UPIM : Unipolar Switching Logic for High Density Processing-in-Memory Applications

Joonseop Sim, Saransh Gupta, Mohsen Imani, Yeseong Kim and Tajana Rosing
University of California San Diego, La Jolla, CA 92093, USA
{j7sim, sgupta, moimani, yek048, tajana}@ucsd.edu

ABSTRACT

Internet of Things (IoT) has built a network with billions of connected devices which generate massive volumes of data. Processing large data on existing systems requires significant costs for data movements between processors and memory due to limited cache capacity and memory bandwidth. Processing-In-Memory (PIM) is a promising solution to address the issue. Prior techniques that enable the computation in non-volatile memory (NVM) are designed on a bipolar switching mode, which suffers from a high sneak current in a crossbar array (CBA) structure. In this paper, we propose a unipolar-switching logic for high-density PIM applications, called UPIM. Our design exploits a unipolar-switching mode of memristor devices which can be operated in 1D1R structure, hence suppresses the sneak current that exists in prior PIM technologies. Moreover, UPIM takes advantages of a 3D vertical crossbar array (CBA) structure to increase memory utilization per unit area for high-density applications. Our evaluation on a wide range of applications shows that the UPIM achieves up to 31.3× energy saving and 113.8× energy-delay product (EDP) improvement as compared to a recent GPGPU architecture. As compared to the state-of-the-art PIM design based on the bipolar switching mode, our design achieves 3.1× lower energy consumption.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**; *Supervised learning*;

KEYWORDS

Processing in-memory, Non-volatile memory

ACM Reference Format:
Joonseop Sim, Saransh Gupta, Mohsen Imani, Yeseong Kim and Tajana Rosing. 2019. UPIM : Unipolar Switching Logic for High Density Processing-in-Memory Applications. In *Proceedings of ACM Conference (GLSVLSI'19).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 DESIGN OVERVIEW

1.1 Memristor Switching Modes

There are two classes of ReRAM switching mode depending on the applied bias polarity. One is 'unipolar', where the switching between high resistance state (HRS) and low resistance state (LRS) is not relevant to the polarity of the operating voltage as shown in Fig. 1(a), and the other is 'bipolar', where the reset switching (LRS → HRS) and set switching (HRS → LRS) take place with the opposite of the bias polarity as shown in Fig. 1(b)

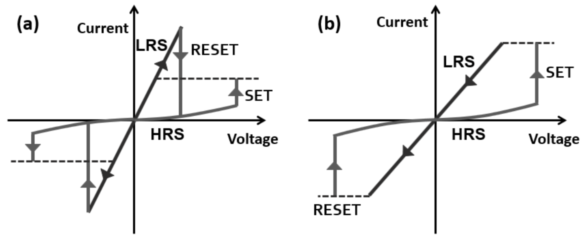
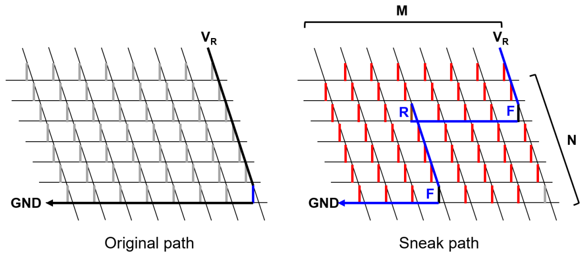


Figure 1: (a) Unipolar and (b) Bipolar switching mechanisms of memristor



Original path

Sneak path

Figure 2: Sneak current influence on the total current in a memory array

1.2 Unipolar-based logic within NVM

Fig. 3(a) shows the basic structure of the proposed UPIM. To simplify the explanation, we show a logic that supports two-input NOR operation, but it can be extended to multi-input logics in a straightforward way. Each unipolar device consists of a memristor device and a diode. The input values are stored in two memristors, R_{IN1} and R_{IN2}, while the other memresistor, R_{OUT} stores the computation result. The logical values are stored in each memresistor as resistance states in the input/output memristors. HRS in either R_{IN12} or R_{OUT} indicates the logical value of 0, while LRS represents 1. In our experiment, we exploit the model shown in

To perform the NOR operation, our design first initializes the R_{OUT} to R_{HRS}. We then apply the V_{IN1} and V_{IN2} voltages to the input memristors and V_{OUT} to the output memristor. Fig. 3(c) shows how the proposed logic performs the NOR operation. In the two-input case, the stored values in the input memristors have four combinations: 00, 01, 10 and 11. When both inputs have high resistance, i.e., '00', the voltage on the BL (V_{BL}) is almost pulled into ground, while the voltage across R_{OUT} (V_{OUT}-V_{BL}) is close to V_{OUT}. Since V_{OUT}-V_{BL} is larger than V_{SET}, it incurs the SET switching of the R_{OUT} to LRS. Note that the applied voltage across the diode is negligible as compared to the voltage applied to R_{OUT} since R_{OUT} is previously initialized as HRS. In all other cases (i.e., 01, 10, and 11), at least one

GLSVLSI'19, May 2019, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

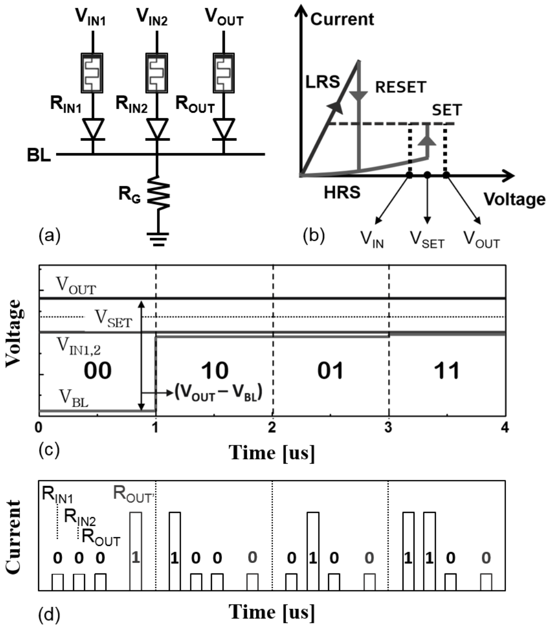


Figure 3: Proposed unipolar–based NOR logic: (a) Schematic of the NOR gate (b) Voltage conditions (c) NOR gate simulation result (d) Resistance behaviors depending on input states

of the input memristors has a low resistance state. Therefore, the V_{BL} voltage has a higher voltage close to V_{IN}. For instance, if the case of '10', where R_{IN1} and R_{IN2} have LRS and HRS, respectively, the net resistance is close to R_{IN1}. Since the voltage ratio of R_{IN1} to R_G is close to zero, ($\approx$ 0.03 in our experiment), V_{BL} is almost V_{IN}. Thus, the R_{OUT} keeps the high resistance state representing the logical 0. Fig. 3(d) shows the resistance behavior of the UPIM NOR gate. R_{OUT} and $R_{OUT'}$ indicate resistance states from the output resistor prior to operation and after applying V_{IN}, respectively. Except for the case of '00' which the SET switching occurs in R_G, all the other cases keep the R_G as low resistance state, presenting NOR operation.

1.3 Integration to 3D CBA structure

The proposed design executes arithmetic functions using NOR operations. Existing NOR-based approaches require additional cells to store intermediate results. The area overhead due to the generated intermediate states is not suitable for high-density applications. In this work, we utilize a 3D structure to minimize the area cost. Fig. 4(a) shows the conventional 2D logic implemented in a memory array. In this structure, the intermediate operation results are stored in the same plane while consuming extra cell area. In contrast, as shown in Fig. 4(b), the 3D structure can store the intermediate results in a different layer. Therefore, the intermediate cell is hidden under/over the memory cells, increasing chip density as compared to the 2D case.

Fig. 5 presents the comparison diagram of 2D and 3D cases. We denote the area of memory cells, which is used to store data, by A_{memory}. A_{logic} and A_{shift} are the areas of intermediate cells for storing logic results and the interconnects, respectively. We define *cell efficiency* as the ratio of the memory area over the total area. In the 2D design, since the intermediate cells take chip area, the cell efficiency is represented by $A_{memory}A_{memory}+A_{logic}+A_{shift}$. In contrast, for the 3D case, the intermediate cells for all arithmetic logic can be completely stacked on the top of the memory cells. If the number of layers is n, the cell efficiency of 3D design is given by $n \times A_{memory}A_{memory}+A_{shift}$. This means that, with the 3D logic stacking, it can achieve high area efficiency.

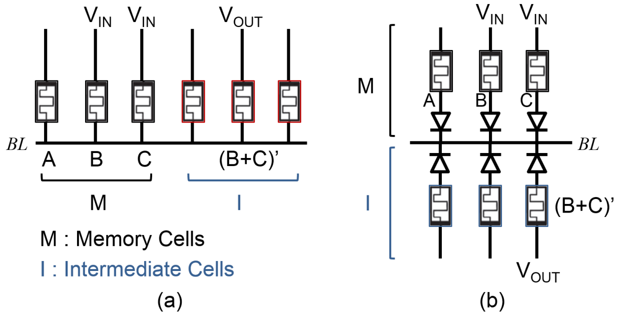


Figure 4: Schematic of (a) Prior 2D and (b) Proposed 3D logic in memory

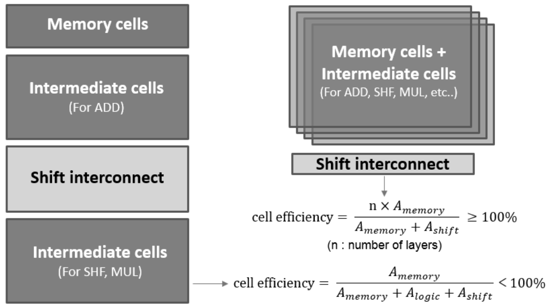

Memory cells
Intermediate cells
(For ADD)
Shift interconnect
Intermediate cells
(For SHF, MUL)
Memory cells +
Intermediate cells
(For ADD, SHF, MUL, etc..)
Shift interconnect
cell efficiency = $\frac{n \times A_{memory}}{A_{memory}+A_{shift}} \geq 100\%$
(n : number of layers)
cell efficiency = $\frac{A_{memory}}{A_{memory}+A_{logic}+A_{shift}} < 100\%$

Figure 5: Diagram of prior 2D (left) and proposed 3D (right) PIM structures

Fig. 6 shows our integration design of 3D logic-in-memory. The V_{IN} and V_{OUT} are applied to wordlines connected to memory cells and intermediate cells, respectively. For example, if the V_{IN} is applied to 'A' and 'B' cell, the result of NOR operation is stored at a cell where the V_{OUT} is applied. As appeared in the figure, the proposed 3D structure can improve the chip density by storing the intermediate results in a different layer compared to the 2D structure. Moreover, a memory layer and a computation layer are paired and they can be stacked with multiple layers. Therefore, our design enables parallel operation with a single input signal. In case of Fig. 6, the UPIM NOR operations of A and B, D and E can be executed in parallel with a single PIM operation.

Table. 1 summarizes the comparison of the proposed UPIM to existing technologies. Since UPIM performs logic operations in 1D1R cell structure, we achieve higher power efficiency than other PIM technologies based on the bipolar switching mode. Moreover, when implementing UPIM into the 3D CBA structure, it further overcomes the issue of the area overhead existing in the 2D PIM approaches.

2 EXPERIMENTAL RESULTS

2.1 Experimental Setup

Performance and energy consumption have been obtained by Cadence Virtuoso and Spectre simulators with 45nm CMOS process technology. We use *VTEAM* memristor model

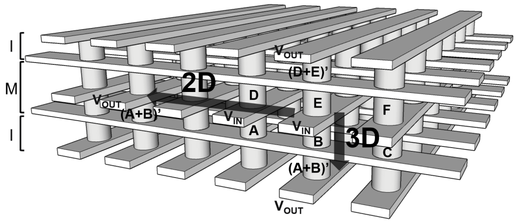


Figure 6: The integrated structure of 3D UPIM

Table 1: performance of proposed UPIM and other technologies

	IMPLY	MAGIC	2D-UPIM	3D-UPIM
Cell Structure	1R	1R	1D1R	1D1R
Condition	$2(V_{cond}, V_{set})$	$1(V_0)$	$2(V_{in}, V_{out})$	$2(V_{in}, V_{out})$
Functions	IMPLY (False)	OR, NOR, NOT, AND, NAND		
Power	High leakage	High leakage	**Low**	**Low**
Density	Low	Low	Low	**High**

2.2 Energy and Performance

As discussed in Section 1.1 and 1.2, our unipolar-based logic is operated in the 1D1R structure, which shows lower static power consumption by reducing sneak current dissipation. Fig. 7 shows the energy and energy-delay product (EDP) improvements of running applications on proposed UPIM and state-of-the-art PIM designs

2.3 Process Variation

The UPIM design uses a configurable resistor, R_G. To make our design robust, we determine the resistor value with consideration of process variation, which most of today's technology suffers. In our experiment, there are two major factors that induce process variation, memristor dimension, and near-far cell difference. The dimension variation comes from a diameter deviation during lithograph and etching process in the formation of pillar memristors. This results in the resistance variation on UPIM, since a memristor resistance with a cylindrical shape has an inverse dependency with its diameter

Fig. 8(b) shows V_{BL} characteristic as a function of R_G, when input values are 00 and 01, considering the factors of the process variation. All V_{BL} transfer curves are presented with dimension variation of 10%, denoted as (H). As R_G increases, the electrical potential in the BL increases due to an escalation of the voltage applied to R_G. $V_{OUT} - V_{BL}$ has to be higher than V_{SET} for the case of 00 and lower than V_{SET} for other cases, i.e., 10,01,11. Thus, the gap between $V_{OUT} - V_{BL}$@10 and $V_{OUT} - V_{BL}$@00 needs to be enough wide for operation stability. The voltage gap, denoted as V_{BL} margin, is tunable by adjusting R_G value. Fig. 8(c) shows the simulation results of the V_{BL} margin for different R_G. We extract an optimized R_G point from the graph of V_{BL} margin with a R_G. Based on this analysis, we choose the optimal R_G value, $R_{G,OPT}$, by 300KΩ to guarantee computation accuracy, despite existing process instability.

2.4 Evaluation for Area Efficiency

We evaluated area efficiency of our design as compared to the MAGIC

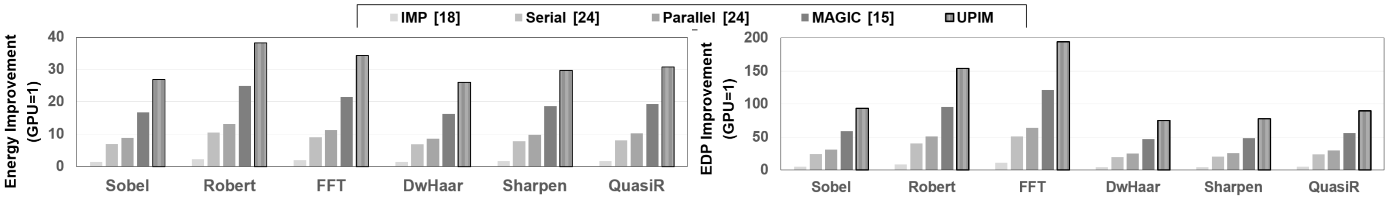
3 CONCLUSION

We present an energy efficient and high-density PIM architecture which enables logic-in-memory based on unipolar-switching memristors. The proposed design resolves the static power issue due to the sneak current by implementing the logic in the 1D1R cell structure. Our design also addresses the low cell-density of other PIM technologies due to extra area consumption for storing computation results by implementing them in 3D CBA. The experimental results show that our design presents 3.1× and 31.3× improvement in energy consumption compared to the state-of-the-art PIM designs and the GPU architecture, respectively.

ACKNOWLEDGEMENTS

This work was partially supported by CRISP, one of six centers in JUMP, an SRC program sponsored by DARPA, and also NSF grants #1730158 and #1527034. Joonseop Sim is partially supported by a Ph.D. fellowship from SK Hynix Inc.

REFERENCES



(a) Energy

(b) Energy-Delay Product

Figure 7: Energy and Energy-delay product improvement of proposed UPIM and state-of-the-art

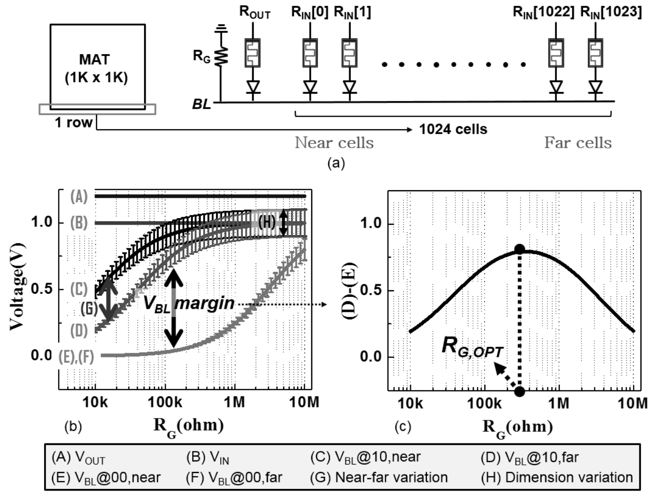


Figure 8: V_{BL} margin and R_G optimization considering process variation

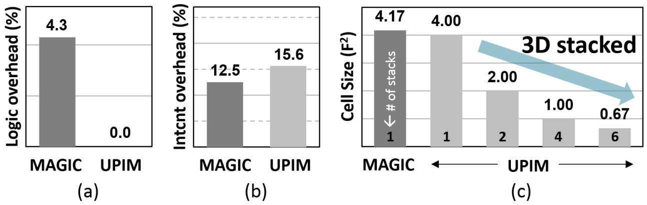


Figure 9: Overhead and cell size comparison between UPIM and MAGIC