國立清華大學 電機工程學系
實作專題研究成果報告

Design and Analysis of a Sense Amplifier Circuit Compatible with RRAM Array
相容於電阻式記憶體陣列之感測電路設計與分析

專題領域：系統組

組　　別：B573

指導教授：葉昭輝

組員姓名：黃柏叡、巫邦碩

研究期間：2025年2月15日 至 2025年10月31日止，共8個月。

1. Abstract

With the rapid development of artificial intelligence (AI), neural network architecture has become increasingly complex, requiring highly parallel Multiply-and-Accumulate (MAC) operations and generating large amounts of intermediate data. Traditional von Neumann architectures suffer from the von Neumann bottleneck, caused by frequent data movement between memory and processing units, which leads to latency and power consumption issues. To alleviate this bottleneck, Computing-in-Memory (CIM) technology has emerged. Among CIM technologies, Resistive Random Access Memory (RRAM) has become an ideal choice due to its non-volatility. The 1T1R (one-transistor—one-resistor) RRAM structure is simple to fabricate and provides a large resistance ratio between high- and low-resistance states, enabling easier data readout. However, the low-resistance state results in larger current, leading to higher power consumption. Reducing current can mitigate this issue, but it also decreases the margin between the input and reference values, making it harder for the sense amplifier (SA) to distinguish the states. Therefore, a highly sensitive SA is required.

This study is based on the Triple-Margin Small-Offset Current-Mode Sense Amplifier (TMCSA) architecture. Using the 40 nm process provided by TSRI, we designed the TMCSA circuit and integrated a Bitline-Clamping Feedback Amplifier (BCFA) and current mirrors as peripheral circuits to achieve a complete RRAM readout flow. Transistor sizing and circuit design were carried out using Cadence Virtuoso, and the circuit performance was simulated across different process corners, temperature ranges (-40°C to 125°C), and Monte Carlo variations. The results show that the proposed design improves RRAM read accuracy and stability, providing an efficient solution for CIM applications.

2. Introduction

The rapid growth of AI has driven increasingly complex neural networks that require parallel MAC operations and generate large intermediate data. Non-volatile memory is used to reduce standby power, but the von Neumann architecture suffers from data-transfer bottlenecks. Computing-in-Memory (CIM) alleviates this by performing computation and storage together. Among memory options, RRAM is promising due to its simple 1T1R structure, large HRS/LRS ratio, and good readability.

However, RRAM's low-resistance state causes high current and power consumption, and lowering the current reduces sensing accuracy, requiring a highly sensitive sense amplifier (SA). This work designs an efficient SA based on the TMCSA architecture,

implemented in a 40-nm TSRI process, and integrates a Bitline-Clamping Feedback Amplifier and current mirrors for full RRAM readout.

Using Virtuoso, we optimize device sizes and simulate process, temperature, and Monte Carlo variations. The design improves the trade-off between RRAM power and read accuracy, providing a reliable CIM solution for AI accelerators.

3. Research Methodology

3.1. RRAM-based Computing-in-Memory (CIM)

CIM is a hardware architecture that uses non-volatile memory to map neural networks directly onto memory arrays. In RRAM-based CIM architectures, the high-resistance state (HRS) and low-resistance state (LRS) of RRAM cells are used to store weights, with the cell conductance G representing the stored value. When a voltage is applied, current flows through the RRAM device: $I_{ij} = V_i \times G_{ij}$.

The bitline (BL) current is the sum of the currents from all RRAM cells connected to that bitline: $I_j = \sum_i V_i \times G_{ij}$. This bitline current represents the Multiply-and-Accumulate (MAC) value, corresponding to vector—matrix computations in neural networks. Since the MAC output directly affects computing accuracy, the conversion between analog and digital signals (Analog-to-Digital Out) has become one of the major research focuses in modern CIM designs and is also a key topic of this project.

3.2. Resistive Random Access Memory (RRAM)

RRAM has two states: HRS and LRS. When bitline (BL) voltage is 1V and sourceline (SL) voltage is 0V, the measured current through an HRS cell is approximately $1\mu A$, while the current through an LRS cell is about $10\mu A$. For simulations, we directly assigned the resistance values of the RRAM: $1000k\Omega$ for HRS and $100k\Omega$ for LRS.

In addition, we adopted a 1T1R RRAM structure, consisting of one transistor and one resistor. An NMOS transistor controls whether the RRAM cell is enabled. The NMOS gate is connected to the timing-controlled wordline (WL), the source is connected to the resistor and the BCFA, and the drain is connected to the PMOS current mirror.

3.3. Bitline-Clamping Feedback Amplifier (BCFA)

To accurately read the current, the voltage across the RRAM cell must be maintained at approximately 1V. Therefore, based on the architecture described in the reference material, we designed a Bitline-Clamping Feedback Amplifier (BCFA) system. Each BCFA consists of 5 transistors and functions as a simple operational amplifier. One input receives the reference voltage (1V), while the other input is connected to one terminal of the resistor. The output drives an additional NMOS switch.

Since our peripheral circuit requires three BCFAs, we combined them into a single

integrated BCFA structure to reduce the total transistor count. Furthermore, the three BCFAs share the same NMOS current source, ensuring that the operational amplifier delivers a more accurate and consistent voltage to the bitlines.

3.4. Triple-Margin Small-Offset Current-Mode Sense Amplifier (TMCSA)

TMCSA is the core component of the circuit. By controlling the transistors and four switches (SW1~SW4), the operation is divided into four timing phases. In the stand-by mode, SW3 and SW4 are turned on. At the same time, DSD and CHD are enabled, and DN1~DN4 are turned on, initializing the voltages of LQ, LQB, DP1, and DP2 to zero.

In Phase 1 (PH1), CHD and DSD are turned off, VDDSA is enabled. As a result, the threshold voltages (VTH) of P1~P4 are sampled and stored at their gate terminals G1~G4. In Phase 2 (PH2), SW1 and CHD are turned on. Since I_{IN} flows through the left branch, the voltage change at G1 is capacitively coupled to G4 through capacitor C1. In addition, P4 is designed to be twice the size of P1, so it conducts $2I_{\mathrm{IN}}$. Similarly, the right branch carrying I_{REF} operates in the same manner, resulting in P3 carrying a current of $2I_{\mathrm{REF}}$.

In Phase 3 (PH3), SW1 and SW3 are turned off and SW2 is turned on. The voltage difference between LQ and LQB is proportional to the current difference, expressed as $I_{LQ} - I_{LQB} = 3(I_{\mathrm{IN}} - I_{\mathrm{REF}})$, providing a triple current-margin advantage.

Finally, in Phase 4 (PH4), the latch enables signal (SAEN) is asserted high, allowing the latch to detect the voltage difference between LQ and LQB and generate the corresponding digital output. Since TMCSA is highly sensitive to timing, we carefully adjusted the rising and falling edges of all control signals to ensure proper operation.

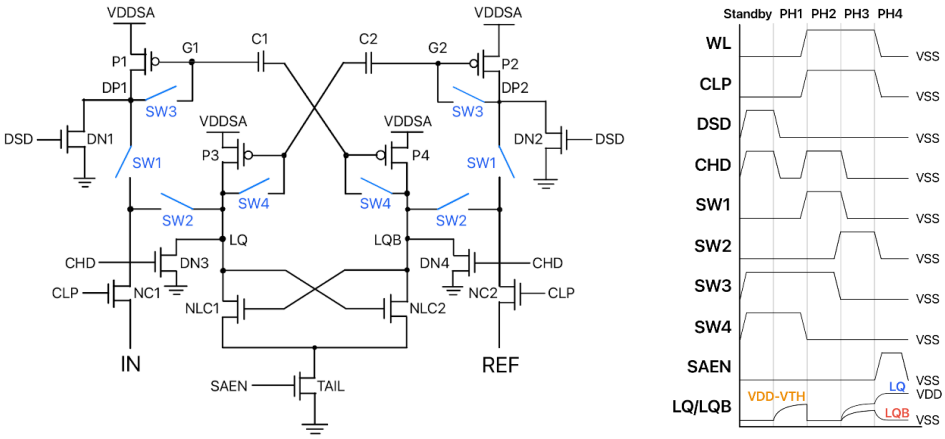

Fig. 1 Sense amplifier circuit (left) and waveform of input and output signals (right) (Drawn by us)

3.5. The Schematic of Sense Amplifier Circuit

To read data from the RRAM, we designed two different architectures during our research. The first allocates a dedicated PMOS current mirror to each RRAM cell to

accurately replicate the current. The second shares a single PMOS current mirror among all RRAM cells, reducing the total number of transistors.

The macro-level circuit diagram of our design is shown in Fig. 2 and consists of three main sections: the input circuit (green), the reference circuit (red and purple), and the TMCSA circuit. The input circuit contains nine RRAM cells. Using the stable 1 V voltage generated by the BCFA system, each RRAM produces a current close to its ideal value. A PMOS current mirror then scales the total input current by a factor of one-fifth. Before entering the TMCSA, an NMOS current mirror further halves the current, so the current delivered to the TMCSA is one-tenth of the total current generated by the nine RRAM cells.

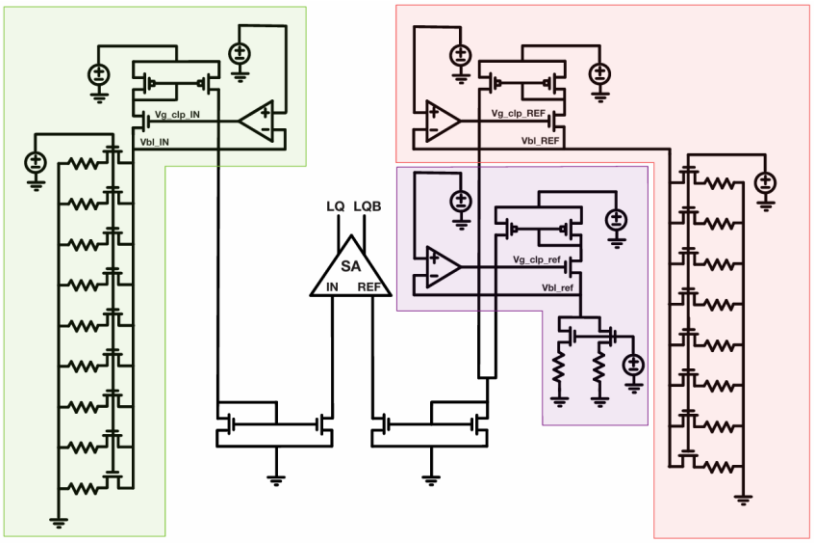

Fig. 2 Sense amplifier circuit architecture (Drawn by us)

4. Experimental Results

4.1. Pre-SIM

In a Convolutional Neural Network (CNN), a commonly used 3×3 convolution kernel corresponds to nine RRAM cells serving as weights. The total current flowing out of the nine RRAM BLs represents the MAC value. As mentioned earlier, the current flowing through a single RRAM cell is $1\mu A$ or $10\mu A$, so the total bitline current from the nine cells ranges from $9\mu A$ to $90\mu A$.

Considering that the current variation range is too large, the sense amplifier (SA) may only maintain high sensing accuracy in either the high- or low-current region, while its performance degrades on the other end. Therefore, we use a current mirror to scale the current down to 0.1× of its original value, reducing the current range to $0.9\mu A$ to $9\mu A$, allowing the SA to maintain a high level of sensing accuracy.

Since the distribution of HRS and LRS in the RRAM array varies with different convolution kernel designs, the bitline current level and sensing yield are also affected.

This study performs two sets of simulations:

4.1.1. Process Corner and Temperature Simulation

Process corners and temperature simulations belong to global variations, which assume that all components in the circuit experience consistent shifts. This allows us to verify whether the circuit can still operate correctly when device characteristics deviate uniformly. The 5 process corners correspond to different transistor speed characteristics: FF, FS, TT, SF, and SS. In addition, to ensure that the circuit can cover a wide range of operating environments, the simulation temperatures are set to −40°C, 27°C, and 125°C.

In the setup, this study selects the two most extreme RRAM state combinations: all HRS (9H0L) and all LRS (0H9L), to evaluate the stability of the SA under boundary conditions. The results show that the proposed circuit architecture successfully produces correct digital outputs for both extreme combinations across all 15 process-corner and temperature conditions. The simulated waveforms show that all the combinations can run correctly, generating right output signal.

4.1.2. Monte Carlo Simulation (MC Simulation)

Monte Carlo simulation belongs to local random variation, which assumes that components in the circuit exhibit random deviations, including channel length (L), channel width (W), threshold voltage (Vth), and so on. This method is used to evaluate the circuit's sensitivity to random variations and to assess its yield.

Therefore, we program the nine RRAM rows with the following ten combinations: 9 HRS, 8 HRS and 1 LRS, 7 HRS and 2 LRS, 6 HRS and 3 LRS, 5 HRS and 4 LRS, 4 HRS e and 5 LRS, 3 HRS and 6 LRS, 2 HRS and 7 LRS, 1 HRS and 8 LRS, and 9 LRS. Each combination is simulated with 1000 Monte Carlo runs to verify whether the overall circuit architecture can maintain its sensing capability under random process variations.

Table 1

# of sensing errors corresponding to each MAC value across 1000 MC simulations.

MAC Value (μA)	9	18	27	36	45	54	63	72	81	90
# of sensing errors	11	2	1	3	8	17	24	39	56	77

(Reference: created by us)

By plotting the yield (number of correct sensing results / 1000) versus the MAC value for these ten configurations, we can observe the trend of yield variation under different bitline current levels.

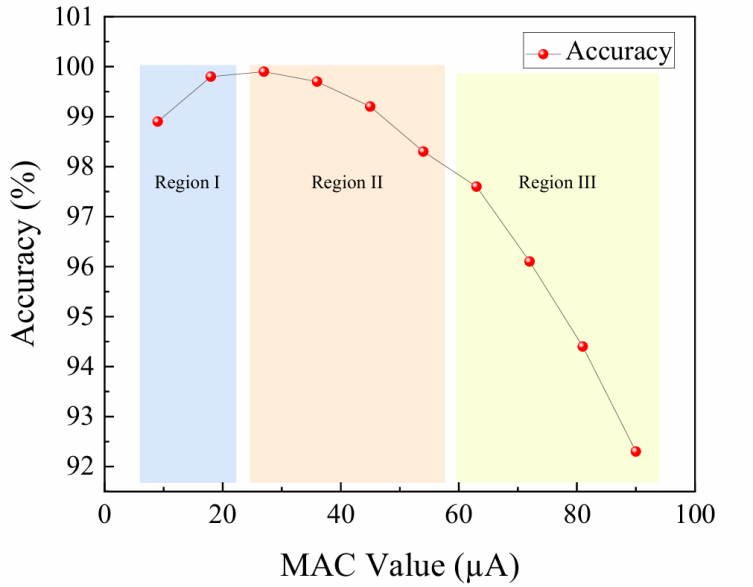

Fig. 3 Plot of yield versus MAC value. (drawn by us)

We observe a distinct distribution pattern that can be divided into three trend regions. The first region spans from $9\mu A$ to $18\mu A$, referred to as Region I. The next region spans from $27\mu A$ to $54\mu A$, referred to as Region II. The last region spans from $63\mu A$ to $90\mu A$, referred to as Region III.

In Region I, the yield decreases as the current becomes smaller. By observing the output current from the SA, we find that this current is smaller than 0.1 times the designed bitline current. In other words, the current replicated by the current mirror has become inaccurate. Because current mirrors are highly sensitive to voltage variations, their output impedance must be designed to be very high to reduce the influence of voltage variations from other parts of the circuit. Therefore, we designed the channel length of the current mirror to be $3\mu m$ to increase its output impedance.

According to the current equation, to maintain the same current, the overdrive voltage must increase. Under this condition, V_{DS} must also increase to satisfy the saturation requirement $V_{DS} \geq V_{GS} - V_{th}$. However, since the load of the current mirror maintains the same voltage across it, meaning V_D is fixed, the transistor can no longer remain in the saturation region and instead enters linear region. This causes the current entering SA to decrease. The decrease in current further reduces the difference between the input current and the reference current, ultimately lowering the sensing success rate.

In Region II, as the MAC value increases, the output current of the SA also increases. At this point, the gate voltage V_G of the current mirror decreases. Based on the saturation condition $V_{DS} \geq V_{GS} - V_{th}$, the drain voltage V_D provided by the SA does not need to be very high to keep the current mirror in saturation. Circuit simulations confirm that the current mirror indeed returns to the saturation region. However, once the current mirror

re-enters saturation, although variations in the SA's load voltage no longer significantly affect the mirror's output current, this introduces an issue in Phase 3 of the SA operation.

In PH3, the SA converts the output current into $2I_{REF} - I_{IN}$. However, because the current mirror connected to the SA remains in the saturation region, $2I_{REF} - I_{IN}$ is quickly forced back toward I_{IN}. This reduces the SA's sensing capability. As a result, we observe that the yield decreases as the MAC value increases in Region II & III. However, the decreasing trends of Region II and III are not the same. Region II is affected by HRS cells, while Region III is affected by LRS cells.

4.2. Post-SIM

We used Virtuoso to draw the layout as shown in Fig. 4. Since we used resistors combined with transistors to emulate RRAM behavior, the resulting layout differs from an actual RRAM array. Thus, only the core block of the sense amplifier is shown here. We then proceeded with R&C extraction, followed by renewed PVT variation simulations and Monte Carlo simulations.

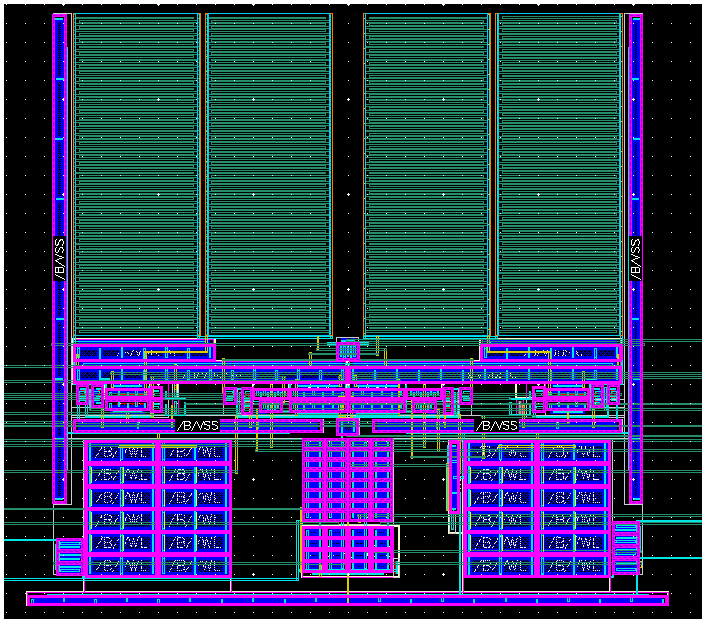

Fig. 4 The layout of the circuit

4.2.1. PVT variation

Like the pre-SIM results, we re-ran the circuit under 15 variation conditions. Among all variation combinations, only the SS corner failed to sense the MAC value as the correct digital output, while all other conditions operated normally.

4.2.2. Monte Carlo simulation

After performing 1,000 Monte Carlo simulations for Code 0~9, we observed an overall improvement in sensing accuracy, except for Code 0, which showed a decline. This is not in line with our expectations.

Table 2

# of sensing errors corresponding to each MAC value across 1000 MC simulations.

MAC Value (μA)	9	18	27	36	45	54	63	72	81	90
# of sensing errors	276	0	0	0	0	5	9	19	45	67

(Reference: created by us)

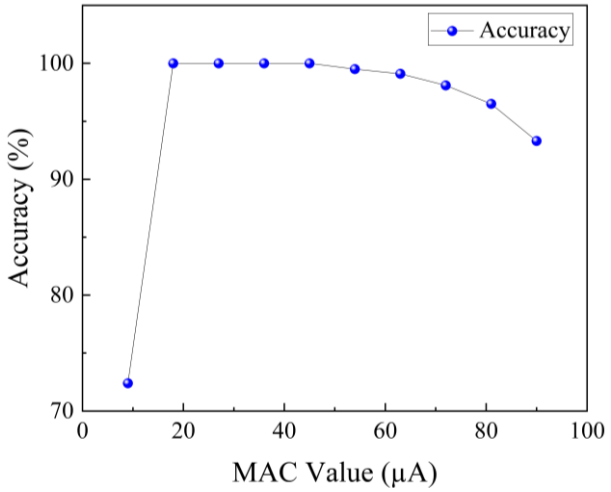

Fig. 5 Plot of yield versus MAC value. (drawn by us)

From Table 2 and Figure 5, there are two points in the post-SIM results that deserve attention. The first is that at the lowest current level, the accuracy is only about 82%, which is much lower than in the pre-simulation and shows a large gap compared with the next current level. The second point is that the accuracy from the second to the tenth combinations is slightly higher than in the pre-simulation, which does not match the expectation that "considering layout effects should reduce accuracy."

We observed that the transistor Vth in the post-SIM is higher than that in the Pre-SIM. A higher Vth not only reduces subthreshold leakage—thereby lowering noise in the SA—but also enhances the offset cancellation mechanism of the SA circuit, resulting in significantly improved sensing performance compared to Pre-SIM. However, since we directly used the built-in resistor from the library to draw the layout, which differs from the actual RRAM structure, this is also one of the reasons for the observed results.

Furthermore, since the layout is not perfectly symmetrical, circuit asymmetry increases. Combined with the impact of parasitic resistance and parasitic capacitance, the current difference between the SA input and the reference becomes larger, which ultimately improves overall sensing accuracy.

5. Conclusion

In this study, we design and analyze TMCSA based on a 40 nm process for RRAM-based CIM architectures. Together with peripheral circuits such as a BCFA, the proposed design successfully enables the readout of MAC results. Process-corner and temperature

simulations verify the circuit remains stable even under the most extreme combinations of HRS and LRS RRAM, demonstrating strong tolerance to process variations.

Further Monte Carlo simulations reveal that the sensing yield is highly correlated with the total bitline current, which can be categorized into three primary regions:

(1) Low MAC Value: The current mirror fails to remain in saturation, leading to increased error rates.

(2) Medium MAC Value: The amplified current difference becomes compressed, dominated by high-resistance RRAM.

(3) High MAC Value: The amplified current difference also becomes compressed, but dominated by low-resistance RRAM.

Overall, the results show the MAC value has a decisive impact on sensing accuracy. The findings provide design insights and potential solutions for future SA and peripheral-circuit optimization, contributing to improved reliability of RRAM-based CIM systems.

6. Reference

[1] Cheng-Xin Xue, Wei-Hao Chen et el., “Embedded 1-Mb ReRAM-Based Computing-in-Memory Macro with Multibit Input and Weight for CNN-Based AI Edge Processors”, IEEE J. Solid-State Circuits, vol. 55, pp. 203-215, 2020.

[2] Yong-Zhi Liu, Po-Rui Huang et el., “Development of 8Kb Embedded RRAM Core Toward Power-Efficient Edge Computing”, Proceedings of 2025 International Electron Devices and Materials Symposium (IEDMS 2025), pp. 92-93, 2025.

[3] D. Saito1, T. Kobayashi et el., “Analog In-memory Computing in FeFET-based 1T1R Array for Edge AI Applications” in 2021 IEEE Symposium on VLSI Technology, pp. 1-2, 2021

7. Review and Reflection

在這次專題研究中，我們首先研讀RRAM和CIM相關的文獻，了解相關的基礎概念。接著與學長討論，並實際參與量測晶片的過程，讓我們更加明白內容。

設計電路時，我們一開始認為只要將結構拼好、設計好各個電晶體的尺寸，就會與想要的結果相去不遠。然而實際上並非如此，在過程中遇到不少問題，像是可讓一部份電路正常運作，但另一塊就會動不了。好在我們仔細排查問題、相互討論，並虛心向學長請教，總算逐一解決問題。我們在專題研究時熟悉了Pads與Virtuoso等軟體，以及學會操作量測電性機台，讓我們不僅學到紙面上的理論，也學會如何操作軟體與儀器。

最後，要感謝葉昭輝教授願意提供機會讓我們做專題，實際完成電路設計的流程；也要感謝劉永智學長不厭其煩地教導我們，並與我們一起找出設計上的問題；還有實驗室的學長姐們，在我們遇到困難時也樂於提供幫助。若沒有這些人的幫助，我們肯定沒辦法順利完成。