A 65nm Time-Domain Computing-in-Memory 6T-SRAM Macro for 8b-MAC Operations for Edge-AI Devices

應用於 AI 設備之65奈米時域記憶體內8b乘加運算的六電晶體靜態隨機存取記憶體架構

Group Number: A405 Advisor: Chang, Meng-Fan Member: Yeh, Yao Kai

Introduction

Background

- Convolutional Neural Network (CNN) is powerful for image sensing and voice recognition application.
- Multiply and Accumulation (MAC) operations is repeatedly performed in CNN. It's of interest to reduce its energy consumption.
- Computing-In-Memory (CIM) is an approach to perform MAC in memory unit, which can save data transmitting energy.
- **Decreasing Signal Margin Problem:** Signal margin of the voltage output decreases exponentially as increasing accumulation number under same power supply. It's a critical problem for the charge sharing [1] and the voltage dividing [2] approaches.

Time-Domain CIM Macro [3] **MAC Output = (8bIN×1bw) × 64 accumulations**

In this project, I implement a time-domain CIM macro with **TSMC 65nm process technology**. It solves decreasing signal margin issue by counting the number of time delay accumulation of the output. However, it has non-uniform delay time caused by process-voltage-temperature (PVT) variation and leakage problem.

Design

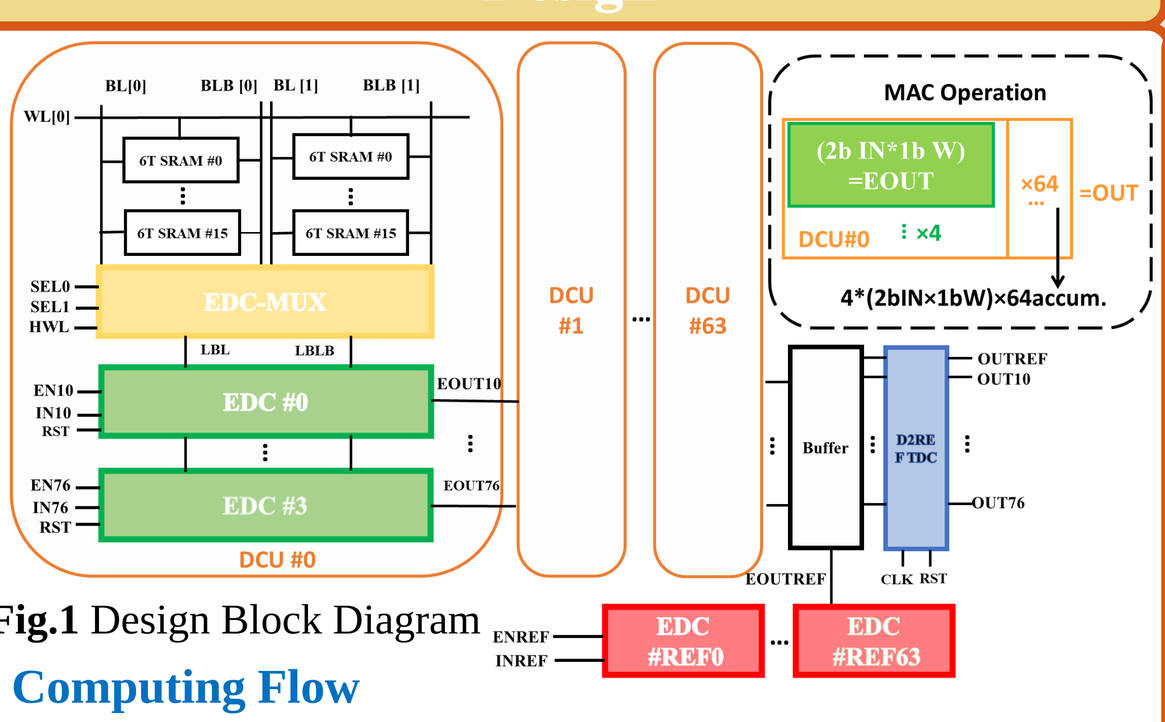

Fig.1 Design Block Diagram

Computing Flow

1. Weights stored in SRAM array are selected by edge delay cell MUX (EDC-MUX) and word line (WL).
2. EDC-MUX passes weight to edge delay cell (EDC) through local bit lines (LBL\LBLB) and multiplies it with input.
3. EN enables first delay computing unit (DCU), and the output enables second DCU, and so on. →Accumulation
4. Dynamic Differential Reference Time to Digital Converter (D2REF-TDC) counts output delay and reference delay, then generates digital output.

New Design: RC-Modulation (RCM) EDC

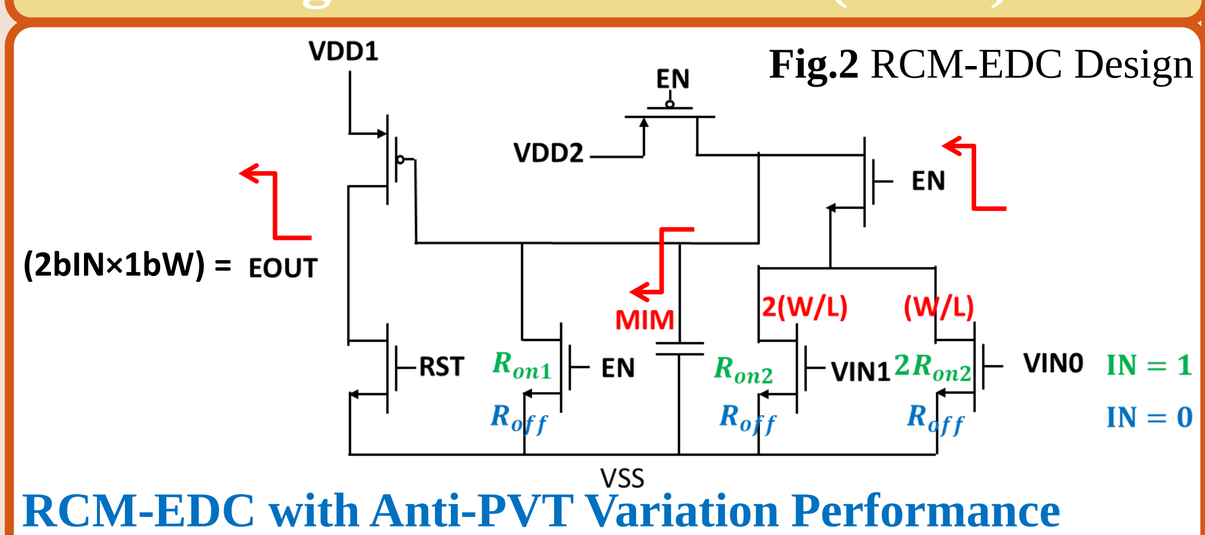

Fig.2 RCM-EDC Design

RCM-EDC with Anti-PVT Variation Performance

- Represent 2b place value with different size.
- Modulates RC time constant with different discharge path.
- Digital input for higher operating voltage and stability.
- Add metal-insulator-metal (MIM) capacitor to dominate node capacitance.

Operation Stages

1. Reset: VDD1 = EN = 0 and RST = 1➔ EOUT = 0

2. Multiply: feed VIN0 and VIN1 with V_{DD} or V_{SS} respectively, corresponding to 2b In × 1b W value.

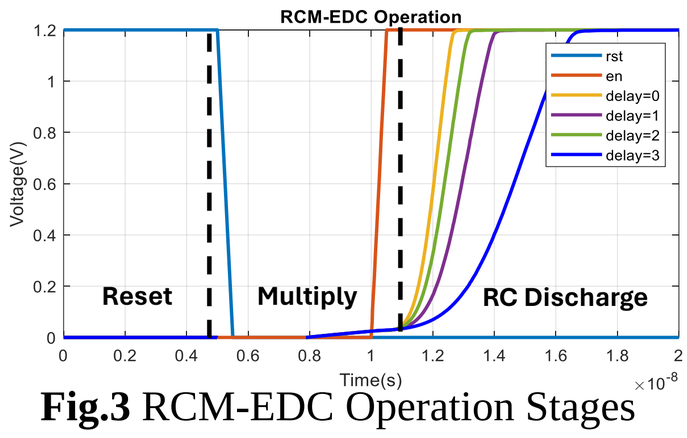

Fig.3 RCM-EDC Operation Stages

3. RC Discharge: capacitor discharge voltage with R_{on1} parallel to different number of R_{on2} and R_{off} corresponding to input value.

Measurement [3]

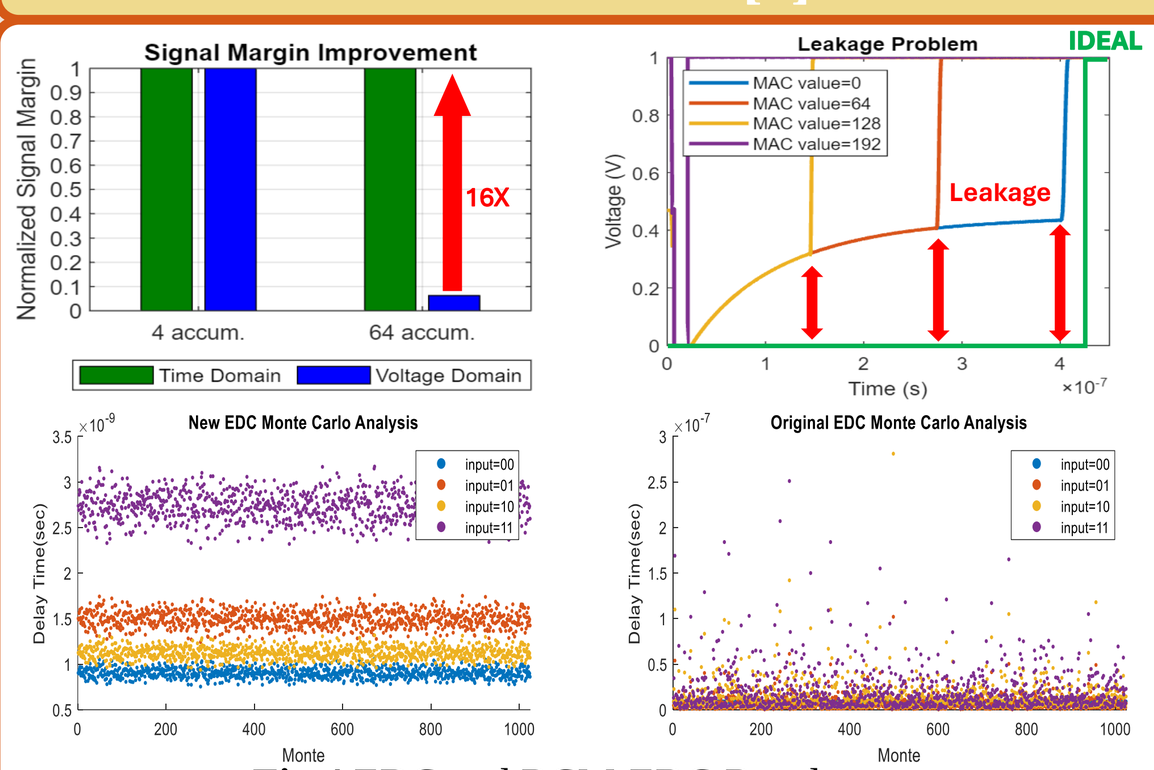

Fig.4 EDC and RCM-EDC Results

EDC

Pros

- Signal margin keeps identical with different number of accumulations. There is no precision loss.

Cons

- EDC suffers from PVT variation severely because transistors operate in small operating voltage with analog input voltages.
- There is leakage problem when EDC is waiting for enable signal. It makes each EDC output rises from different starting point and has inaccurate delay time.

RCM-EDC

- Computes with shorter delay time.
- Smaller PVT variation influence.

Conclusion

- Time-domain design solves decreasing signal margin issue, EDC has leakage problem and PVT variation influence.
- Time-domain approach is suitable for result-concentrate and high-precision data set.

New Design: RCM-EDC

- shorter delay time
- smaller PVT variation influence

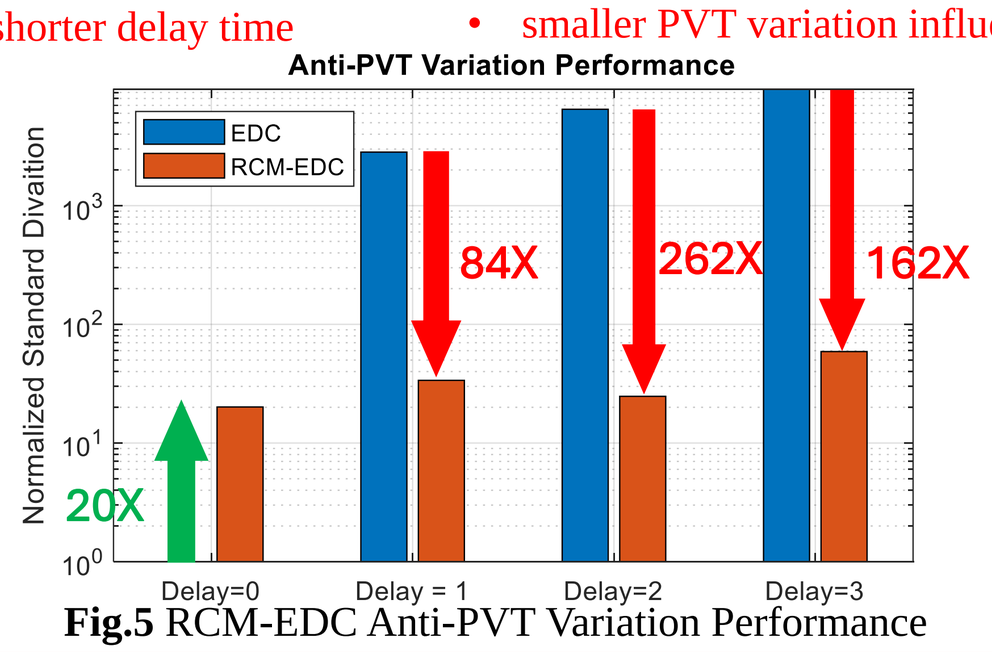

Fig.5 RCM-EDC Anti-PVT Variation Performance

[1] A. Biswas and A. P. Chandrakasan, "Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications," in 2018 IEEE International Solid-State Circuits Conference - (ISSCC), 2018, pp. 488–490. DOI:10.1109/ISSCC.2018.8310397.
[2] X. Si, Y.-N. Tu, W.-H. Huang, et al., "15.5 A 28nm 64Kb 6T SRAM Computing-in-Memory Macro with 8b MAC Operation for AI Edge Chips," in 2020 IEEE International Solid-State Circuits Conference - (ISSCC), 2020, pp. 246–248. DOI: 10 . 1109 /ISSCC19947.2020.9062995.
[3] P.-C. Wu, J.-W. Su, Y.-L. Chung, et al., "A 28nm 1Mb Time-Domain Computing-in-Memory 6T-SRAM Macro with a 6.6ns Latency, 1241GOPS and 37.01TOPS/W for 8b-MAC Operations for Edge-AI Devices, year=2022," in 2022 IEEE International Solid-State Circuits Conference (ISSCC), vol. 65, pp. 1–3. DOI: 10.1109/ISSCC42614.2022.9731681.
[4] X. Si, W.-S. Khwa, J.-J. Chen, et al., "A Dual-Split 6T SRAM-Based Computing-in-Memory Unit-Macro With Fully Parallel Product-Sum Operation for Binarized DNN Edge Processors," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 66, no. 11,pp. 4172–4185, 2019. DOI: 10.1109/TCSI.2019.2928043.
[5] X. Si, J.-J. Chen, Y.-N. Tu, et al., "A Twin-8T SRAM Computation-in-Memory Unit-Macro for Multibit CNN-Based AI Edge Processors," IEEE Journal of Solid-State Circuits, vol. 55, no. 1, pp. 189–202, 2020. DOI: 10.1109/JSSC.2019.2952773.11