Process Variation Resilient Current-Domain Analog In Memory Computing

Kailash Prasad, Sai Shubham, Aditya Biswas and Joycee Mekie
Department of Electrical Engineering, Indian Institute of Technology Gandhinagar, India
{kailash.prasad, shubham.sai, adityab, joycee}@iitgn.ac.in

***Abstract*—In-Memory Computing (IMC) has emerged as one of the energy-efficient solutions for data and compute-intensive machine learning applications. Analog IMC architectures have high throughput, but limited bit precision. Process variation further degrades the bit-precision. This work proposes an efficient way to track process variation and compensate for it to achieve high bit-resolution, which, to the best of our knowledge, is first such proposal. PV tracking is achieved by using an additional SRAM column and compensation by a non-conventional word-line driver. The proposed circuit can be augmented to any analog IMC architecture to make it resilient to process variations. To demonstrate the versatility of the proposal, we have implemented and analyzed 2-bit dot product operation in IMC architectures with six different SRAM cell configurations, and 2-bit, 4-bit, and 8-bit dot product on 6T SRAM IMC. For these, we report a reduction of 4× to 14× in the standard deviation of statistical variations in bit-line voltage for different SRAM cells, increase in the bit-resolution from 2 bits to 4 bits or 6 bits.**

I. Introduction

Machine learning applications using Deep Neural Networks (DNNs) require billions of multi-bit multiply-accumulate (MAC) operations. In memory computing (IMC) has emerged as one of the promising alternatives to the von-Neumann architecture as it allows computations to be performed in memory cells directly providing multifold benefits in computation time and energy. Analog IMC (AIMC) architectures, such as [1], perform the computation in the analog domain, and use analog to digital converters (ADCs) to obtain the final results in digital form. Although AIMCs have high throughput, they are energy efficient only if low resolution ADCs are used resulting in low bit precision [2]. The bit precision is further impacted due to process variations in current-domain AIMC. This work presents a novel method to track process variation and compensate for it. To the best of our knowledge, this is the first proposal that dynamically tracks the process variation and compensates for it. Our proposal requires minimal area overhead (less than 0.2%) and can be augmented with most existing current-domain AIMC architectures with ease.

II. Proposed CWP design

Fig. 1 shows the proposed additional column-based word-line driver for process variation tracking (CWP) circuit and its connection to the SRAM array of an AIMC architecture. Some of the peripherals of the AIMC architecture (including replica column) have not been shown. The control signal AC_WL pulse turns on the four word-lines (WLs) of the additional column (AC) which stores bit-0. AC_BL which was precharged to supply voltage (VDD) discharges to different voltages based on the process variation. For *fast* corner, AC_BL discharges more and for *slow* corner it discharges less. The WL driver used to drive the SRAM array has AC_BL as its supply. This ensures that a PV compensated wordline (CWL) given to the SRAM cells. CWL voltage is higher for *slow* process, and lower for *fast* process. A detailed timing diagram of CWP-AIMC is shown in Fig. 2. We perform 2-bit dot product and observe the results of 2000 point Monte Carlo simulations to capture the effects of process variations. On application of AC_WL pulse, AC_BL falls to different values depending on the process variation. CWL0 and CWL1 which are generated from the word-line drivers to compensate for the process variations, are turned on simultaneously to perform dot

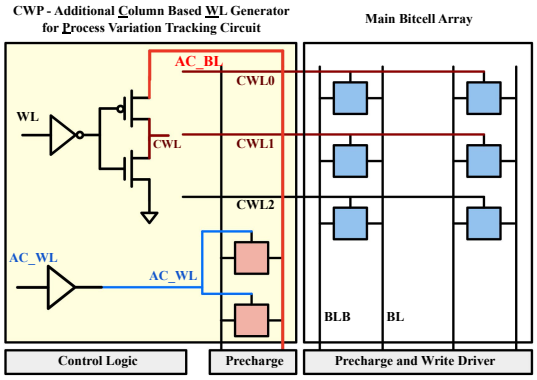

Fig. 1. Proposed CWP-AIMC architecture

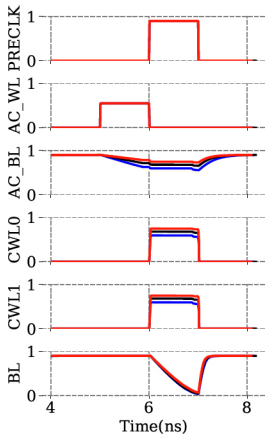

Fig. 2. Timing Diagram of CWP-AIMC

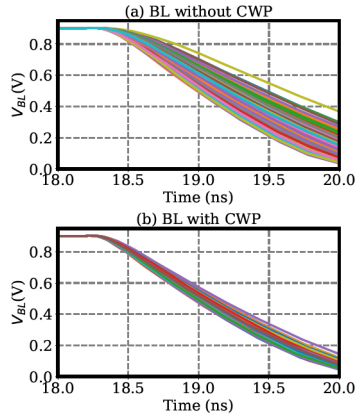

Fig. 3. BL Discharge in (a) AIMC and (b) CWP-AIMC for 2-bit dot product

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.

978-3-9819263-7-8/DATE23/© 2023 EDAA

product of bits stored in Row0 and Row1. The throughput of AIMC is very high as all the columns simultaneously perform the computation. The output of the dot product appears on BL as analog voltage which is then given to ADC to get the final digital value. The 2-bit dot product $(a_0.b_0 + a_1.b_1)$ will result in a 2-bit result. As an example, a_0 and a_1 are given on the wordlines (WL) and b_0 and b_1 are stored in the IMC array. For 4-bit dot product, 4 WLs are turned on and 4-rows store four bits. The BL voltage corresponds to the result of 4-bit dot product, and is converted into 4-bit output through ADC. Similarly, n-bit dot product will result in an n-bit output. Fig. 3(a) shows large variations in BL voltage. This limits the bit-resolution. However, BL voltage shows much lesser variation after applying CWP. Fig. 3(b) shows how process variations impacts the AIMC. Thus, PV tracking and compensation is achieved.

III. Impact of PV-Resilient IMC

We have implemented the complete 128x128 array CWP-AIMC architecture in CMOS 28nm and carried out detailed post-layout simulations. We have performed a 2-bit, 4-bit and 8-bit analog dot product, where the data stored is in a column-major (transposed) fashion [3]. To perform the analog dot product, we turn on multiple WLs simultaneously. The WL voltage is under-driven to $550mV$ to prevent read-disturb and to slow down the BL discharge for proper sampling. This value corresponding to *fast* corner which leads to maximum BL discharge.For 2-bit, 4-bit and 8-bit dot product, the WL pulses are adjusted to adjust the BL discharge. We have analyzed the proposed CWP-AIMC with six different SRAM cells, namely 6T, 6T-DWL [4], 7T [5], CONV8T [6], LIU-D9T [6], and NOG-D10T [6]. Fig. 4 shows the large variations in BL voltage observed across corners for various AIMC architectures. In fact, the difference between BL voltage across corners is quite large, ranging from 350mV to 450mV, the standard deviation (SD) in the variation ranging from 26.43 mV to 28.23 mV. In contrast, CWP-AIMC reduces the variations in BL discharge significantly, ranging from 20 mVto 40 mV, as shown in Fig. 5, and SD ranging from 1.93 to 6.59 for 2-bit dot product for the architectures using six different SRAM cells. We also show comparison of multi-bit dot products for 6T SRAM based AIMC architecture. As captured in Fig. 6, BL dicharge variations for 4-bit and 8-bit dot products also reduces significantly with CWP-AIMC. The SD variation for 4-bit and 8-bit reduces from 29 mV and 31 mV, to 3.1 mV and 6.1 mV, respectively. This, in turn, increases the bit-resolution as well. We have calculated the bit-resolution of analog computation for all the cells under 6σ process variation where,

$$\text{Resolution bits} = log_2 \frac{(\text{Dynamic Range})}{6\sigma}$$

Table I shows the comparison AIMC architectures and CWP-AIMC architectures of six different SRAM cells. The dynamic range for the analog dot product is fixed to 600mV (900mV-300mV), as the computation becomes non-linear below 300mV. In CWP-AIMC, we see an increase in resolution

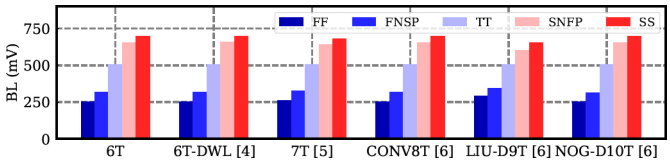

Fig. 4. Bitline Voltage (in mV) across corners for AIMC with different SRAM cells

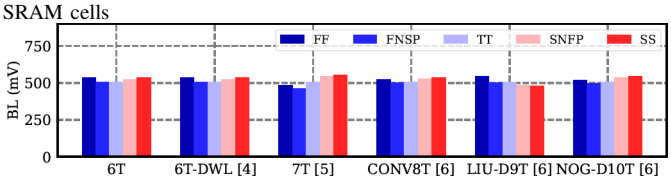

Fig. 5. Bitline Voltage in (mV) across corners for CWP-AIMC with different SRAM cells

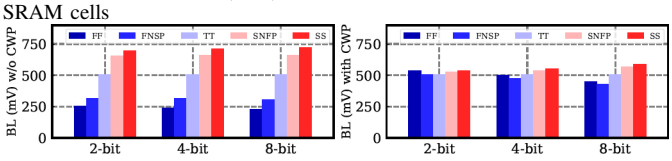

Fig. 6. Bitline Voltage (in mV) across corners for 2-bit, 4-bit and 8-bit dot product in 6T SRAM AIMC (left) and CWP-AIMC (right)

TABLE I
Standard Deviation of BL voltage and Resolution for various memories without and with SCL

	Variation captured as SD(σ) (mV)			Resolution (bits)		
Cell	w/o CWP	with CWP	Reduction	w/o CWP	with CWP	Improvement
6T	27.28	1.93	14.2×	2	6	4
6T-DWL [4]	27.29	2.00	13.6×	2	6	4
7T [5]	28.23	6.59	4.3×	2	4	2
CONV8T [6]	27.21	2.22	12.2×	2	6	4
LIU-D9T [6]	26.43	5.70	4.6×	2	4	2
NOG-D10T [6]	27.22	2.12	12.9 ×	2	6	3

to 4-bit or 6-bit for different SRAM cells used in AIMC. As IMC architectures are targeted largely to DNN applications, the increased resolution directly relates to increased accuracy [7] in ML applications.

IV. Conclusions

The additional column-based WL driver for process variation tracking (CWP) technique proposed in this work effectively reduces the effects of process variations in current-domain analog IMC (AIMC) architectures. We show that CWP-AIMC architectures achieves 4-bit or 6-bit resolution for 2-bit dot product as compared to 2-bit resolution in original AIMC architectures due to large process variations. Further, our proposed techniqe can be easily augmented with any existing current-domain AIMC.

References

[1] Z. Jiang *et al.*, "C3sram: An in-memory-computing sram macro based on robust capacitive coupling computing mechanism," *JSSC*, 2020.

[2] C.-J. Jhang *et al.*, "Challenges and trends of sram-based computing-in-memory for ai edge devices," *TCAS I*, 2021.

[3] M. Kang *et al.*, "Deep in-memory architectures in sram: An analog approach to approximate computing," *Proceedings of the IEEE*, 2020.

[4] P. K. Bharti *et al.*, "Compute-in-memory using 6t sram for a wide variety of workloads," in *ISCAS*, 2022.

[5] H. Jiang *et al.*, "Cimat: A compute-in-memory architecture for on-chip training based on transpose sram arrays," *TC*, 2020.

[6] S. Ahmad, N. Alam, M. Hasan, and B.-S. Kong, "A comprehensive review of design challenges and techniques for nanoscale sram: A cell perspective," 2022.

[7] J. K. Devnath *et al.*, "A mathematical approach towards quantization of floating point weights in low power neural networks," in *VLSID*, 2020.