, MTR/2019/001605, SPR/2020/000450 and Semiconductor Research Corporation (SRC), through contract 2020-IR-2980.

References

[1] Mu *et al.*, "SRAM-Based In-Memory Computing Macro Featuring Voltage-Mode Accumulator and Row-by-Row ADC for Processing Neural Networks," *TCAS I*, 2022.

[2] Jiang *et al.*, "C3SRAM: An In-Memory-Computing SRAM Macro Based on Robust Capacitive Coupling Computing Mechanism," *JSSC*, vol. 55, no. 7, pp. 1888–1897, 2020.

[3] B. Iqbal *et al.*, "A Process and Data Variations Tolerant Capacitive Coupled 10T1C SRAM for In-Memory Compute (IMC) in Deep Neural Network Accelerators," in *AICAS*, 2022, pp. 459–462.

[4] B. Zhang *et al.*, "A 177 TOPS/W, Capacitor-based In-Memory Computing SRAM Macro with Stepwise-Charging/Discharging DACs and Sparsity-Optimized Bitcells for 4-Bit Deep Convolutional Neural Networks," in *CICC*, 2022, pp. 1–2.

[5] Kang *et al.*, "Deep In-Memory Architectures in SRAM: An Analog Approach to Approximate Computing," *IEEE Proc*, vol. 108, no. 12, pp. 2251–2275, 2020.

[6] Zhao *et al.*, "Configurable Memory With a Multilevel Shared Structure Enabling In-Memory Computing," *TVLSI*, pp. 1–13, 2022.

[7] Zhang *et al.*, "In-Memory Multibit Multiplication Based on Bitline Shifting," *TCAS II*, vol. 69, no. 2, pp. 354–358, 2022.

[8] Prasad *et al.*, "Analysis of word line shaping techniques for in-memory computing in srams," in *ICECS*. IEEE, pp. 1–6.

[9] Ali *et al.*, "IMAC: In-Memory Multi-Bit Multiplication and ACcumulation in 6T SRAM Array," *TCAS I*, vol. 67, no. 8, 2020.

[10] Z. Chen *et al.*, "DCT-RAM: A Driver-Free Process-In-Memory 8T SRAM Macro with Multi-Bit Charge-Domain Computation and Time-Domain Quantization," in *CICC*, 2022, pp. 1–2.

[11] Karl *et al.*, "17.1 a 0.6 v 1.5 ghz 84mb sram design in 14nm finfet cmos technology," in *ISSCC Digest*. IEEE, 2015, pp. 1–3.

[12] J. Zhang *et al.*, "In-memory Multibit Multiplication Based on Bitline Shifting," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2021.

[13] Chen *et al.*, "CAP-RAM: A Charge-Domain In-Memory Computing 6T-SRAM for Accurate and Precision-Programmable CNN Inference," *JSSC*, 2021.

[14] S. Jain *et al.*, "±CIM SRAM for Signed In-Memory Broad-Purpose Computing From DSP to Neural Processing," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 10, pp. 2981–2992, 2021.

[15] K. Lee, S. Cheon, J. Jo, W. Choi, and J. Park, "A charge-sharing based 8t sram in-memory computing for edge dnn acceleration," in *2021 58th ACM/IEEE Design Automation Conference (DAC)*, 2021, pp. 739–744.

[16] V. Nautiyal *et al.*, "Self-timed shaper circuit for wide memories in advanced cmos technologies," in *ISCAS*, 2018, pp. 1–5.