國立清華大學 電機工程學系

實作專題研究成果報告

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

專題領域：系統組

組　　別：A405

指導教授：張孟凡

組員姓名：葉曜暟

研究期間：2023年7月1日至　2024年5月8日止，共10個月

Department of Electrical Engineering, National Tsing Hua University
Special Topic on Implementation Research Report

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

Major Category: System Groups

Group Number: A405

Advisor: Chang, Meng-Fan

Members: Yeh, Yao Kai

Research Period: From (2023/07/01) to (2024/05/08). (10 months)

Abstract

Among many kinds of neural networks, Convolutional Neural Network (CNN) is a popular neural network in deep learning. It's because CNN is powerful for image sensing and speech recognition. In CNN, Multiply and Accumulations (MAC) between weights and inputs are performed in all layers. Weights are usually stored in memory units, and inputs are fed by users.

MAC operation is performed in memory unit in computing-in-memory (CIM) macro, and SRAM is suitable for this approach. Existing voltage-domain approaches like charge sharing [1] and voltage dividing [2] can correctly perform MAC operation, but they suffer from trade-offs between output precision and signal margin.

In this project, I implement a time-domain CIM 6T-SRAM macro for 8b-MAC operation [3] with TSMC 65nm process technology. It computes MAC result by counting delay time accumulation and keeps same signal margin as increasing accumulation. It consists of delay computing unit (DCU), reference delay generator, and dynamic differential-reference time-to-digital converter (D2REF-TDC).

However, edge delay cell (EDC) in DCU suffers from process-voltage-temperature (PVT) variation and leakage issue. It makes delay time nonuniform with same input value, which degrades inference accuracy. Moreover, long computing cycle with larger macro slows down inference speed.

In order to reduce PVT-variation influence, I design RC-Modulation EDC (RCM-EDC) with shorter and more stable delay time. Currently, RCM-EDC has smaller PVT-variation influence, and its standard deviation of time delay is hundreds of times smaller than EDC's.

摘要

在眾多神經網路中，卷積神經網路 (Convolutional Neural Network, CNN) 在深度學習 (Deep Learning) 尤其受歡迎。這是因爲 CNN 有強大的影像辨識和聲音辨識能力。在 CNN 中，權重 (Weight) 和輸入資料的乘加運算 (Multiply and Accumulation, MAC) 會在每一卷積層運作。權重通常會存在記憶體裡面，而輸入則會由使用者輸入。

在記憶體內運算 (Computing-in-Memory, CIM) 架構中，MAC 運算會在記憶體裡面進行，而靜態隨機存取記憶體 (SRAM) 很適合此方法。目前出現的電壓域計算方法像是電荷共享 (Charge Sharing) [1] 和分阻電壓 (Voltage Dividing) [2] 都可以正確執行 MAC 運算，但是他們都會受輸出精度和信號裕度 (Signal Margin) 之間權衡的限制。

這次專題中，我運用台積電 65 奈米製程實作了之時域記憶體內 8b 乘加運算的六電晶體靜態隨機存取記憶體架構 [3]。此架構透過計數累加的時間延遲計算 MAC 結果，並可以隨著更多的累加維持信號裕度。此架構由計算延遲單元 (Delay Computing Unit, DCU)、延遲基準產生器、動態差分基準時間數位轉換器 (Dynamic Differential-Reference Time-to-Digital Converter, D2REF-TDC) 組成。

然而，邊界延遲器 (edge delay cell, EDC) 受到製程、電壓、溫度差異 (PVT variation) 和漏電問題影響。這讓同樣的輸入值產生不同的延遲時間，並降低推論精準度。此外，更大的架構所產生的長計算週期也拖慢了推論速度。

爲了減少 PVT variation 影響，我設計了可調整 RC 的 EDC(RC-Modulation EDC, RCM-EDC) 並有更短且穩定的延遲時間。目前，RCM-EDC 有比較小的 PVT variation 影響，其延遲時間的標準差比 EDC 小數百倍。

1 Background

Convolutional Neural Network

Convolutional Neural Network (CNN) is powerful for image sensing and voice recognition. In CNN, an input matrix, which could represent image or voice signal, is is passed through a series of convolution filters. Each filter, which is associated with a set of weights, represents an activation function such as ReLU, leaky ReLU, or Sigmoid.

In each layer, Multiply and Accumulation (MAC) is performed on input data and weight, and the result is passed to next layer. This process continues until the final layer of the network. These output from the final layer are then processed through Max Pooling and Fully Connected Layers to determine the most possible result.

Given that MAC operation is performed repeatedly in the process of a CNN, it's of interest to reduce its energy consumption.

Near Memory Computing and Computing In Memory

Near Memory Computing (NMC) performs computations near a memory unit. Sensing amplifiers (SA) located nearby process the results directly. NMC usually uses non-volatile memory (NVM) with small area, such as MRAM and ReRAM. As a result, NMC has a higher Area Efficiency (AF) than CIM. However, due to the slower speed of NVM compared to SRAM, NMC has lower speed than CIM.

Computing In Memory (CIM) performs computations directly in memory unit. Its memory type is usually SRAM, which has larger area but faster speed compared to NVM. Hence, CIM usually has a faster speed but lower AF compared to NMC.

CIM/NMC-Memories operate in two modes: (1) Memory mode and (2) CIM/NMC mode. In memory mode, the memory stores weights like a normal storage process. In CIM/NMC mode, the system performs MAC operations and obtain convolution results. Both computing methods can save data transmission energy by reducing transmitting path to buffer, thereby increasing Energy Efficiency (EF).

Voltage-Domain SRAM-CIM Approach

It's intuitive to represent values using different voltage levels. It's the concept of voltage-domain approach. Voltage-Domain SRAM-CIM operation struggles with trade-off between output precision and power consumption.

The supply voltage is divided in to segments, each representing a specific value. The size of these segments shouldn't be too small to maintain adequate signal margin. As accumulation increases, the signal margin decreases, and this decrease is exponential under a fixed power supply. Under fixed power supply V_{DD}, the relationship between output precision and signal

margin can be expressed by equation 1. This poses a significant challenge because output value often includes a wide range of values.

$$\frac{\mathbf{V}_{\mathrm{DD}}}{\mathbf{2}^{\text{bit precision}}} = \textbf{signal margin} \tag{1}$$

2 Research Methodology

Time-Domain SRAM-CIM Approach [3]

I implement a time-domain SRAM-CIM macro for 8b-MAC operation with TSMC 65nm process technology, as shown in Fig.1. This macro computes MAC result of 8b input, 1b weight, and 64 accumulations by counting delay time accumulation and keeps same signal margin as increasing accumulation. Multiplications of weights and inputs are performed in edge delay cells (EDC) by generating different delay time. Delay Computing Units (DCU) then pass delayed enable signal to next DCU. This time accumulation process is continued until the final DCU of the macro. Finally, dynamic differential-reference time-to-digital converter (D2REF-TDC) converts different delay time to digital output.

However, edge delay cell (EDC) in DCU suffers from process-voltage-temperature (PVT) variation and leakage issue. It makes delay time nonuniform with same input value, which degrades inference accuracy. Moreover, long computing cycle with larger macro slows down inference speed.

Currently, I attempt to design new EDC with higher anti-PVT variation performance, address EDC leakage issue, and shorten computing cycle of D2REF-TDC.

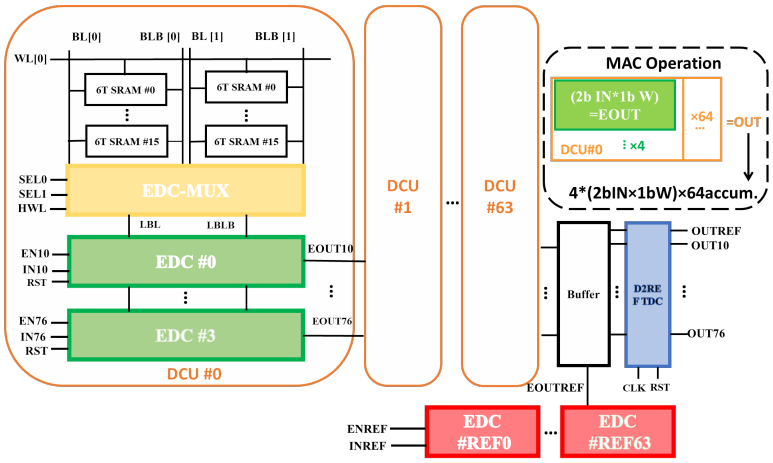

Fig. 1 Time-Domain SRAM-CIM Macro Block Diagram

3 Experimental Results

3.1 Measurement

8b Input∗64b Weight MAC Operation

Source Name	2b Input	Voltage (V)	Time Delay	Total Delay	Digital Output
V_{REF}	00	0	t_0	$64t_0$	$64t_0$
V_{in76}	00	0	t_0	$64t_0$	0
V_{in54}	01	0.688	$t_0+\Delta t$	$64t_0+64\Delta t$	64
V_{in32}	10	0.699	$t_0+2\Delta t$	$64t_0+128\Delta t$	128
V_{in10}	11	0.718	$t_0+3\Delta t$	$64t_0+192\Delta t$	192

Table 1: 8b Input∗64b Weight MAC Result Table

Signal Margin

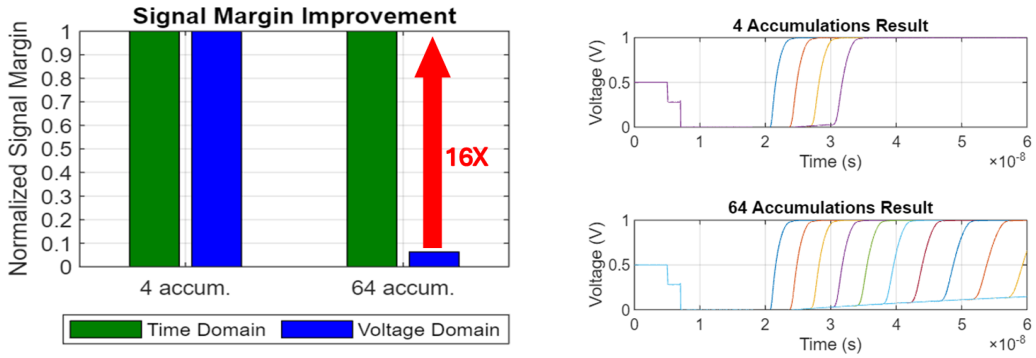

Fig. 2 Signal Margin Improvement [3]

Leakage Problem

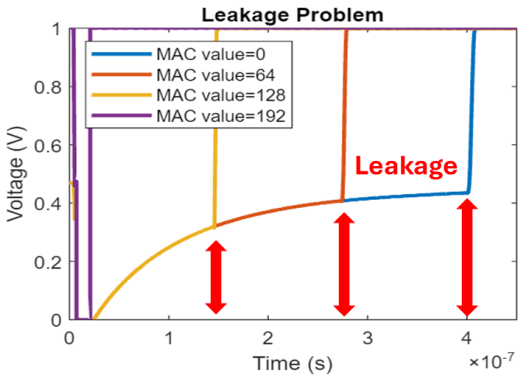

Fig. 3 Leakage Problem

4 RC-Modulated Edge Delay Cell

I design a RC-Modulated Edge Delay Cell (RCM-EDC) as shown in Fig.4.RCM-EDC apparently has more separate and shorter time delay, which means it has stronger anti-PVT variation performance.

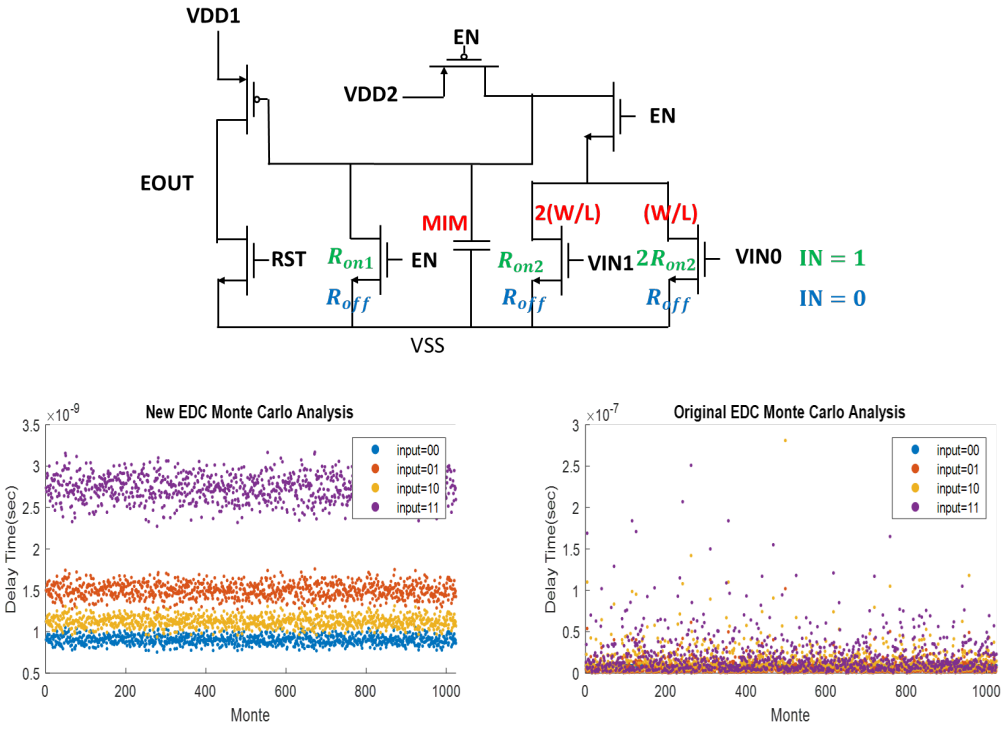

Fig. 4 RC-Modulated (RCM) EDC Scheme and Performance

5 Conclusion

In this Project, EDC wisely feed weight value through to gate port, this can avoid 6T SRAM read disturb problem, and it use EDC-MUX to increase computing capacity with lower area

cost. D2REF-TDC can save power by deleting overlapped counted time delay. Although time-domain design can indeed solve signal margin and output precision issue, but it has leakage and long computing time problems. Moreover, it also suffers from PVT variation.

Beyond these potential problem, time-domain approach is an effective way to solve signal margin issue. As a result, time-domain is suitable for result-concentrate dataset and high precision data set.

To solve PVT-variation issue of EDC. I design a new RCM-EDC which generates different delay time by different equivalent discharge resistance generated by different parallel input on/off resistors. This makes RCM-EDC has stronger immunity to PVT variation (Fig.5).

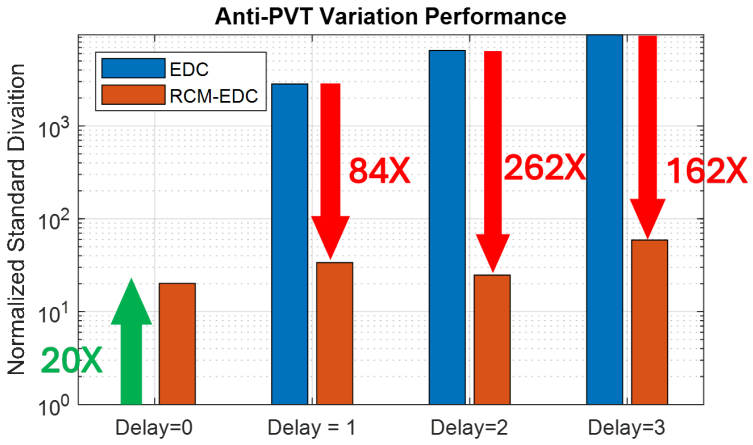

Fig. 5 RCM-EDC Anti-PVT Variation Performance

7 Reference

[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," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 1–3. DOI: 10.1109/ISSCC42614.2022.9731681.

8 Review and Reflections

In this project, I learn knowledge of different types of memory, skills of EDAs, and presentation skills. These are the things I've never touched before, and now I have the basic concept

of them.

In the beginning, I read a lot of essays slowly and ask a lot of help from mentor. During this project, I discuss different design with my classmates, know how to catch the important part of the paper, and try to implement what I read. There are some challenges during the process, but I put a lot of effort and get some helps from others. Eventually, I successfully get the right results and find out the pros and cons of the design. It impresses me the fact that we need to think the potential challenges that are not mentioned in what we read.

Finally, I can use what I read to design and name a new circuit with better performance I wants. I failed many times and deletes the wrong paths I have tried. It's a new experience that I've never had.

I'm really appreciate that professor Chang, Meng-Fan gives me this opportunity to conduct this project. I'm also grateful to the seniors and mentor in this project. They provide resource, point out the thing I miss, and tell me the better way to work efficiently, They really help me a lot when I face difficulties and challenges.

After project, I know that there are still many knowledge that I must figure out. I hope this experience will help me out in the future when I face similar challenges.