國立清華大學 電機工程學系
實作專題研究成果摘要

A Monolithic 3D 8T-SRAM Computing-in-Memory Macro with Write Ability Improvement

利用三維單晶八顆電晶體的靜態隨機存取記憶體進行記憶體內運算與記憶體寫入功能優化

專題領域：系統組

組別：A303
指導教授：張孟凡教授
組員姓名：田任鈞
研究期間：2022年7月1日至2022年4月30日止，共10個月

// Abstract

The growth of artificial neural network calculations has resulted in a significant load on processing data on AI edge devices. However, processors that employ von Neumann architecture waste a lot of time and power transferring data between memory and CPU. To address this issue, Computing-in-Memory (CIM) based on Monolithic 3D (M3D) fabrication has been proposed, which can reduce power consumption, latency, and chip area when performing tasks like Convolution Neural Networks (CNN).

While M3D fabrication can reduce the layout area by stacking transistors, its M3D VIA, which connects upper and lower layers, comes with considerable parasitic capacitance and resistance. This issue makes the writing stage in 8T SRAM harder than planar 8T SRAM.

Therefore, in this project, we took cues from previous papers and designed an area-efficient two-layer Monolithic 3D (M3D) 8T SRAM cell with a Negative Bit-Line (NBL) structure that supports Multiply and Accumulate (MAC) operations. The M3D 8T SRAM layout showed a 45% area reduction compared to a planar 8T SRAM cell. The NBL structure also improved the write success rate of the memory cell from 83.2% to 99.9% in low voltage (VDD = 0.8) operation.

We used CIC018 0.18um technology to fabricate the 45nm technology as described in the reference paper.

Introduction

The block diagram of the M3D 8T SRAM macro is presented in Fig. 1. The macro consists of a 64-row by 64-column SRAM array, 64 columns of NBL, eight 8-to-1 multiplexers, and eight 3-bit Analog-to-Digital Converters (ADCs).

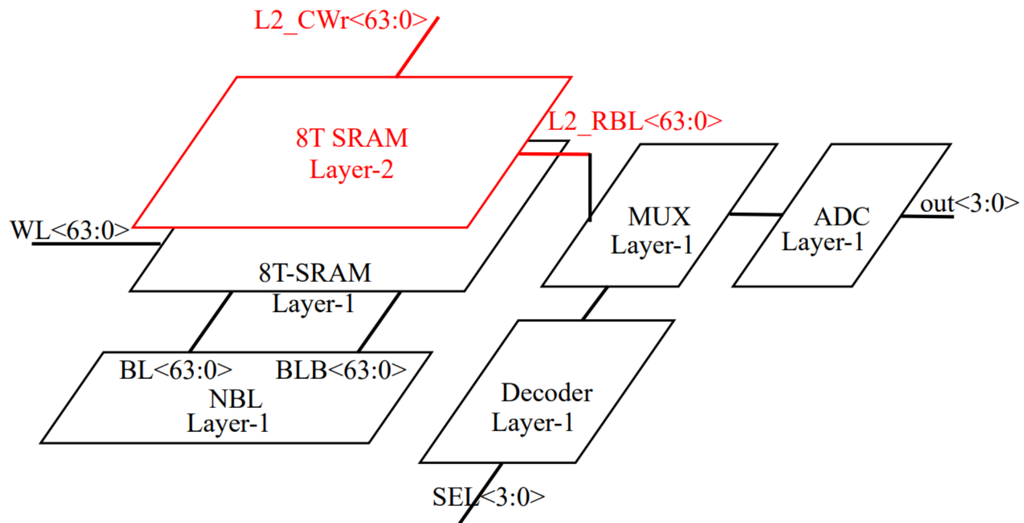

Fig. 1 Block diagram of M3D 8T SRAM

M3D 8T-SRAM Architecture

Srinivasa *et al*. [1] have developed an M3D 8T SRAM cell that exhibits 45% greater area efficiency when compared to the conventional 2D 8T SRAM cell, as shown in Fig. 2. The M3D VIA in the 45nm technology exhibits a parasitic capacitance of 0.5fF and a parasitic resistance of 2Ω. In our project, we have scaled up the M3D cross-section area by 16 times and the M3D VIA length by 4 times to convert the data to the 0.18um technology. Consequently, the corresponding capacitance becomes 8fF, and the resistance becomes 0.5Ω in 0.18um technology.

According to the reference paper, the lower layers are processed using a standard thermal budget, while the upper layers are processed using a low-thermal budget to prevent the layer-1 transistors from melting. The layer-2 transistors exhibit a smaller I_{on} compared to the layer-1 transistors. The width is four times that of the layer-1 transistors to obtain the same I_{on}. In our project, we have disregarded the low I_{on} of the transistors in the upper layer and used regular transistors as bottom layer since we sense the MAC value mostly by the voltage of L2-RBL.

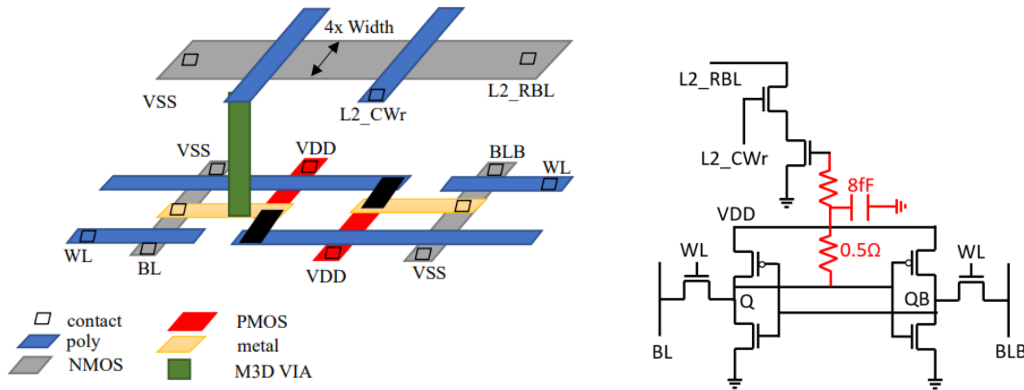

Fig. 2 layout and schematic of M3D 8T SRAM

Negative Bit-line (NBL) Architecture

The M3D SRAM's parasitic capacitance leads to longer write discharging processing, resulting in a higher chance of write operation failure. To address this issue, Mukhopadhyay *et al*. [2] propose a technique that uses a transient negative voltage at the low-going bit-line during writing to reduce write failure. Fig. 3 illustrates the NBL structure and the operation waveform.

During phase 1(PH1), the writing operation is the same as in conventional SRAM, where D discharges BL to low level. At the same time, BIT_EN charges to high level.

During phase 2(PH2), the high-to-low transition of BIT_EN induces an undershoot on BL and BLB due to capacitance coupling of C_{boost}, pulling down the BL voltage to a negative voltage, which improves the writing success rate.

However, the undershoot can also affect BLB. To avoid this, two PMOS transistors with minimum size can be employed. When BL is pulled down, the PMOS transistors will enter saturation mode, charging BLB to a high level.

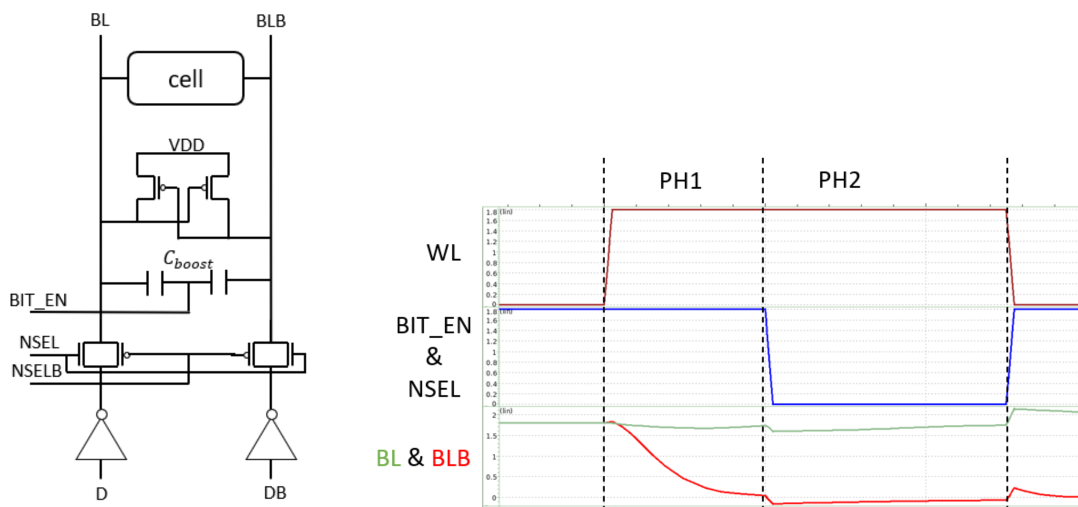

Fig. 3 NBL Structure

3-bit Analog to Digital Converter (ADC) Architecture

We propose using a 3-bit register divider type ADC to accurately distinguish the up to 7 times MAC value. Fig. 4 illustrates the MAC procedure and the structure of the ADC. Each SRAM cell stores weight data, and multiple L2-CWr activations enable MAC operations. The analog MAC value is stored in L2-RBL.

To read out the analog MAC value, we use a 3-bit ADC comprising six Sense Amplifiers (SA) and a Priority Encoder (PE). Two P-type input SAs are employed to receive low-level voltage input, with each SA sensing the difference between the input and the ideal reference voltage. The Encoder then takes the sensing result and encodes the MAC value into a binary number.

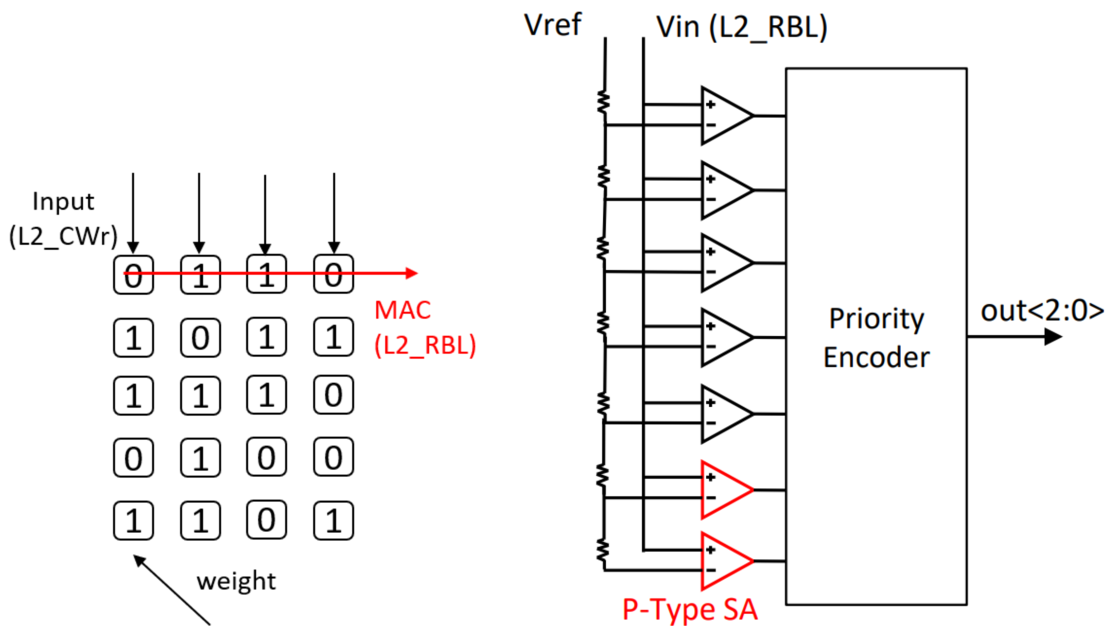

Fig. 4 MAC scheme and ADC structure

Result

In this project, we have successfully implemented a MAC=7 operation. Fig.5 illustrates the digital-to-analog conversion of the 3-bit values stored in L2_RBLs. We then employ an ADC to obtain the MAC value. The voltage difference between 001 and 000 is measured to be 161mV, which is the minimum voltage difference that in each interval level.

We make the SRAM macro operate at a low voltage of 0.8V, and we have conducted 1024 Monte-Carlo simulations. The results of low voltage writing operations are presented in Table 1. It is observed that the NBL structure significantly improves the writing success rate, from 82.3% to 92.6%. As we increase the ratio of C_{boost} to C_{load}, the memory cell's writing ability is enhanced.

All the writing specification results and discussions are summarized in Table 2. The write time and the write energy are increased due to the loading of extra capacitance on the BL and BLB. Read disturb, Hold Static Noise Margin (HSNM), and Read Static Noise Margin (RSNM) remains the same as the standard 6T and 8T SRAM.

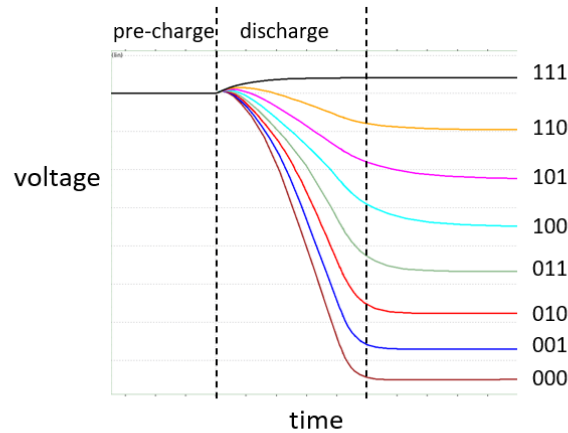

Fig. 5 digital to analog representation of 3-bit values

Table 1
success writing rate of M3D 8T SRAM macro

	write successful rate
normal structure	83.2%
Cboost/Cload=0.125	92.6%
Cboost/Cload=0.25	94.4%
Cboost/Cload=0.5	97.8%
Cboost/Cload=1	99.6%
Cboost/Cload=2	99.9%

Table 2
writing specification of M3D 8T SRAM macro

	Cboost/Cload	write time (ps)	% change	write energy (μW)	% change
std. 6T	-	70.1	0	107.4	0
std. 8T	-	93	32.60%	124.4	15.80%
	1	95.3	35.90%	124.7	16.10%
	2	95.8	36.70%	131.7	22.60%
	4	97	38.40%	147.9	37.70%

Conclusion

M3D fabrication enables stacking multiple layers of transistors, leading to increased area efficiency. In our project, we present a M3D 8T SRAM macro that performs separate write and read operations in different layers and performs a MAC=7 operation. To address the writing bottleneck, we adopt the NBL structure, resulting in a significant improvement in the writing success rate from 83.2% to 99.9%. Moreover, the single cell area is reduced by 45% compared to conventional 2D 8T SRAM cells. The TOPS/W of the macro is 0.0354.

Reference

[1] S. Srinivasa, X. Li, M. -F. Chang, J. Sampson, S. K. Gupta and V. Narayanan, "Compact 3-D-SRAM Memory With Concurrent Row and Column Data Access Capability Using Sequential Monolithic 3-D Integration," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 26, no. 4, pp. 671-683, April 2018, doi: 10.1109/TVLSI.2017.2787562.

[2] S. Mukhopadhyay, R. M. Rao, J. -J. Kim and C. -T. Chuang, "SRAM Write-Ability Improvement With Transient Negative Bit-Line Voltage," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 19, no. 1, pp. 24-32, Jan. 2011, doi: 10.1109/TVLSI.2009.2029114.

[3] V. P. -H. Hu, C. -W. Su, C. -C. Yu, C. -J. Liu and C. -Y. Weng, "Monolithic 3D SRAM Cell with Stacked Two-Dimensional Materials Based FETs at 2nm Node," 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Korea, 2021, pp. 1-5, doi: 10.1109/ISCAS51556.2021.9401245.

[4] C. -J. Jhang, C. -X. Xue, J. -M. Hung, F. -C. Chang and M. -F. Chang, "Challenges and Trends of SRAM-Based Computing-In-Memory for AI Edge Devices," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 68, no. 5, pp. 1773-1786, May 2021, doi: 10.1109/TCSI.2021.3064189.

心得感想

在這一年的專題訓練中，我們受到教授和學長極為紮實的指導，從上學期的背景調查、確定專題研究主題、論文調查、寒假作業、論文實作，一直到開發新的想法，每一個階段都使我在記憶體領域的知識量得到了提升。從一開始對論文中的專有名詞一無所知，到後來能夠迅速理解論文內容中的電路重點，更讓我對記憶體這個領域充滿了興趣。

在本次專題報告中，我運用了整年所學的知識，從搜尋到的論文中得到了啟發，改善了原本的記憶體電路寫入功能，並實現了基礎的乘加運算。在這個過程中，雖然曾一度懷疑自己是否能在專題競賽截止前完成電路，但隨著自己投入的時間越來越多，我對自己所建構的電路越來越熟悉，也逐漸找到了解決問題的方法。最終，當我看到電路運作正常，產生出與預期一致的波形圖時，我感到了極大的成就感。

最後，我想特別感謝兩位博班學長：洪哲民學長以及溫戴豪學長。他們總是非常及時地回應我的問題，並提供一些電路解法與優化方式。在這一年的專題研究中，我們收穫良多。張孟凡教授也常常在會議中提醒我們除了要考慮到電路本身模擬的情況，還要充分考慮晶片量產時的些微偏差變化。因為即使一點點的光罩誤差，也有可能導致晶片在測試時報廢。這些經驗讓我認識到自己在電路專業領域中的知識量需要再繼續提升。期許在實驗室所學到的內容能對未來的研究和讀書計畫有一定的幫助。