國立清華大學 電機工程學系

實作專題研究成果報告

A 65 nm ReRAM Compute-in-Memory Structure Using Residue Shared ADC for AI Edge Device

使用65奈米製程殘差共享類比數位轉換器之電阻式隨機存取記憶體內運算架構

專題領域： 系統組

組　　別： A521

指導教授： 張孟凡

組員姓名： 孫德安

研究期間： 113年07月01日至114年05月05日止，共10個月

Abstract

AI edge devices that rely on convolutional neural networks (CNNs) demand a significant amount of computation. The conventional von Neumann computing architecture suffers from considerable latency due to frequent data transfers between memory and the CPU, resulting in a bottleneck that limits overall computing speed. The concept of compute-in-memory (CIM) has been proposed to address this issue.

Resistive random-access memory (ReRAM) is commonly used in non-volatile compute-in-memory (nvCIM) architectures. Implementing parallel analog multiply-and-accumulate (MAC) operations directly within memory cell arrays can greatly reduce latency and improve energy efficiency. This project focuses on the analog-to-digital converter (ADC), a peripheral circuit in the CIM structure, which converts the analog value generated by MAC operations into digital signals.

In this project, we first implemented a residue-shared ADC (RS-ADC) scheme from reference research [1] with TSMC 65nm process technology. The RS-ADC readout only 2-bits MSB information of the analog current value in each cycle, while accumulating the remaining LSB information in voltage domain. This residue-shared concept can reduce toggling rate for LSB digitization, enhancing energy efficiency.

During the implementation of the RS-ADC, we encountered a challenge: the RS-ADC failed to sense correctly in voltage-mode when the input voltage was too low. To solve this problem, we proposed a multi-bit-friendly voltage sense amplifier (multi-bit-friendly VSA). This design incorporates both NMOS and PMOS input differential pairs to ensure proper functionality across the full voltage range from 0 to VDD.

Experimental results show that the proposed multi-bit-friendly VSA reduces sensing latency and sensing offset in Monte Carlo simulations, but at the cost of higher power consumption. Nevertheless, it achieves up to a 6.46× improvement in performance compared to the original VSA scheme.

摘要

涉及卷積神經網路(CNNs)的 AI 邊緣運算應用中，需處理大量的乘積運算(MAC operation)。傳統的馮·諾伊曼架構下，大量的中間數據必須在 CPU 與記憶體之間頻繁傳輸，成為提升運算速度的瓶頸。為了解決馮·諾伊曼瓶頸，因此提出了記憶體內運算（Compute-In-Memory）這樣的概念。

電阻式隨機存取記憶體（ReRAM）常被應用於非揮發性記憶體內運算（nvCIM）架構中。其原理為直接在記憶體單元陣列中平行執行大量的類比乘加運算，以大幅降低延遲並提升能效。本專題著重於類比數位轉器（ADC）的部分，這是一種記憶體內運算架構中的周邊電路，用於將乘加運算所產生的類比電流值轉換為數位輸出。

在本專題中，我們首先參考文獻 [1]，使用台積電 65 奈米製程實作出殘差共享類比數位轉換器（Residue-shared ADC）架構。此 RS-ADC 在每個運算週期中只讀取類比電流值的 2 位元最高有效位（MSB）資訊，並將剩餘的最低有效位（LSB）資訊以電壓形式累積於電容當中，進行跨運算週期的累加。這種殘差共享的概念能大幅降低將 LSB 數位化時所需的讀取次數，因此能提升能量效率。

在實作 RS-ADC 的過程中，我們遇到了一個挑戰：當輸入電壓過低時，RS-ADC 在電壓模式下無法正確地運作。為了解決這個問題，我們提出了一種適合於多位元讀取的電壓感測放大器（multi-bit-friendly VSA）。此設計同時使用 NMOS 與 PMOS 作為輸入差動對，當輸入電壓太小以至於打不開 NMOS 差動對時，仍可以藉由另一組 PMOS 差動對進行電壓感測。如此一來，能確保輸入電壓在 0 到 VDD 的所有範圍都能正常運作。

實驗結果顯示，我們改良的多位元友善電壓感測放大器（multi-bit-friendly VSA）不只確保在所有電壓範圍中都能正常運作，經過蒙地卡羅(Monte Carlo)模擬後，也發現有著其他優勢:更低的感測延遲以及更低的感測裕度(sensing offset)。儘管以能耗提升作為代價，但在與先前的感測放大器架構比較中，其性能最多仍可提升達 6.46 倍

1. Introduction

1.1 Background

In AI edge computing applications involving convolutional neural networks (CNNs), a large amount of multiply-and-accumulate (MAC) operations must be processed. Under the traditional von Neumann architecture, data must first be fetched from memory and then transferred to the CPU for computation. This results in frequent data movement between the memory and CPU, which becomes a bottleneck to improving computational speed.

The concept of non-volatile compute-in-memory (nvCIM) has been proposed to overcome this von Neumann bottleneck. Implementing parallel analog multiply-and-accumulate (MAC) operations directly within memory cell arrays can greatly reduce latency and improve energy efficiency. In addition, its non-volatile nature allows data to be retained during power-off states, while also reducing standby power and wake-up latency.

Resistive random-access memory (ReRAM) is a type of non-volatile memory. This project focuses on its two readout modes: current-mode and voltage-mode. Current-mode readout offers high accuracy but also incurs higher power consumption due to the DC current generated within the memory cell array. In contrast, voltage-mode readout consumes less power, but its accuracy in multi-bit readout is limited by the supply voltage (VDD) headroom.

Research [2] proposed **hybrid-mode sense amplifier (HMSA)**, combining pros of two sensing mode. Which sensing 2-bits MSB by current-mode to ensure accuracy, and sensing 3-bits LSB by voltage-mode to reduce energy consumption. However, this computing flow must sense 5-bits in each cycle, which means HMSA needs to toggle 5 times and reset to its initial state after each toggling. We can see that energy consumption is large due to multi-bit readout. Hence, research [1] based on previous work, proposed **residue-shared ADC (RS-ADC)** scheme, which stack 3-bits LSB information in voltage-domain through a capacitor over multiple cycles, then eliminates LSB readout toggling times to reduce the overall sensing power.

1.2 Motivation and Purpose

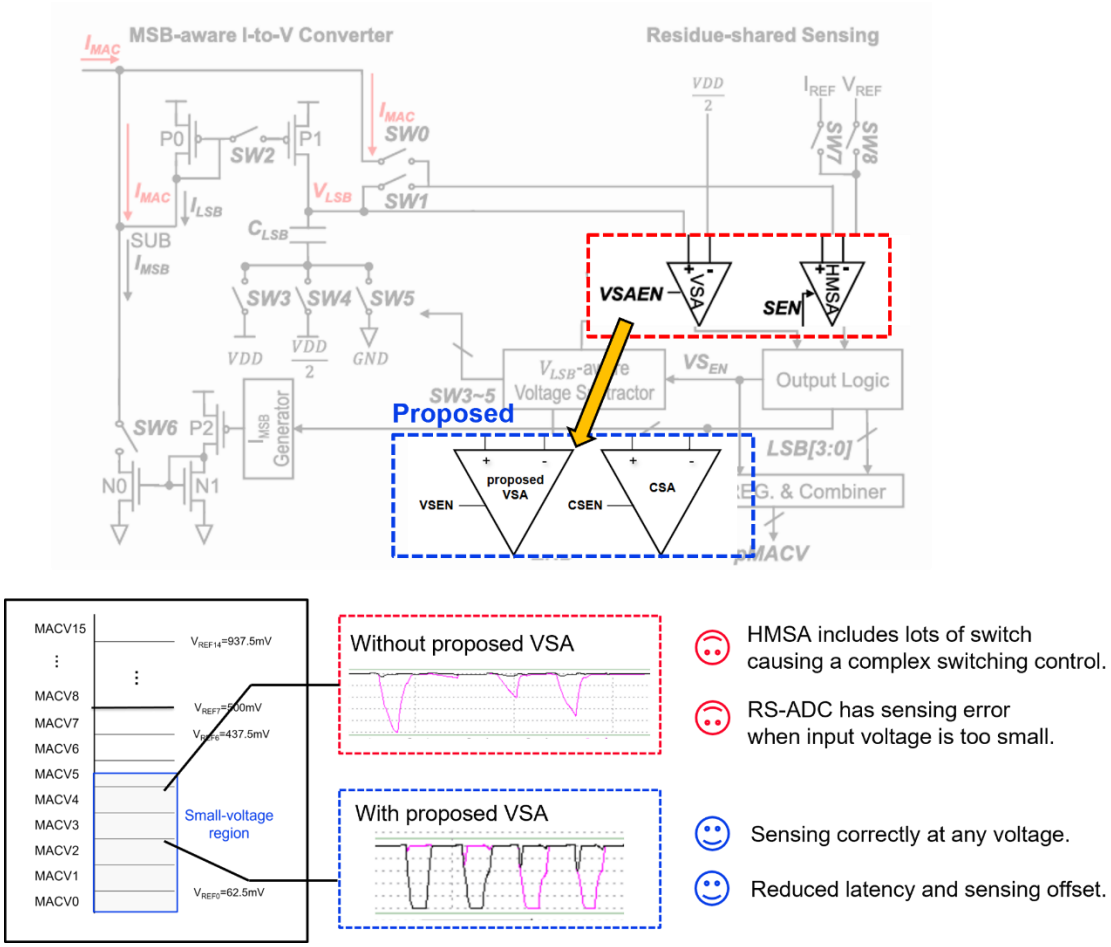


Fig. 1 Original RS-ADC structure and its challenges

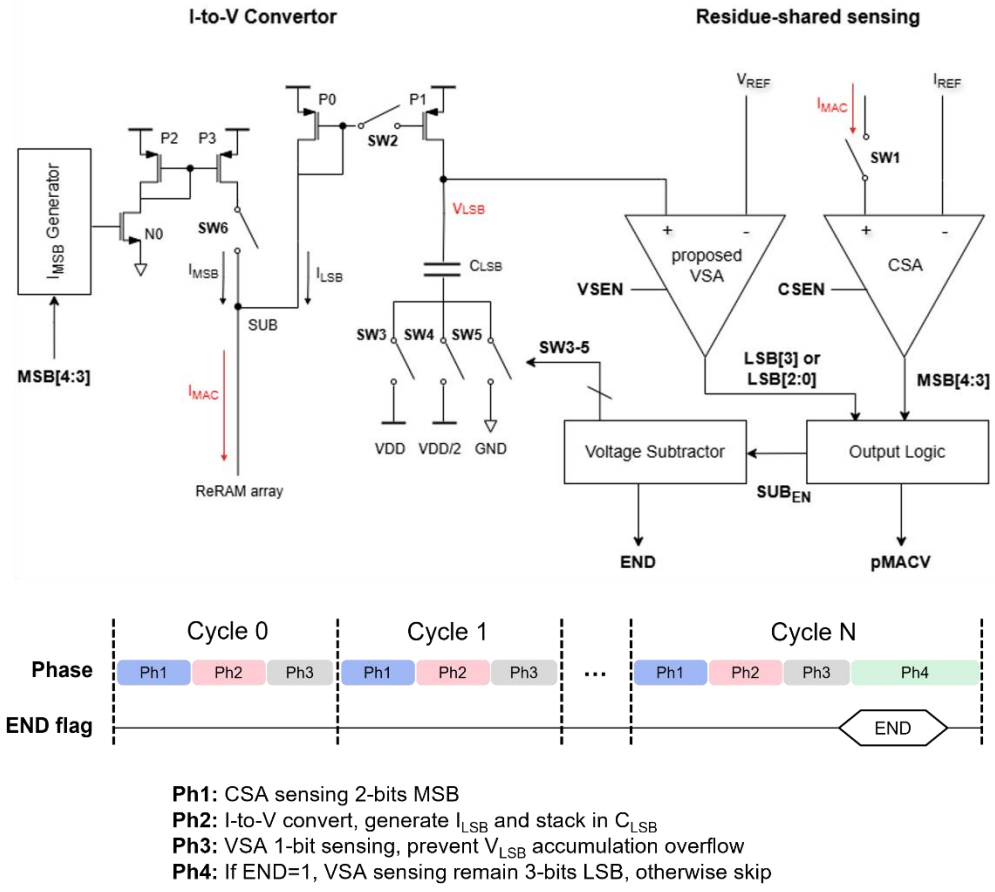
When implementing the RS-ADC scheme in [1], We found the following challenges: 1) The use of HMSA includes lots of switch, causing a complex switching control and unexpected coupling effect; and 2) The RS-ADC failed to sense correctly in voltage-mode when the input voltage dropped below approximately 400 mV. This issue occurred when the corresponding MAC value was too small, causing the voltage to fall into the small-voltage region, as shown in Fig. 1.

To solve these problems, we modified the RS-ADC structure by **replacing the HMSA with two separate SA**: a current-sense amplifier (CSA) and a voltage-sense amplifier (VSA). Moreover, we also proposed a **multi-bit-friendly VSA** that can operate correctly under VDD/2.

2. Research Methodology

2.1 High Level Architecture and Workflow

Fig. 2 shows our modified 5-bits RS-ADC structure in this project, comprising I-to-V converter and Residue-shared sensing two part. **I-to-V converter part** contains SW1-SW6, an I_{LSB} current-mirror pair (P0/P1), an I_{MSB} current-mirror pair (P2/P3), an I_{MSB} generator, an integration capacitor (C_{LSB}) for LSB accumulation. **Residue-shared sensing part** contains a CSA, a multi-bit-friendly VSA, a voltage subtract controller, and other control circuits.



Ph1: CSA sensing 2-bits MSB
Ph2: I-to-V convert, generate I_{LSB} and stack in C_{LSB}
Ph3: VSA 1-bit sensing, prevent V_{LSB} accumulation overflow
Ph4: If END=1, VSA sensing remain 3-bits LSB, otherwise skip

Fig. 2 Modified RS-ADC structure and workflow

The overall work flow can be divided into four phases. Phase1 (Ph1) involves CSA current-mode sensing, yielding 2-bits MSB output (DOUT[4:3]). Phase2 (Ph2) involves I-to-V converter, which first generates I_{LSB} ($I_{LSB} = I_{MAC} – I_{MSB}$), then converts I_{LSB} into voltage-domain (V_{LSB}) and stack into C_{LSB} to do voltage accumulation across multiple cycles. Phase3 (Ph3) involves 1-bit sensing to prevent

overflow due to voltage accumulation. Phase4 (Ph4) involves VSA voltage-mode sensing, yielding 3-bits LSB output (DOUT[2:0]) when END signal is activated; otherwise if END=0, it will skip Ph4 and return to Ph1.

The following will first introduce more details about modified RS-ADC structure and operation, which the different is that we replace original HMSA with a single CSA and a proposed VSA. Note that with this modification, we can avoid complex switching control. The last part of this chapter will introduce proposed multi-bit-friendly VSA. Note that this can operate correctly under VDD/2 and increase accuracy during multi-bits sensing.

2.2 Residue-shared ADC operation

To simulate current-mode ReRAM compute-in-memory, we set the ratio of R_{HRS}:R_{LRS} = 1:10. The corresponding current will be I_{HRS}=10uA and I_{LRS}=100uA, which represent 0 and 1, respectively. Since we assume 32 wordline (WL) activated in each cycle, after doing multiply-and-accumulate (MAC) operation by compute-in-memory, the I_{MAC} is range from 320uA to 3110uA, which represent MAC value (MACV) from 0 to 31. Inform that ideal current is set to simulate I_{MAC} in this project.

In Ph1, SW1 is on and I_{MAC} flow into CSA to sensing 2-bits MSB. We used current-mode sensing by conventional latch-type CSA, ensuring the accuracy of 2-bits MSB. Note that 2-bits MSB are read out every cycle and store in D flip-flop (DFF) as DOUT[4:3] since they are the most important part in a MAC value.

In Ph2, SW1 is off and SW2/SW6 is on to activate the I-to-V converter scheme. The I_{MSB} generator will generate I_{MSB} according to DOUT[4:3] from Ph1. The corresponding current I_{MSB} is the source of node SUB and I_{MAC} from ReRAM array is the sink, alternatively. Hence, node SUB will occur current subtraction, generating $I_{LSB} = I_{MSB} - I_{MAC}$. The current of I_{LSB} is coped by current-mirror pair P0/P1 and the analog $MACV_{LSB}$ is stacked in C_{LSB} by charge accumulation, representing 3-bits V_{LSB}[3:0] in voltage domain.

In Ph3, 1-bit detection of the V_{LSB} is deduct by VSA to prevent overflow during voltage accumulation across multiple cycles. If it found out that V_{LSB} is over VDD/2, an enable signal SUB_{EN} will be activated. Then voltage subtract controller will control SW3-SW5 to execute voltage subtraction via DC coupling by switching the voltage source of the C_{LSB}. For example, if SW3 is suddenly off and SW4 is on, the voltage

source of C_{LSB} is change from VDD to VDD/2, causing a coupling effect ($\Delta V = -VDD/2$) to V_{LSB}, complete the voltage subtraction.

However, we can only operate coupling effect twice since there are only three voltage sources: VDD, VDD/2, and GND. After the last coupling effect, the SUB_{EN} will not execute voltage subtraction but generate END signal and move on to Ph4.

In Ph4, the END signal will determine whether to sensing remain 3-bits LSB information store in voltage-domain (V_{LSB}) by VSA, and store in DOUT[2:0]. The END signal becomes active when two coupling effect completed. The END signal generator can be implemented by a DFF, whose data line is connected to SW5 control signal and is triggered by SUB_{EN}. After 3-bits LSB sensing, a group of accumulation is completed and output logic generate a partial MAC value (pMACV).

2.3 Proposed multi-bit-friendly VSA

When performing LSB multi-bit sensing, 4-bits information (MACV0-MACV15) are represented in the voltage domain, requiring VDD to be divided into 16 levels for readout. When the MACV is too small (Small-voltage region showed in Fig. 3), the corresponding voltage may fall below the threshold voltage (V_{th}) of NMOS, making it hard to turn on the input differential pair. Hence, we proposed a multi-bit-friendly VSA that can sense correctly in the small-voltage region.

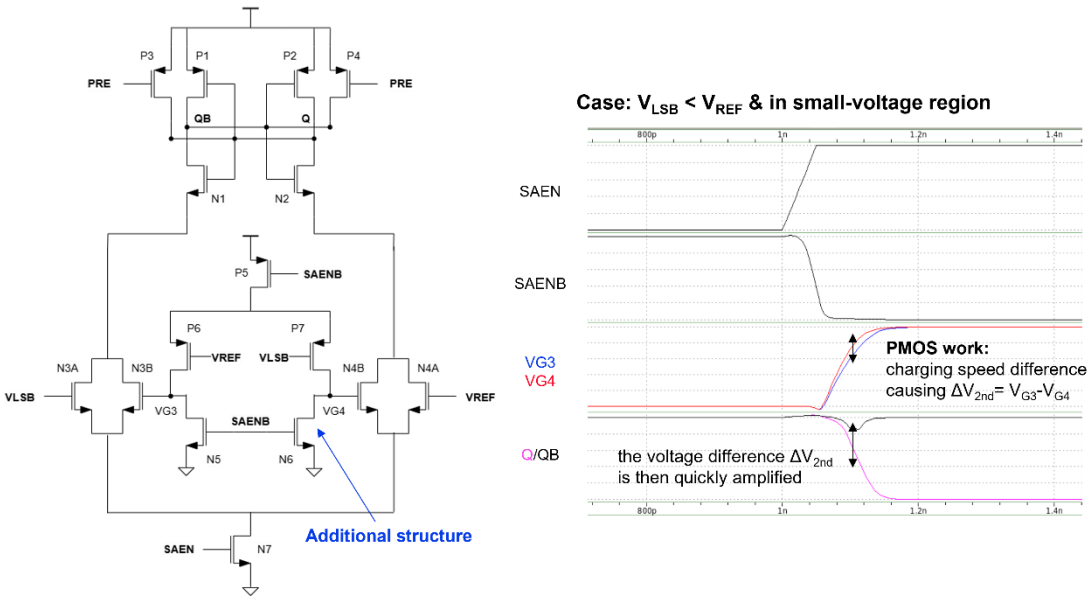


Fig. 3 Proposed multi-bit-friendly VSA scheme and waveform

The proposed multi-bit-friendly VSA is illustrated in Fig. 3, which is optimized based on conventional latch-type VSA in research [3]. It comprises a cross-couple

latch (P1/N1 and P2/N2), two precharge PMOS (P3, P4), a discharge NMOS (N7), and a basic input differential pair (N3A, N4A). The new part of the multi-bit-friendly VSA comprises an additional differential pair (N3B, N4B), an enable PMOS (P5), two charging PMOS (P6, P7), and two initial NMOS (N5, N6).

The central idea of the proposed multi-bit-friendly VSA is to connect the input voltages V_{LSB} and V_{REF} to both NMOS and PMOS transistors, ensuring that the VSA can function properly across the entire voltage range from 0 to VDD. When operate in the small-voltage region, the input differential pair N3A/N4A fails to function properly (V_{LSB}, V_{REF} are below V_{th} of NMOS), but PMOS charging pair P7/P8 still work. The voltage difference in 1st stage ($\Delta V_{1st}= V_{LSB}-V_{REF}$) cause the charging speed difference of V_{G3} and V_{G4}, which are the input voltage of differential pair N3B and N4B, respectively. The voltage difference in 2nd stage ($\Delta V_{2nd}= V_{G3}-V_{G4}$) then cause the voltage difference between Q and QB, then cross-couple latch provides positive feedback to amplify this small difference and complete the sensing process.

In summary, when the input voltage is too low to turn on the NMOS differential pair, an additional PMOS differential pair can be used alternatively. In this way, proper operation across the entire input voltage ranges from 0 to VDD can be ensured.

3. Experimental Results

3.1 Multi-bit-friendly VSA Performance

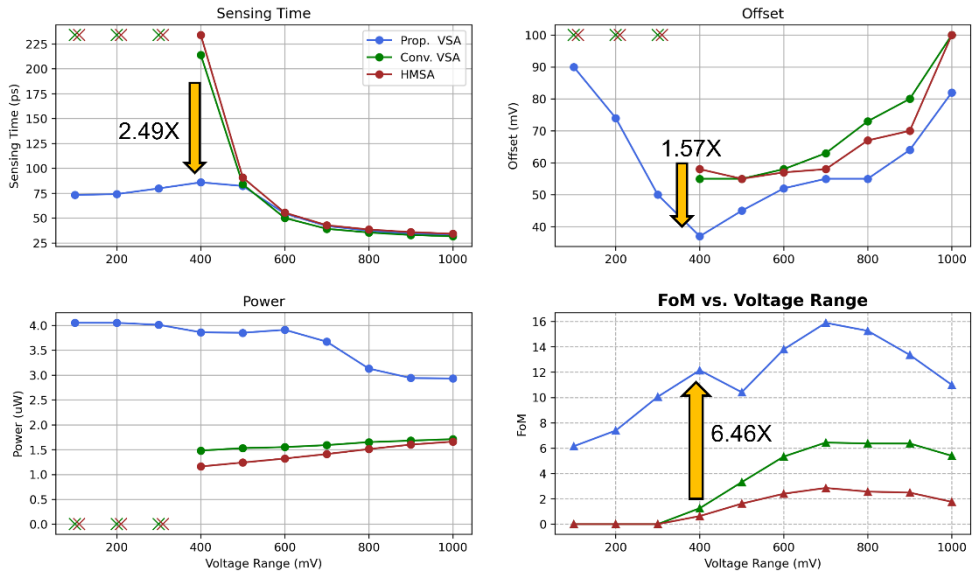


Fig. 4 Performance comparison of different VSA types

Fig. 4 shows the three types of VSA compare with sensing time, offset, power. The offset is defined as the minimum voltage difference between V_{LSB} & V_{REF} with 99.9% yield in 1024 Monte Carlo simulations. The last picture illustrates FoM within the voltage range from 0 to VDD, which is defined as fallow:

$$FoM = \frac{1}{sensing\ time\ (ps) \times power\ (uW) \times offset\ (mV)}$$

Note that FoM is bigger the better.

As noted previously, there is an issue when the VSA operates in the low-voltage region (approximately 0 to 400 mV). Both the conventional latch-type VSA and the HMSA fail to sense correctly in this region (As you can see, the sensing time increases dramatically as the operating voltage decreases, eventually leading to sensing error), which is indicated by a cross mark in Fig. 4.

Only proposed multi-bit-friendly VSA can operate correctly in this region with the help of additional PMOS charging pair (P6, P7) mentioned before. Although this comes at the cost of higher power consumption, the resulting FoM of multi-bit-

friendly VSA is best across the entire voltage range from 0 to VDD.

3.2 RS-ADC operation waveform

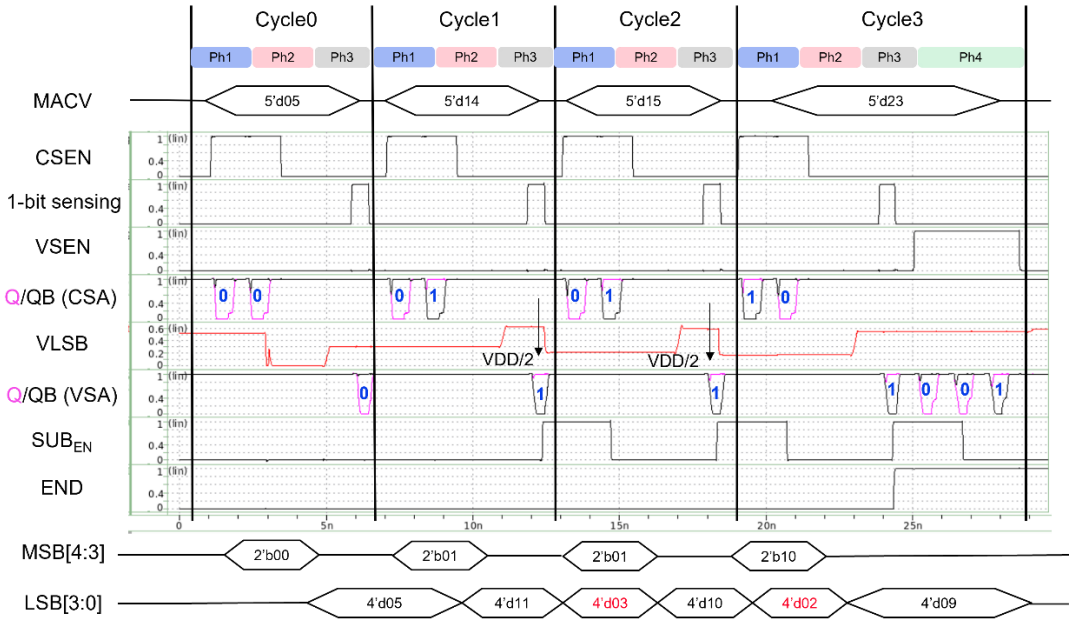


Fig. 5 Illustrative computation of the RS-ADC over four cycles.

Fig. 5 shows the waveform of RS-ADC over four computing cycles. In this case, we assume that MAC values of 5, 14, 15, and 23 are sequentially input over four cycles.

The MACV in **Cycle1 is 5** and the 2-bits MSB readout "00" in binary form during Ph1. The value for LSB is calculated as: 5 - 0 × 8 = 5. This residue is then stored in V_{LSB} during Ph2.

The MACV in **Cycle2 is 14** and the 2-bits MSB readout "01" in binary form. The residue value for LSB is calculated as: 14 - 1 × 8 = 6. This value is added to the previous V_{LSB}, resulting the value 11. During Ph3, 1-bit sensing detects that the value exceeds 8 (i.e., the voltage surpasses VDD/2), which triggers the SUB_{EN} signal and coupling effect occurs. After subtraction, the value becomes 11 − 8 = 3, which is highlighted in red.

The MACV in **Cycle3 is 15** and the 2-bits MSB readout "01" in binary form. The residue value for LSB is calculated as: 15 - 1 × 8 = 7. This value is added to the previous V_{LSB}, resulting the value 10. SUB_{EN} is trigged by 2nd time, yielding a value

of 10 – 8 = 2.

The MACV in **Cycle4 is 23** and the 2-bits MSB readout “10” in binary form. The residue value for LSB is calculated as: 23 - 2 × 8 = 7. This value is added to the previous V_{LSB}, resulting the value 9. Since two coupling effect are done, the END signal is activated, resulting in Ph4 LSB sensing and output the value 9 in binary form: “1001”. The final pMACV is computed as follow:

MSB readout:	(0 + 1 + 1 + 2) × 8
V_{LSB} 1-bit detect:	(1 + 1) × 8
LSB readout:	9

The resulting pMACV = (0 + 1 + 1 + 2) × 8 + (1 + 1) × 8 + 9 = 57, which equals the sum of four input MAC values = 5 + 14 + 15 + 23 = 57. Note that with residue-shared operation, the total toggling rate can be reduced during multi-bit sensing.

4. Conclusion

In this project, we have successfully implemented the residue-shared ADC (RS-ADC), which aims for lower toggling rate and energy consumption. With our optimized, the proposed multi-bit-friendly VSA ensure the sensing functionality even operate with small voltage. It also shortens the latency and sensing offset at the cost of higher power consumption. Nevertheless, the FoM shows an improvement of up to 6.46× compared to the original VSA scheme.

5. Reference

[1] H. -H. Hsu et al., "A 22 nm Floating-Point ReRAM Compute-in-Memory Macro Using Residue-Shared ADC for AI Edge Device," in IEEE Journal of Solid-State Circuits, vol. 60, no. 1, pp. 171-183, Jan. 2025

[2] H.-H. Hsu et al., “A nonvolatile AI-edge processor with SLC–MLC hybrid ReRAM compute-in-memory macro using current–voltage hybrid readout scheme,” IEEE J. Solid-State Circuits, vol. 59, no. 1, pp. 116–127, Jan. 2024.

[3] B. Wicht, T. Nirschl and D. Schmitt-Landsiedel, "Yield and speed optimization of a latch-type voltage sense amplifier," in IEEE Journal of Solid-State Circuits, vol. 39, no. 7, pp. 1148-1158, July 2004

6. Review and Reflection

在這一年的專題中，我們在實驗室的帶領下接受了一連串扎實的訓練。在暑假時，我們初次接觸並學習使用 Hspice, Laker 等電路設計的 EDA tool。開學後，我們開始大量閱讀記憶體內運算的 paper 與期刊，讓我們對記憶體內運算電路有更進一步的了解。且藉由一周一次的 meeting 中同學輪流報 paper，認識到 SRAM, ReRAM, MRAM 等各種不同記憶體的運算架構。而在暑假時，我們使用了先進製程嘗試畫了一顆 offset-canceling sense amplifier 的 layout，雖然是一項巨大的挑戰，但也是一個非常難得的學習機會，並且讓我的最後兩周暑假過得相當刺激。

在這一學期，我們選定一篇 paper 作為實作目標，並開始努力發想創新的架構，最後完成這次的專題研究。在做專題的過程中，我學習到了電路模擬的技巧並更加熟悉記憶體電路架構，除此之外，我也學習到了許多電路以外的事情，比如說實驗室非常重視的 High level 敘事技巧，在報 paper 的同時學習如何讓其他人更能快速理解，這在往後做研究或許也是非常重要的能力。這一年中實在是獲益良多。

特別感謝這一年指導我的 mentor，在研究忙碌之餘還總是能解答我的疑惑，並且與我討論專題遇到的困難，在專題實作的過程中給了我需多有用的建議。也感謝博班學長不辭辛勞地帶領我們完成專題，教導我們完成各種培訓。另外，實驗室的學長姐與專題同學也都非常友善，在輪流報 paper 的過程中，我也從他們身上學習到許多。

最後感謝張孟凡教授的指導，也感謝教授讓我有這個機會能進到實驗室學習，並且順利地完成了這一年的專題研究。