電機資訊學院 2024 作 BRAIN PLUS HAND 實作專題競賽

A 180nm 6T-SRAM Computation-in-Memory Macro with high area efficiency SAR-ADC reusing parasitic capacitance of SRAM-cells for AI Edge chip

使用180奈米製程高面積利用效率逐次逼近類比數位轉換器之靜態隨機存取記憶體內運算電路架構

組別：EECS002　組長：王怡萱　組員：林彥華

I. Abstract

In Machine Learning, training a model requires tons of convolution operation, the previous computer architecture takes lots of time and power in transporting data between memory and CPU. To increase energy efficiency, CIM(Computation in Memory) architecture do most of the basic MAC(Multiply and Accumulate) operations inside the memory, which not only increases the computation speed but also reduces power consumption.

Our project construct a SRAM-CIM architecture which combine the originally separated 6T-SRAM array and SAR-ADC to reuse parasitic capacitance of SRAM-cells. We achieve the reduction of 73.7% on area usage(w/o SRAM cells), and also 12% improvement on TOPS/W, successfully increase the energy efficiency per chip area.

SRAM-CIM Macro

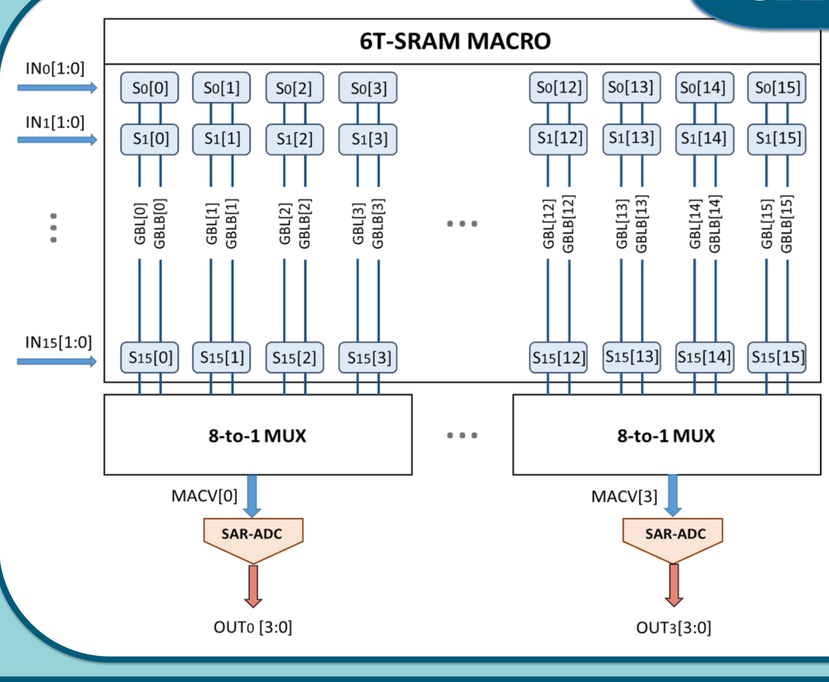


Fig. 1 shows CIM macro, and Fig.2 shows waveform.
(1) phase 1: multiply INxW in sub-array, IN is the input signal, and W is stored in SRAM.
(2) phase 2: accumulate done by charge sharing between sub-arrays.
(3) phase 3: sensing Analog result of MAC is sent to SAR-ADC for the ADC conversion.

II. Proposed Scheme

Fig. 3 Structure of new work. 1 unit composed of 4 column and 1 SAR-ADC.

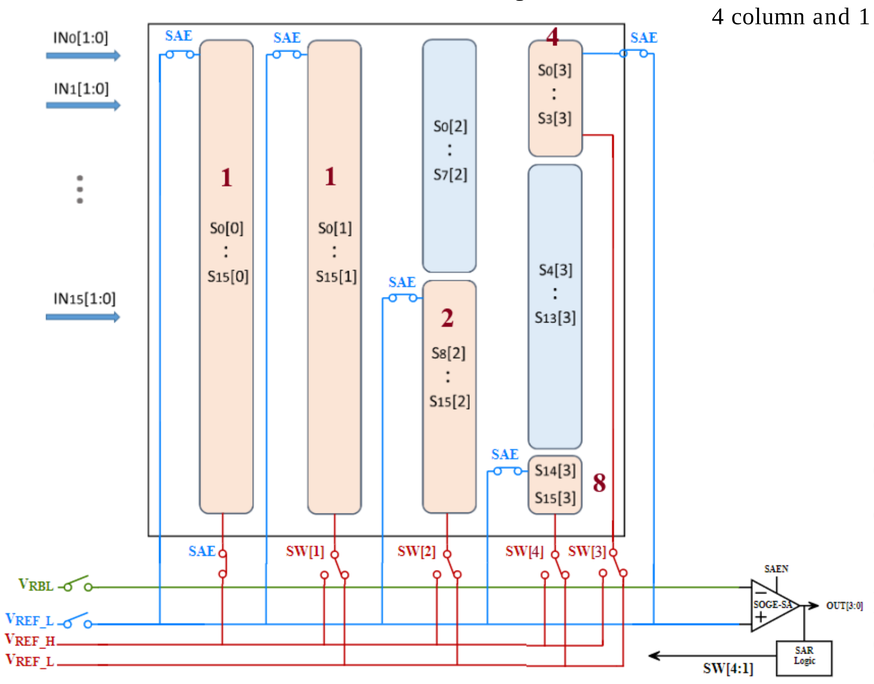
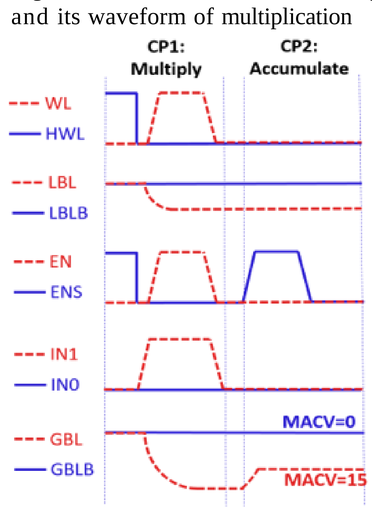


Fig. 2 The structure of the sub-array and its waveform of multiplication



Multiply		Result
1b-IN	1b-W (LBL)	IN x W
0(1.800V)	1(1.800V)	0(1.480V)
1(0.520V)	1(1.800V)	1(0.397V)
Don't care	0(0.000V)	0(1.480V)

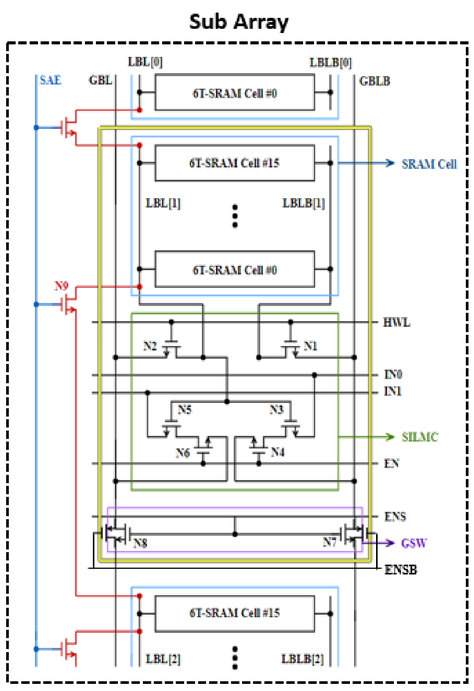


Fig.2 and Fig.3 shows the scheme we proposed. Our goal is to further utilized the parasitic capacitance of SRAM cells in ADC evaluation phase, and make make improvements on (1) Area efficiency (2) precision of successive approximation.

We implement the idea by (1) adding LBL(Local BitLine)-connecting switch(N9) into sub array , N9 can connect the separated LBLs when SAE=1. (2) The orange blocks are the sub array groups with LBLs connected, the number of connected sub-array is 16 : 16 : 8 : 4 : 2, because we mainly use gate capacitance of N9s as capacitor unit, the number of cap. in series will corresponding to capacitance ratio 1 : 1 : 2 : 4 : 8, then the 5 groups can be used as the capacitor array in SAR-ADC. The blue circuit with switches is added in order to connect the capacitor in one common end which is sent into sense amplifier of SAR-ADC. Instead of using purely NMOS series as capacitor array, connect it by LBLs with small SRAM parasitic capacitance help calibrate the capacitance ratio to better match the ideal one, and this calibration is made using no additional resource.

In sub-Array structure, We also replace NMOS with transmission gate at GSW(N7/N8) to improve the offset of MACV.

III. Result and Conclusion

We construct schematics and run the pre- simulation to test the performance(Fig.4), and draw the layout to compare the area. You can see that the area and access time drop but power increases as a trade-off, while it's acceptable because TOPS/W for new work increase by 12%.

In Fig.5, you can see that the voltage offset between each MACV is larger. In Fig. 6, shows that we achieve a better match of ideal cap ratio in SAR-ADC, provide a better approximation precision.

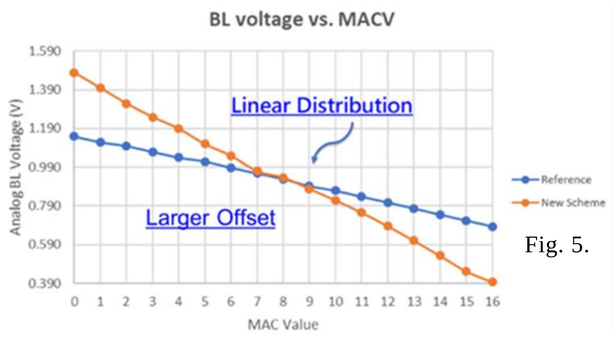
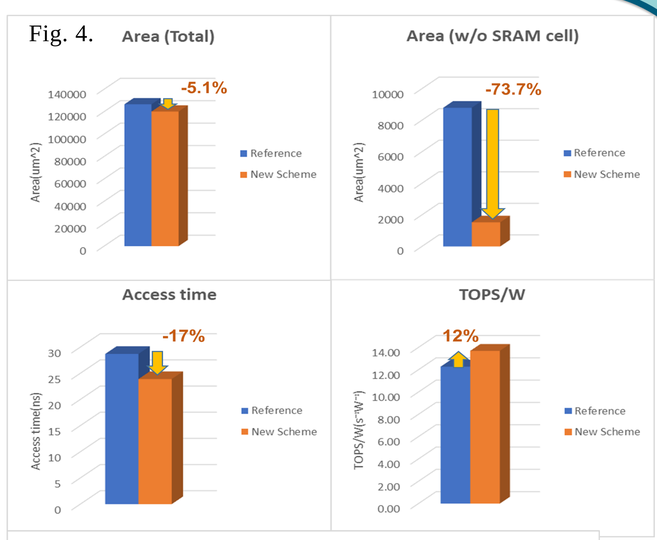


Fig. 5.

Fig. 4.



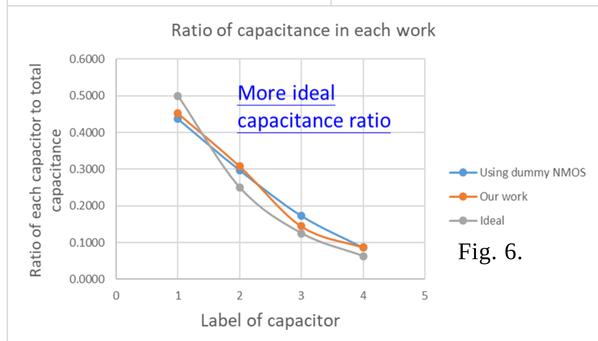


Fig. 6.

Reference:

[1] J. -W. Su et al., "16.3 A 28nm 384kb 6T-SRAM Computation-in-Memory Macro with 8b Precision for AI Edge Chips," 2021 ISSCC.

[2] J. -W. Su et al., "15.2 A 28nm 64Kb Inference-Training Two-Way Transpose Multibit 6T SRAM Compute-in-Memory Macro for AI Edge Chips," 2020 ISSCC