A DRAM-Based Process-in-Memory Using Data Redundancy and Differential Bit-Line Computation

Hyein Yoon[1], Donghwan Kim[2], Giwoo Lee[2] and SeongHwan Cho[a,2]
[1]Samsung Electronics, Korea
[2]Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Korea
E-mail : [1]hyein0311hy@gmail.com

***Abstract* - This paper presents a novel Dynamic Random Access Memory (DRAM) - In Memory Computing (IMC) structure achieving high throughput without altering the existing cell configuration. Multiple Word Line (WL) activations are utilized to enhance the throughput of Multiply and Accumulate (MAC) operations. The issue of data destruction during simultaneous WL activations is addressed by employing the adjacent MAT and differential operation of the sense amplifier. Moreover, the problem arising from non-ideality in WL switches is alleviated through the differential bit-line computation operation. Consequently, an accuracy of 99.01% was achieved on the MNIST dataset.**

***Keywords*—DRAM-Based Process-In-Memory, Multiply-and-accumulate (MAC), Multiple word-line activation**

I. INTRODUCTION

Due to recent advancements in machine learning algorithms, the increasing number of parameters used in high-complexity applications such as video classification and object detection is causing traditional processors to consume significant power and latency in data computation due to the von Neumann bottleneck. Among various processor-level solutions, Process-in-Memory (PIM) stands out for alleviating the Von Neumann bottleneck through massive parallel processing and reduced data volume passing through I/O interfaces. From the perspective of memory devices used in Process-in-Memory (PIM), various memory devices such as Flash, DRAM, and SRAM (Static Random Access Memory) can be candidates. Flash memory offers a large capacity, but its operation speed is limited by long read times. SRAM can operate at high speeds but has limited capacity due to low cell density. DRAM, with its significant gigabit capacity and high bandwidth, can be considered as a suitable candidate for PIM.

In previous studies on DRAM-based Process-in-Memory (PIM), there were suggestions for bank-level architecture modifications [1], [2] and DRAM cell structure modifications [3]. [1] and [2] effectively demonstrated DRAM-based computation by introducing alterations to the DRAM architecture at the bank level and integrating processing elements within the same level. Nevertheless, the constraint of activating only one word-line at a time results in a limitation in computational bandwidth. Moreover, the arrangement of processing elements at the bank level proves infeasible for networks such as Convolutional Neural Network (CNN) that require extensive data reuse. This is attributed to the challenges posed by latency and energy considerations. [3] modifies DRAM cell by adding transistors or a capacitor to the 1T1C structure and demonstrated cell-level PIM in embedded DRAM (eDRAM). However, modifying cell structure in DRAM process is difficult to be implemented, which limits the feasibility of this works.

In this work, PIM in DRAM with high throughput using multiple word-line access without modifying the cell structure is proposed. To address the data destruction issue due to charge sharing among multiple cells while multiple word-lines are activated, data recovery is employed using the adjacent MAT. Also, to remove the computation errors caused by switch non-ideality, differential operation between two adjacent bit-lines is proposed. Note that the proposed architecture can be successfully implemented in DRAM process since the architecture is designed to fit the DRAM architecture while not modifying DRAM cell structure.

TABLE I. Comparison of Memory Device Characteristics

Memory	Cell size	Capacity	Internal Bandwidth
SRAM	160 F^2	100 Mb	1 Tbps
DRAM	6 F^2	16 Gb	100 Gbps
NAND Flash	< 1 F^2	1 Tb	100 Mbps

II. PROPOSED DRAM-PIM

A. Basic Concept of the MAC operation

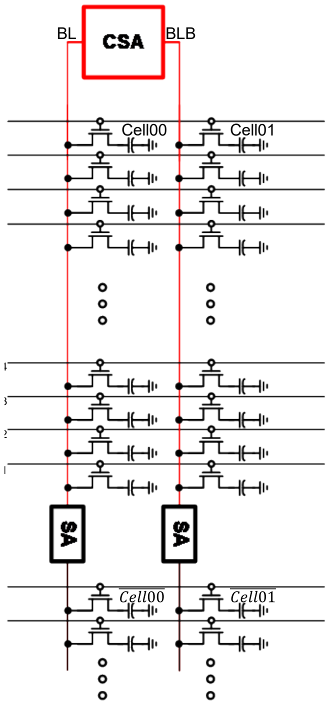
The operation of MAC can be divided into two main processes;

1. multiplication of input(IN) and weight(W)
2. accumulation of multiplications

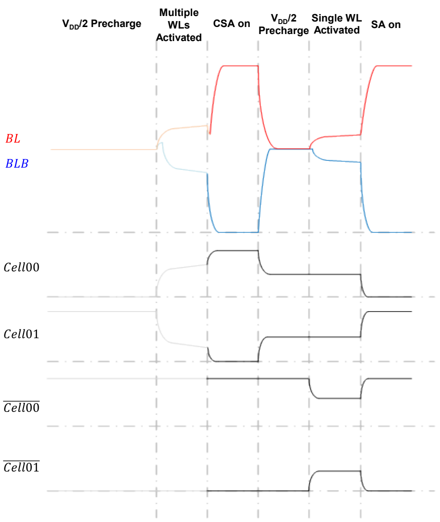
In this paper, we propose DRAM PIM structure plotted in Fig. 1, which performs MAC operation exploiting 1T1C DRAM cells. The basic operations of the proposed structure

a. Corresponding author; chosta@kaist.ac.kr

Manuscript Received Aug. 21, 2023, Revised Nov. 26, 2023, Accepted Dec. 5, 2023



(a)

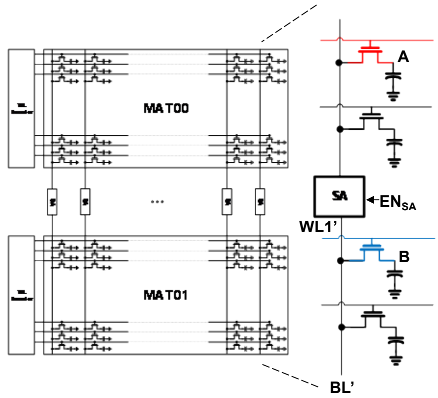


(b)

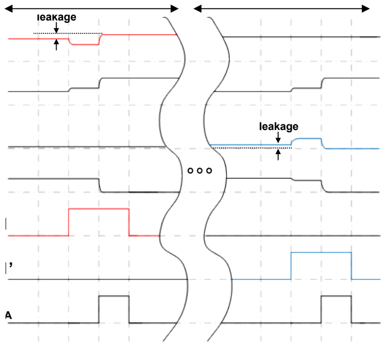
Fig. 1. MAC Operation of Proposed DRAM PIM; (a) Circuit Configuration (b) Timing Diagram for MAC Operation

are highly similar to that of DRAM. In this structure, the stored data in DRAM cell can be charge-shared with Bit-Line (BL) capacitor only for the case that WL signal is high. In case of WL voltage is low, the pass transistor doesn't transfer the stored data to BL, resulting no changes in V_{BL}. These operation matches to the multiplication, where the WL signal and the stored data match to input and weight, respectively. The voltage after charge-sharing V_{BL} is represented,

$$V_{BL} = \frac{C_p * \frac{V_{DD}}{2} + C_C * N_{INPUT=1\&WEIGHT=1} * V_{DD}}{C_p + C_c * N_{INPUT=1}} = \frac{1}{2} * V_{DD} + \Delta V$$

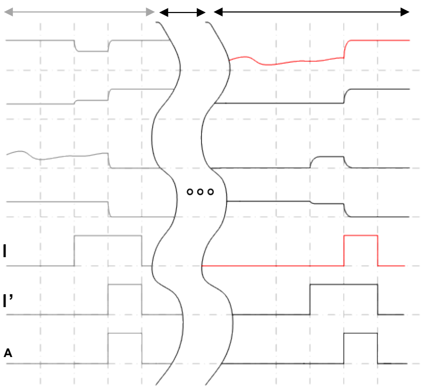


(a)

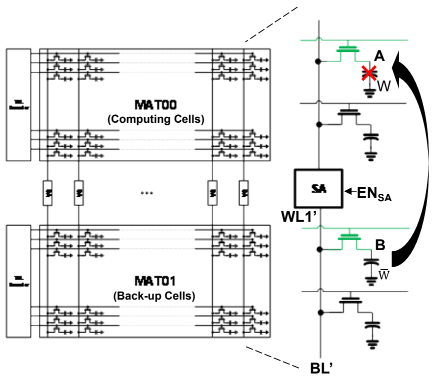


(b)

Fig. 2. Conventional DRAM; (a) Circuit configuration (b) Timing diagram for the refresh and read operations



(a)



(b)

Fig. 3. Proposed DRAM IMC Structure; (a) Circuit Configuration (b) Timing diagram for the MAC and recover operations

where C_p is BL parasitic capacitance, C_C is cell capacitance, V_{DD} is supply voltage and N is the number of arbitrary targets. V_{BL} is compared with V_{ref}, resulting either 1 or 0 depending on ΔV. The proposed structure can perform multiple WL activation, which enables the accumulation of the multiple multiplication results. In this structure, it has an advantage of fast operation time thanks to multiple WL access without any modifications of DRAM cell structure, enabling application of the proposed structure directly to commercial DRAM.

B. Data Recovery Using Adjacent MAT

DRAM structured with a 1T1C configuration possesses a characteristic where refresh operations are essential due to the presence of leakage current in the cell. As shown in Fig. 2(a), if cell A connected to BL holds the data '1', and cell B connected to BL' holds the data '0', over time, as depicted in Fig. 2(b), each cell tends to deviate from its originally stored data due to leakage current. Therefore, periodic activation of WL is necessary to execute refresh operations. In the case of conventional DRAM, when refreshing a cell located at MAT00, the corresponding WL of MAT00 is activated, while the WL situated in MAT01, which shares the same sense amplifier with MAT00, remains inactive.

In conventional DRAM, due to the limitation of activating only one WL at a time, the throughput of MAC operations is inevitably lower. In the proposed DRAM IMC structure, as shown in Fig. 3(a), the objective is to overcome the drawback of low MAC operation throughput while maintaining a physically identical configuration to traditional DRAM. This is achieved by multiple WL activations to enhance the throughput of MAC operations. When multiple WL activations are applied, a critical drawback arises due to charge sharing among multiple cells situated within the same BL. This leads to the destruction of stored data. In the proposed structure, a solution has been devised by utilizing the differential operation between MAT00, where the operation is conducted, and the neighboring MAT01 along with the sense amplifier (SA).

Looking at Fig. 3(b), the proposed DRAM IMC structure employs a total of three stages: Copy - MAC - Recover, to carry out MAC operations and data recovery. In the initial 'Copy' operation, the WLs of both the targeted MAT00 for the operation and the neighboring MAT01 are each activated. This action copies the data stored in MAT00 to MAT01. During this process, due to the differential behavior of the sense amplifier, the data copied into the cells of MAT01 becomes the complementary value of the original data stored in MAT00. Subsequently, in the 'MAC' operation, multiple WLs are activated in MAT00 to perform the MAC operation, leading to the destruction of the data initially stored in MAT00. Finally, during the 'Recovery' operation, the same as the first 'Copy' operation, the WLs of MAT00 and MAT01 are each activated, allowing the data stored in MAT01 to be returned to MAT00. During this operation, since the data stored in MAT01 is the complementary value of the original data in MAT00, the updated data in MAT00 after recovery is identical to the original data stored.

The above operation is similar to the refresh or read operations in conventional DRAM, except for the simultaneous activation of WLs of the target MAT and the adjacent MAT. Therefore, the proposed structure achieves high throughput with minimal overhead and enables both computation and data recovery to be carried out efficiently.

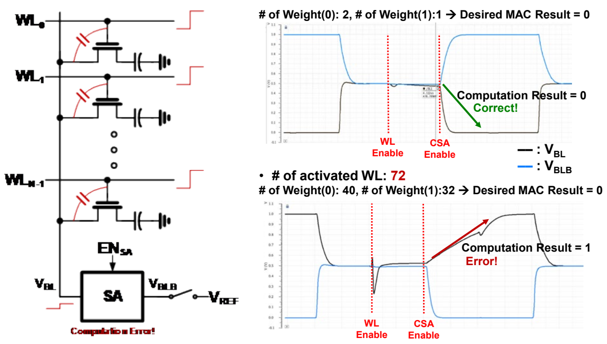


Fig. 4. Presence of Switching Non-ideality

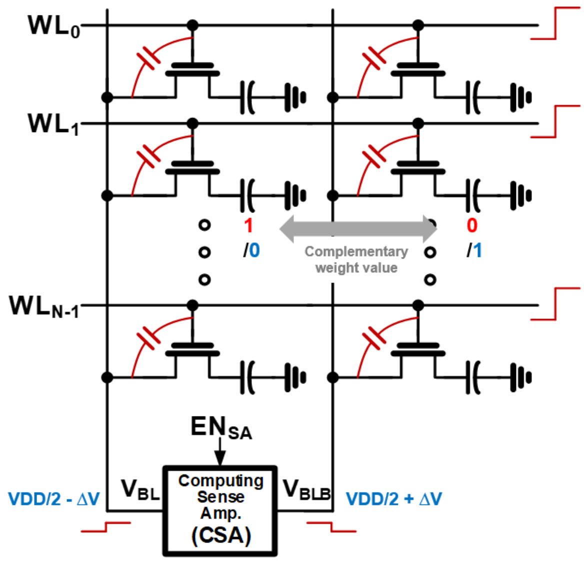


Fig. 5. Proposed Differential Bit-Line Computation

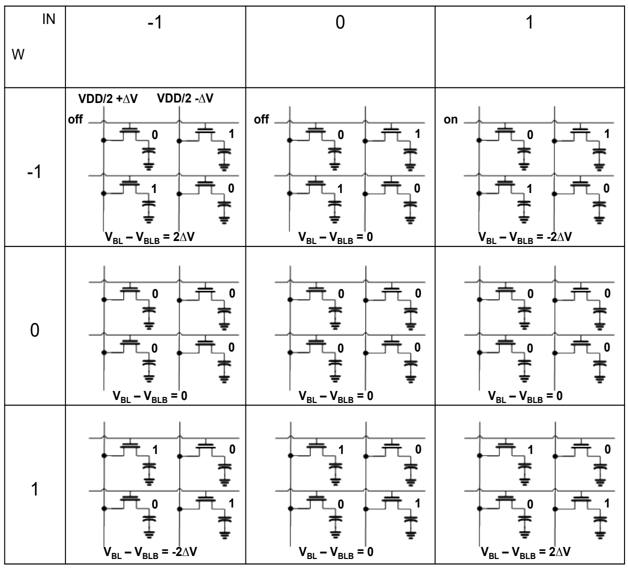


Fig. 6. Ternary Input/Weight Operation

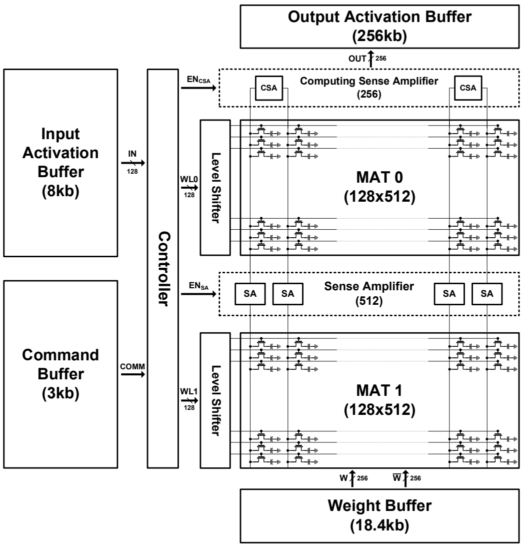


Fig. 7. Overall Architecture

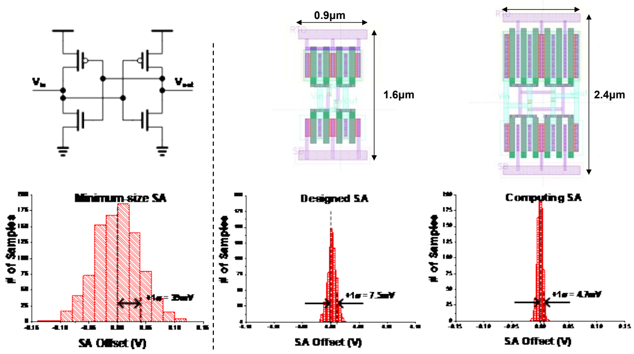


Fig. 8. Offset voltage result of the MonteCarlo simulation of the core sense amplifier and computing sense amplifier

C. Differential Bit-Line Computation

During the proposed MAC operation, there exists the non-ideality of switch transistor such as clock feedthrough and charge injection. The impact of the switching non-ideality can be more severe as the number of activated WL increases, which is depicted in Fig. 4. The non-ideality impact can cause the computation error, resulting the degradation of the computing accuracy. Therefore, in this paper, we propose the differential MAC operation, which can suppress the effect of the switching non-ideality using 2 adjacent BLs, as plotted in Fig. 5. In differential MAC operation, the presence of switching non-ideality in both adjacent BLs can lead to a cancellation of switching effects, thereby reducing the switching effects. The adjacent BLs have complementary weight values, thus doubling the sensing margin. Because 2x2 cell array containing 2WLs and 2BLs is a minimum calculation unit, each weight and input can have ternary states. All cases for each ternary input/weight are described in Fig. 6. Therefore, the proposed differential bit-line computation with ternary input/weight mode enables cancelling the switching non-ideality as well as increasing the throughput exploiting the ternary states.

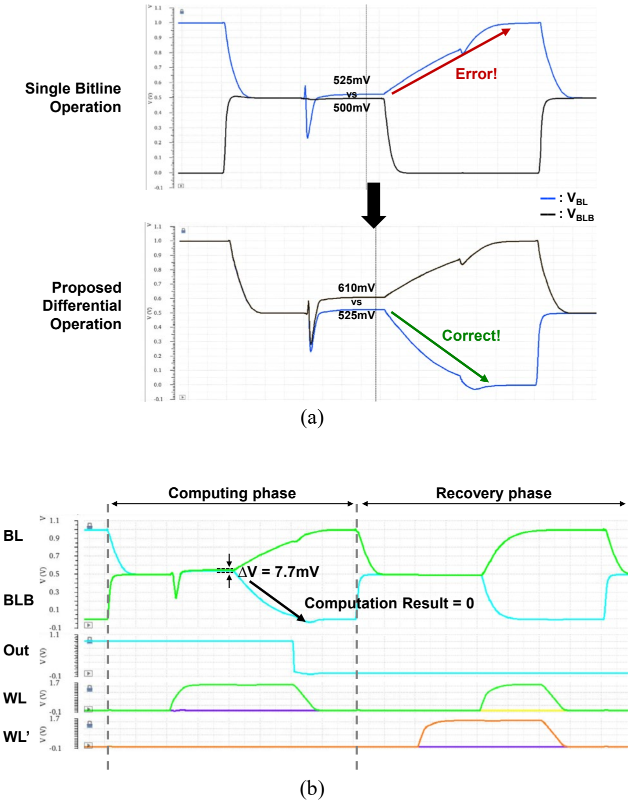


Fig. 9. Simulation Result; (a) Comparison of the simulation result between single bit-line operation and differential bit-line operation (b) Transient Result including both computing and recovery phase

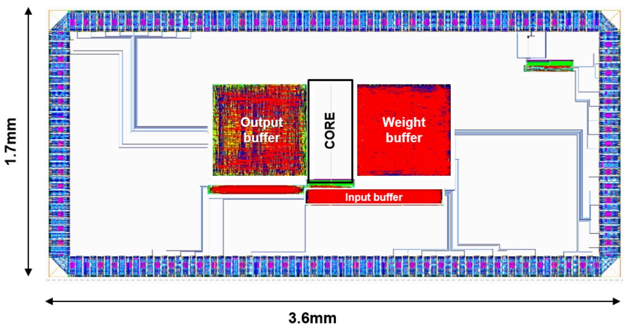
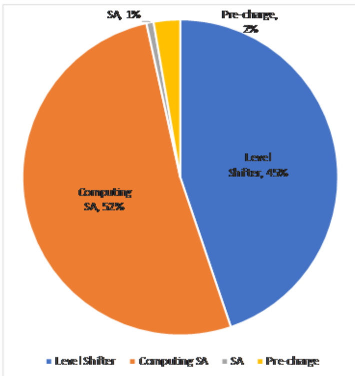


Fig. 10. Chip Photo of The Proposed DRAM PIM system



Level Shifter	0.909mW
Computing SA	1.050mW
SA	0.015mW
Pre-charge	0.055mW
Total	**2.021mW**

Fig. 11. Power breakdown of the core system

TABLE II. Comparison Table

	ISSCC'21	TCAS1'21	ASSCC'21	This work
Tech	65nm	65nm	28nm	28nm
Bitcell	1T1C eDRAM	4T2C eDRAM	8T2C SRAM	1T1C DRAM (Logic)
VDD	1-1.2	0.5-0.7	0.7-1.0	1.0-1.6
Precision (bit)	Input 8/weight 8 /output 8	Input 1/weight 1.5	Input 1/weight 1.5	Input 1-1.5 /weight 1-1.5
On-chip Mem. Size	16kb	16kb	16kb	128kb
Model	CNN	CNN	CNN	CNN
Dataset	Cifar-10	Cifar-10	MNIST/Cifar-10	MNIST/Cifar-10
Accuracy	80.1	82.8	97.37/81.17	99.01/77.63
Throughput (GOPS)	4.71	N/A	716	2048(Binary) /1536 (Ternary)
Energy Efficiency (TOPS/W)	4.76	552	1607	1280(Binary) /757 (Ternary)

III. SIMULATION RESULTS AND DISCUSSIONS

The overall architecture of the proposed DRAM PIM is shown in Fig. 7. The entire system operates at a frequency of 250MHz. The input buffer and weight buffer store the data for MAC operation. The input and weight data are given by the corresponding buffer and are multiplied and accumulated. Finally, the result for MAC operation is transferred to computing sense amplifier where the binary decision is done for BNN. The output activation buffer stores the final output and transfers the data out of the chip. The main component of the proposed system is a sense amplifier. The offset of sense amplifier can be a severe problem because of its mismatch characteristics. In our work, the standard deviation of the sense amplifier is less than 7.5mV by sizing the transistor. The distribution and layout are depicted in Fig. 8. In Fig. 9(a), by exploiting the proposed differential bit-line computation, the calculation can be performed correctly. In detail, the presence of non-ideality of transistor switch leads to 110mV differential output difference compared to that of single-ended bit-line computation. The detailed operation of differential bit-line computation is illustrated in Fig. 9(b). As can be seen in the figure, the result was correctly decided even though there is only 7.7mV difference between BL and BLB. Followed by computing phase, the data is recovered by the complementary data as mentioned in II-B chapter. The chip is fabricated in 28nm technology. The area of the chip is $6.12mm^2$, as can be shown in Fig. 10. The system consumes 2.021mW while the level shifter, computing SA, SA and pre-charging consume 0.909mW, 1.05mW, 0.015mW and 0.055mW, as can be seen in Fig. 11. It should be noticed that the computing SA consumes 51.9% of the total system power. As a result, the proposed DRAM-PIM shows the best throughput of 2048(binary)/1536(ternary) GOPS. This is calculated under the assumption that the utilization of MAT is 100%. In binary, it's 128 (number of Word Lines) x 512/2 (number of differential Bit Lines) x 250MHz/4 (4 clocks used for computation) = 2048GOPS. In ternary, it's 128/2 (two Word Lines per input) x 512/2 x 250MHz/4 x 1.5 (ternary) = 1536GOPS. Also, the proposed DRAM-PIM results in accuracy of 99.01% against MNIST dataset, which is a best performance among the compared paper listed in the comparison table. It should be noted that the proposed DRAM-PIM exploits the existing DRAM system without any modifications of DRAM cell structure(1T1C).

IV. CONCLUSION

The proposed DRAM-PIM performs MAC operations effectively exploiting existing DRAM structure with multiple WL activation. The proposed system achieves an accuracy of 99.01% against MNIST dataset, while the throughput of the system is 2048 GOPS. This research is meaningful that the proposed structure can be applied to existing DRAM without any modification of DRAM cell structure(1T1C).

ACKNOWLEDGMENT

The chip fabrication and EDA tool were supported by the IC Design Education Center (IDEC), Korea. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-00826, High resolution intelligent Radcomm system)

REFERENCES

[1] S. Lee et al., "A 1ynm 1.25V 8Gb, 16Gb/s/pin GDDR6-based Accelerator-in-Memory supporting 1TFLOPS MAC Operation and Various Activation Functions for Deep-Learning Applications," 2022 IEEE International Solid- State Circuits Conference (ISSCC), 2022, pp. 1-3

[2] Y. -C. Kwon et al., "25.4 A 20nm 6GB Function-In-Memory DRAM, Based on HBM2 with a 1.2TFLOPS Programmable Computing Unit Using Bank-Level Parallelism, for Machine Learning Applications," 2021 IEEE International Solid- State Circuits Conference (ISSCC), 2021, pp. 350-352

[3] S. Xie, C. Ni, A. Sayal, P. Jain, F. Hamzaoglu and J. P. Kulkarni, "16.2 eDRAM-CIM: Compute-In-Memory Design with Reconfigurable Embedded-Dynamic-Memory Array Realizing Adaptive Data Converters and Charge-Domain Computing," 2021 IEEE International Solid- State Circuits Conference (ISSCC), 2021, pp. 248-250

[4] R. W. Erickson and D. Maksimovic, *Fundamentals of Power Electronics*, Springer, 2001.

[5] M. He et al., "Newton: A DRAM-maker's accelerator-in-memory (AiM) architecture for machine learning." 2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO). IEEE, 2020.

[6] Y. Kim et al., "Input-splitting of large neural networks for power-efficient accelerator with resistive crossbar memory array." Proceedings of the International Symposium on Low Power Electronics and Design. 2018.

[7] J. Kim et al., "Area-efficient and variation-tolerant in-memory BNN computing using 6T SRAM array." 2019 Symposium on VLSI Circuits. IEEE, 2019.

Hyein Yoon received the B.S. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2020, and M.S degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2022. Her research interest includes DRAM processor in memory for machine learning. Especially, she is currently conducting the research on smart scheduler in NAND flash memory in Samsung Electronics.

Donghwan Kim received the B.S. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2019. He is currently pursuing the integrated master's and doctoral degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. His research interest includes memory IC and PIM for area efficient application.

Giwoo Lee received the B.S. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2021. He is currently pursuing the integrated master's and doctoral degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. His research interest includes memory IC and PIM for low power application.

SeongHwan Cho received the B.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1995, and the M.S. and Ph.D. degrees in EECS from MIT, Cambridge, MA, USA, in 1997 and 2002, respectively.

In 2002, he joined Engim, Inc. Acton, MA, USA, where he was involved in data converters and phased-locked loop (PLL) design for IEEE 802.11abg WLANs. Since 2004, he has been with the School of EE, KAIST, where he is currently a Professor and the Department Head of semiconductor system engineering. He was with Marvell Inc., Santa Clara, CA, USA, from 2011 to 2012, and Google, London, U.K., from 2016 to 2017, as the Research Scientist. His research interests include analog and mixed-signal circuits for high-speed communication, low-power sensors, memory, and machine learning.

Prof. Cho was a co-recipient of the 2009 IEEE Circuits and System Society Guillemin-Cauer Best Paper Award and the 2012 ISSCC Takuo Sugano Award for Outstanding Far-East Paper. He has twice received Outstanding Lecturer Award from KAIST. He has served on the Technical Program Committee on several IEEE conferences, including ISSCC, Symposium on VLSI and A-SSCC. He has served as an Associate Editor for IEEE JOURNAL OF SOLID-STATE CIRCUITS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS, and Distinguished Lecturer of the IEEE Solid-State Circuits Society.