A 65nm 6T-SRAM Ping-Pong Weight Updating Computing-in-Memory Macro with Adaptive Precision ADC for AI edge chips

65奈米製程同步權重更新及計算與自適性精度類比數位轉換器之記憶體內運算架構

指導教授：張孟凡　　組員：李昀達　　組別：A399

Abstract

On-device training face challenges with time and power for data transfer under the conventional architecture. Computing-in-Memory (CIM) architecture minimizes data movement, enabling parallel multiply-and-accumulate (MAC) operations directly within the memory. Deploying neural network models on CIM processors face challenges from insufficient weight storage to voltage dividing CIM accumulation affecting MAC values' linearity and margin. The ping-pong CIM scheme [1] allows simultaneous CIM and write operations, while an adaptive bitline header paired with an adaptive precision ADC balances accuracy and performance.

Research Methodology

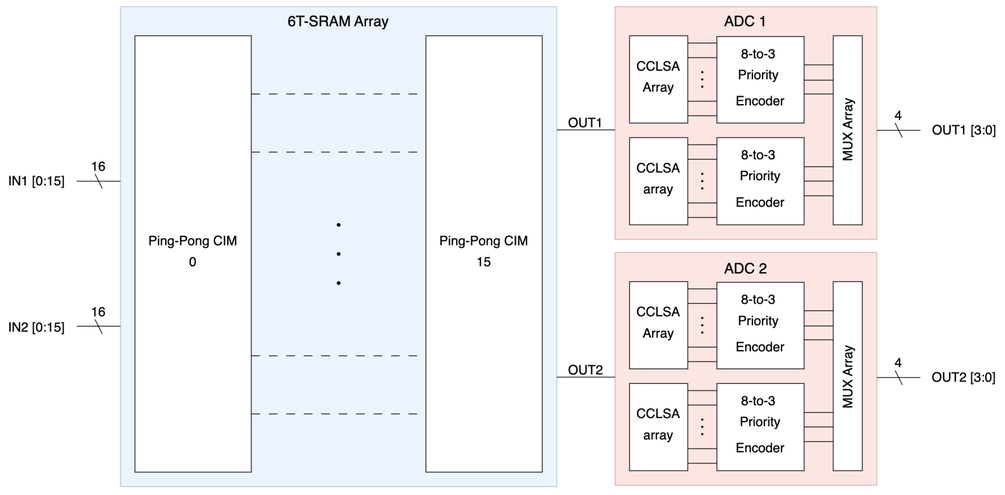

Fig. 1 Overall architecture of ping-pong CIM with adaptive precision ADC.

Fig. 1 represents the overall architecture of this design, while Fig. 2 shows the circuit and operation of the Ping-Pong CIM bank. Two homogeneous regions A and B bitwise memory units (BMU) support simultaneous weight update and CIM operations for 1-bit weight × 2-bit input in the local computing cell (LCC).

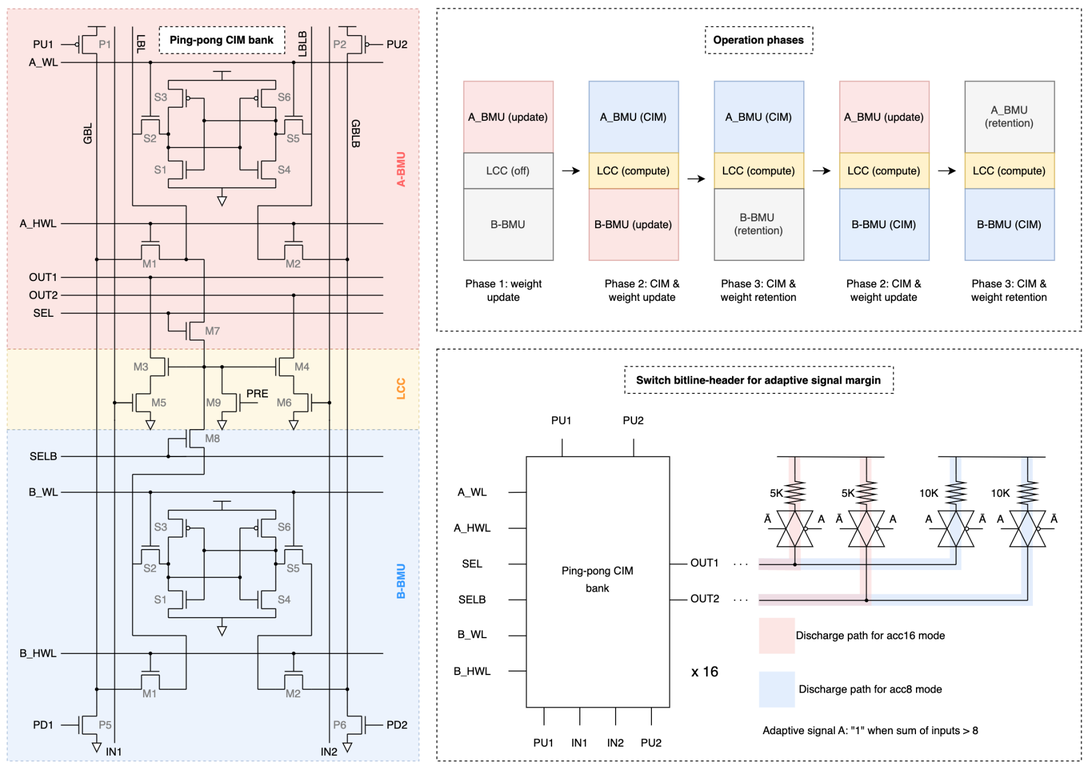

Fig. 2. Ping-pong CIM circuit and operation phases.

Operations are split into three phases: In phase 1, weight data are written into A-BMU while B-BMU stays in retention. In phase 2, A-BMU executes CIM operations with weight, B-BMU writes the next section of weight data. In phase 3, if the CIM operation is still running in A-BMU, B-BMU goes into retention and waits for the next cycle of operation. Phase 2 and phase 3 alternate between A-BMU and B-BMU for CIM operations to complete MAC computation.

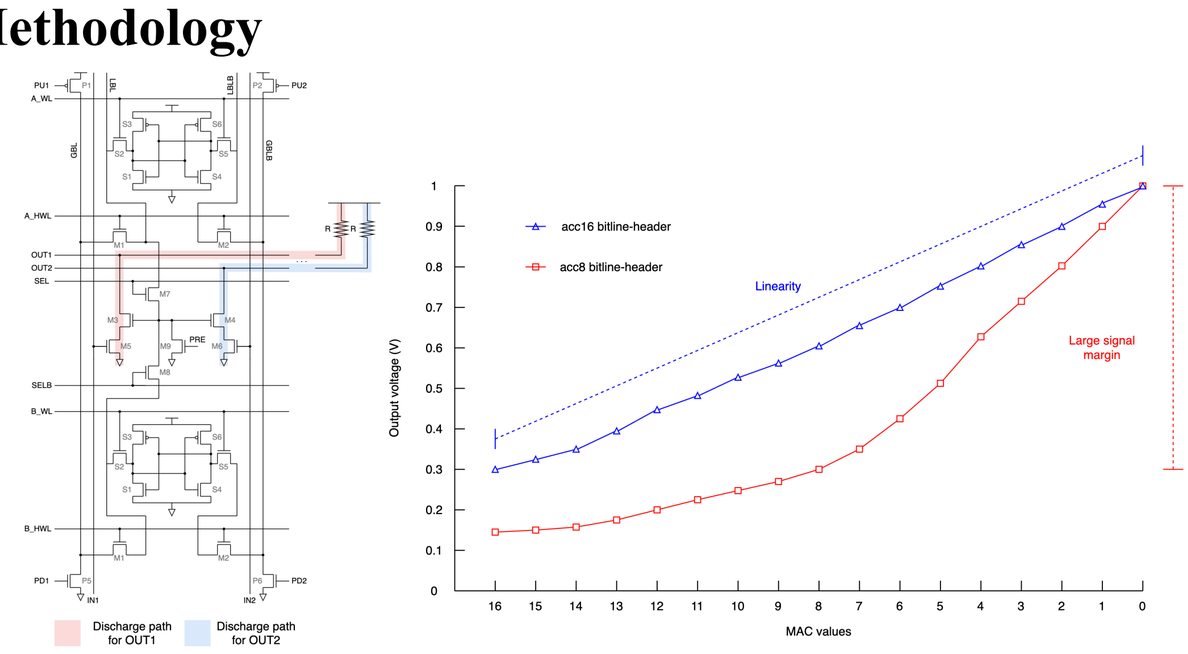

Fig. 3. Discharge path and MAC value comparison between acc8 and acc16 mode.

The voltage dividing method is used for row accumulation. The pull-down path with internal resistance R_0 is used for voltage dividing with bitline header resistance R. The total resistance from ping-pong CIM cells is inversely proportional to multiplication results equaling to "1". In acc8 mode, the accumulation is at most 8, thus we use 10k ohms resistance to increase the signal margin for MAC values 0~8. In acc16 mode, accumulation results range from 0~16, resistance is set to 5k ohms to split the full range voltage equally for correct MAC values, preserving linearity and margin as presented in Fig. 3.

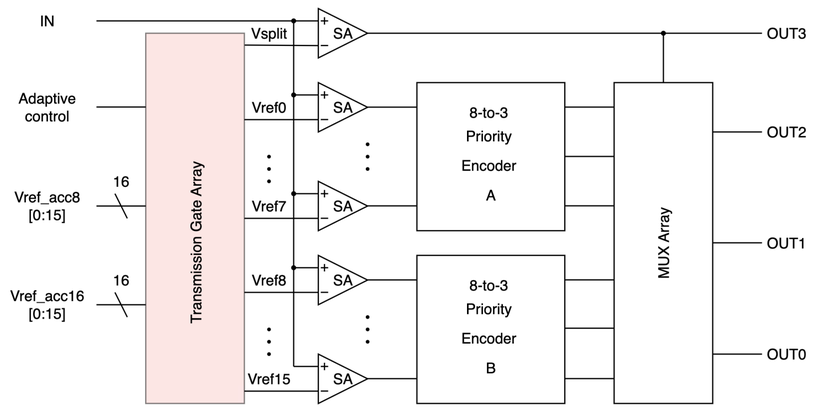

Fig. 4. Adaptive precision ADC structure.

The adaptive ADC compares accumulation voltage output with reference voltages using current-controlled latch sense amplifiers [3]. Priority encoders are used for 3-bit digital output, shown in Fig. 4. MUX array selects between the output of two priority encoders according to the OUT3 signal from an additional sense amplifier. Transmission gates switch the sense amplifiers' reference voltages between acc8 and acc16 modes for 3-bit or 4-bit sensing.

Experimental Results and Conclusion

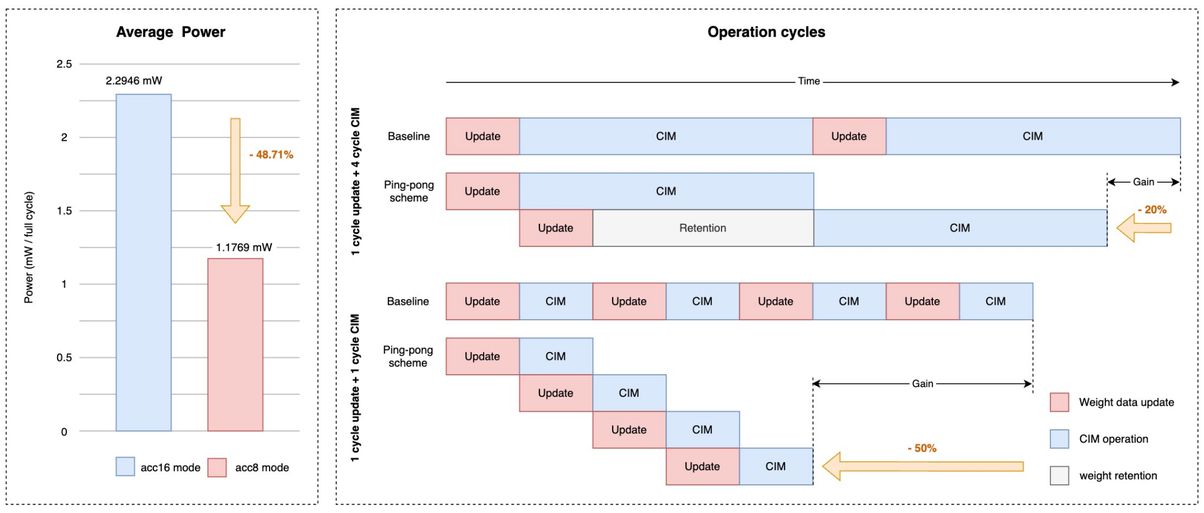

Fig. 5. Power and operation cycle comparison.

In acc8 mode, deactivating one priority encoder and 8 sense amplifiers reduces power consumption by 48.71%, seen in Fig. 5. This work shows potential for higher power savings in datasets with input distributions clustering closer to 0. Ping-pong CIM design achieves a reduction in operation time by 20%, under 4 cycles of MAC operation per CIM phase and 1 cycle weight updating. For a single cycle of MAC per phase, a 50% reduction could be achieved, depicted in Fig. 5. This work shows potential in datasets with high weight value variations for efficient time savings compared to conventional CIM structure.

Reference:

[1] J. Yue, X. Feng, et al., "A 2.75-to-75.9TOPS/W Computing-in-Memory NN Processor Supporting Set-Associate Block-Wise Zero Skipping and Ping-Pong CIM with Simultaneous Computation and Weight Updating" ISSCC 2021

[2] J. Yue, Y. Liu, et al., "An Energy-Efficient Computing-in-Memory NN Processor With Set-Associate Blockwise Sparsity and Ping-Pong Weight Update" JSSC 2023

[3] J. Su, X. Si, et al., "Two-Way Transpose Multibit 6T SRAM Computing-in-Memory Macro for Inference-Training AI Edge Chips" JSSC 2022

[4] T. Kobayashi, K. Nogami, et al., "A current-controlled latch sense amplifier and a static power-saving input buffer for low-power architecture" JSSC 1993