A 65nm In-Memory Computing 8T-SRAM with Digital Sparsity Optimization Architecture

指導教授：張孟凡　組員：俞芷妍、張匀馨　組別：A407

Abstract

To speed up the computing speed, Computing in Memory (CIM) can not only store data but also perform operations in the memory, reducing data transmission times and delays, thereby improving computing efficiency and saving energy.

In this project, we implement and modify both Analog CIM and Digital CIM, and further combine their advantages by implementing a new CIM architecture.

Circuit Architecture - Comparison

1. Analog 8T SRAM-CIM with Charge Sharing

- Input is realized with the number of RWL pulse. 4-b weight is realized using charge sharing with binary-weighted capacitors. Compensation caps are isolated after RBL sampling.
- When input pulses are present on RWL and SRAM data is 1, RBL voltage drops. After charge sharing, the voltage across the computation caps is proportional to MAC value.
- Tuning switch size and placement (RC balance), optimize linearity between MAC value & final voltage.

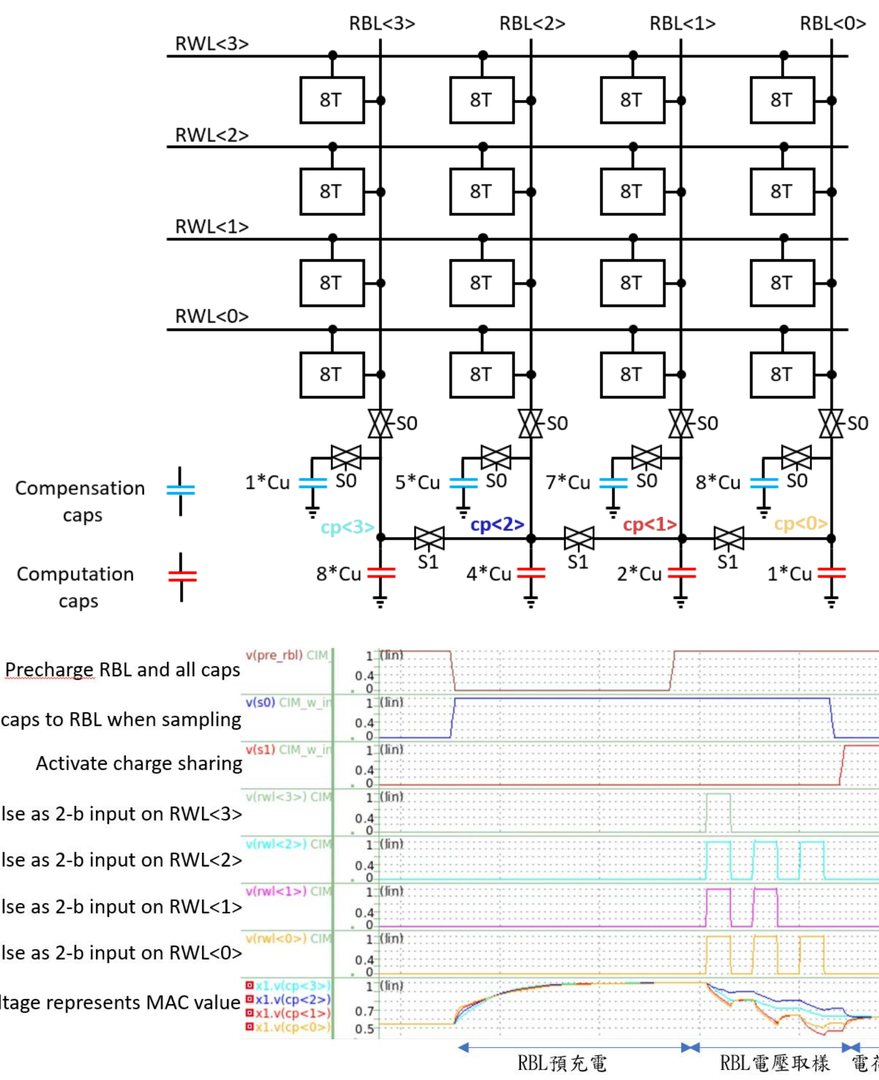


Fig.1 8T-SRAM CIM Schematics and Operation waveform

2. Digital Bitline-Transpose CIM (BLT-CIM)

- Introduction:
  a) CIM Controller with 3b controlling signals decides the working state is weight writing or input feeding.
  b) 16 × 256 SRAM-CIM Array is designed with 6T SRAM and a 4T NOR gate is connected to Accumulator for computation.
  c) Macro Accumulator helps performing shift-accumulation for the bit-serial inputs.
- Modification
  a) Sparsity Optimization will skip the current computation when detecting zero input.

3. Simulation Results

	Total cycle number	Operations	Power (W)	Energy Efficiency (TOPs/W)	Accuracy
Analog CIM	4	$(4+3)\times 3$	4.56×10^{-4}	1.15	Lower
Digital CIM	13	$(4+3)\times 3$	1.54×10^{-4}	1.05	Higher

Table.1 Comparison Between Analog CIM and Digital CIM
(Modify Digital's Specification to Analog's Specification)

Circuit Architecture - Combination

We optimize our circuit performance by combining and adjusting architectures from two works. Input data are first rearranged in Digital domain, thus reducing the number of computation that need to be conducted in the Analog CIM. Bellows are the component we add.

- 8T SRAM Array with RWL Pulse Generator
  a) Converts the clock signal into three distinct pulse signal using delay cells.
  b) Input data as the enable signal to each pulse signal, producing RWL pulses that can be used in Analog CIM.

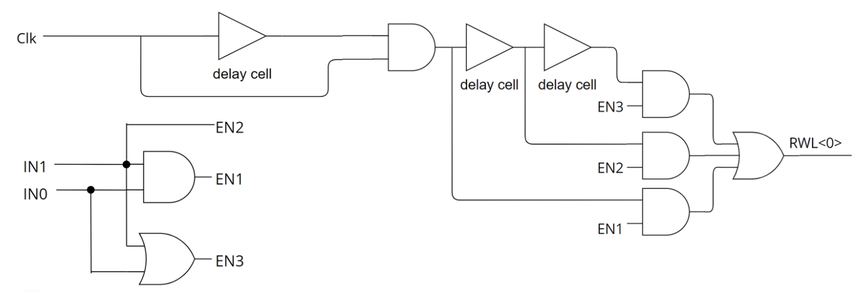


Fig. 2 RWL Pulse Generator

- Sparsity Optimization
  a) Detects "All zero input pattern" and skip this computation during the Analog CIM.

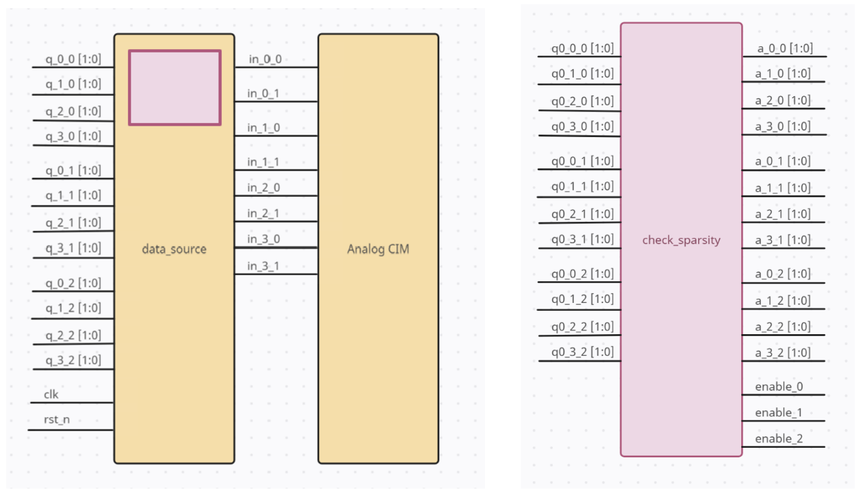


Fig. 3 Block Diagram of Sparsity Optimization & Analog 8T-SRAM CIM

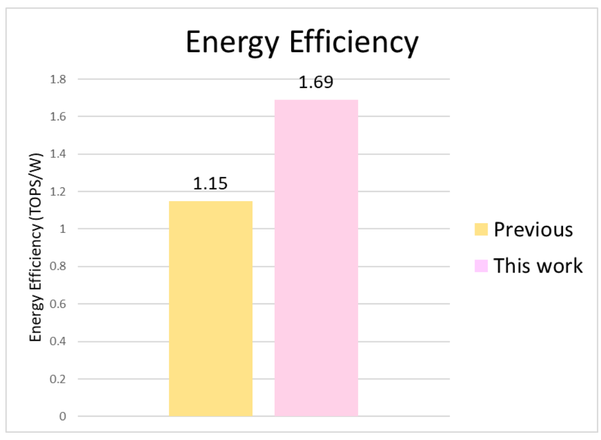
Experimental Result & Conclusion

Considering to combine the better Energy Efficiency advantage of Analog CIM with the input-first processing mechanism of Digital CIM, we merged the modified Analog 8T SRAM CIM and Sparsity Optimization structure with following result:

	Total cycle number	Operations	Power (W)	Energy Efficiency (TOPs/W)
Previous	4	$(4+3)\times 3$	4.56×10^{-4}	1.15
This work	3	$(4+3)\times 3$	4.14×10^{-4}	1.69

Table.2 Comparison Between With and Without Sparsity Optimization

1. Comparing Analog CIM and Digital CIM, we observed that the Analog 8T SRAM CIM has better energy efficiency, and through improvements it also has better linear performance and better accuracy. On the other hand, the Digital BLT-CIM has great accuracy and spatial mapping capabilities. In addition, Sparsity Optimization can also effectively speed up the calculation speed.
2. Combining the modified Analog 8T SRAM CIM and Sparsity Optimization structures, it can be observed that the CIM operation speed is accelerated, and the Energy Efficiency is improved by 46%.



Energy Efficiency is improved by 46%

Fig.4 Comparison of Energy Efficiency before and after Sparsity optimization

Reference:

[1] Qing Dong, Mahmut E. Sinangil, Burak Erbagci, Dar Sun, Win-San Khwa, Hung-Jen Liao, Yih Wang, Jonathan Chang, "A 351TOPS/W and 372.4GOPS Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine-Learning Applications", ISSCC2020

[2] F. Tu et al., "A 28nm 15.59μJ/Token Full-Digital Bitline-Transpose CIM-Based Sparse Transformer Accelerator with Pipeline/Parallel Reconfigurable Modes," 2022 IEEE International Solid-State Circuits Conference (ISSCC)

[3] Mahmut E. Sinangil, Burak Erbagci, Rawan Naous, Kerem Akarvardar, Dar Sun, Win-San Khwa, Hung-Jen Liao, Yih Wang, Jonathan Chang, "A 7-nm Compute-in-Memory SRAM Macro Supporting Multi-Bit Input, Weight and Output and Achieving 351 TOPS/W and 372.4 GOPS", JSSC, VOL. 56, NO. 1, pp. 188-198, JANUARY 2021.