# A High Accuracy Low Power Convolution Operator with 12T SRAM for CNN

Tae Seob Oh
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, South Korea
souliu@skku.edu

YoungGun Pu
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, South Korea
hara1015@skku.edu

Kang-Yoon Lee
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, South Korea
klee@skku.edu

***Abstract*—** To get high accuracy, various weight should be stored in memory. For this, this paper presents tri-state weight static random access memory(SRAM). 12T SRAM is a form of power gating on the conventional 10T SRAM. By using power gating, the inverter can be turned off. The new weight (0) can be stored in 12T SRAM when inverter is turned off. The operator fabricated in a 0.18-µm CMOS process dissipates 172.3µW with the supply of 1.8V while convolution. Even without sizing, the writing margin is better than the conventional SRAM and the accuracy is improved by 23.2%.

***Keywords—12T SRAM, convolutional neural networks (CNN), Convolution, processing in memory (PIM), Tri-state weight***

## I. Introduction

Artificial intelligence(AI) has changed most of us. Through these features, we effectively block spam [1] and early detect pendemic disease such as COVID-19 [2]. As AI is applied more and more in our lives, the computation of AI is moving from the server, the center of network, to devices in the edge of network, which called “Edge AI”. The devices in edge of the network refers to all Internet-of-Things(IoT), including smartphones and TV etc. There is many reason for this phenomenon. First, the amount of computation for AI is too large that it is difficult to perform only with server. Second, as the input data and the result of the computation become more complex, the amount of data that needs to communicate with the sever increases. Third, “Edge AI” makes immediate judgment without communication with server.

To apply AI to the edge of network, low power and low area are essential. Most IoT devices are small and have a small amount of battery. However, most AI should work almost all time and have to be judged immediately. The proposed structure is optimized for “Edge AI”. This structure can reduce the area because the write margin is good without SRAM sizing, and the leakage current is reduced through power gating, and the weight is saved with the SRAM turned off to reduce power consumption. In addition, by storing the weights of the three states, it improves the accuracy of convolution, an important operation in CNN.

This paper is organized as follows. Section II explains the characteristic of convolution operator. Section III presents the architecture of convolution operator. Section IV explains the simulation processes. Section V shows the result of simulations. the conclusion and the discussion will be in Section VI

## II. The characteristic of convolution operator

### *A. 12T SRAM*

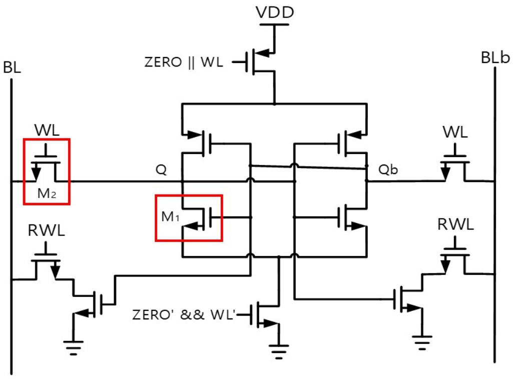


Fig. 1. Schematic of 12T SRAM

Fig. 1. shows the schematic of 12T SRAM. It is based on conventional 10T SRAM [3]. The additional NMOS and PMOS are placed between the power supply and inverter. These two MOSFET controlled with ‘Zero’ signal and ‘WL’ signal. ‘Zero’ signal becomes high when the weight that will be stored in SRAM is zero. ‘WL’ signal becomes high when the writing operation is started. ‘Zero`’ ‘WL`’ is the opposite of ‘Zero’ and ‘WL’ respectively. The advantages of 12T SRAM is low power, more weight, and the small area. By using power gating method, 12T SRAM can reduce leakage current. The additional MOSFET are high threshold voltage(HVT) MOSFET. So the leakage current of 12T SRAM is much smaller than conventional 10T SRAM. The inverter in SRAM can be turned on and off due to the MOSFET between the inverter and the power supply. The new weight (0) is stored when inverter is turned off. When the inverter is turned off, the inside of SRAM becomes floating. Then weight of 0 can be stored. The size of $M_1$ should be larger than $M_2$ because of writing stability. It can be ignored by turning off inverter while

978-1-7281-6476-2/21/$31.00 ©2021 IEEE

writing. Also 12T SRAM includes advantage of 10T SRAM such as reading SNM.

### B. *Writing operation of 12T SRAM*

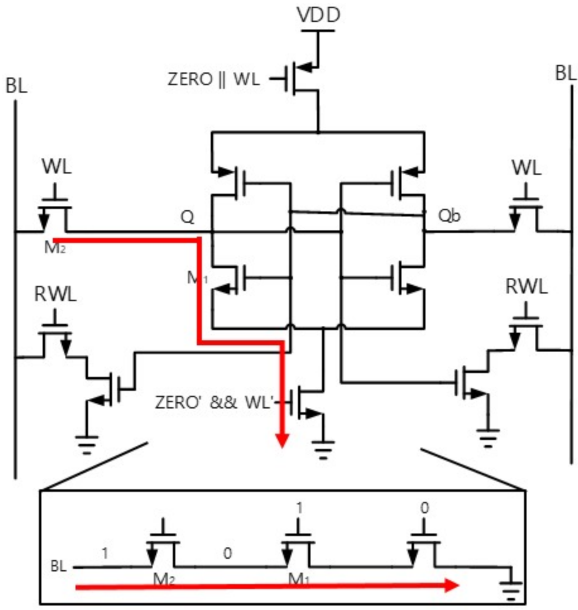


Fig. 2. Writing operation of 12T SRAM

Fig. 2. shows the writing operation for 0 to 1. For the conventional 10T SRAM, M1 is directly connected with ground. So the sizing issue occur. But, for 12T SRAM, M1 is not connected with ground while writing operation. For this reason, 12T SRAM can change saved data more efficiently, and the area of 12T SRAM will be smaller than conventional.

### C. *Reading operation of 12T SRAM*

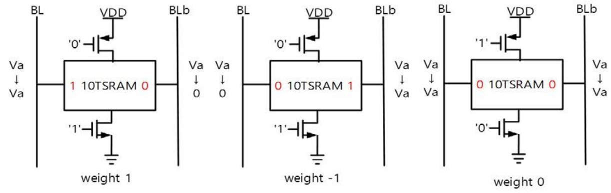


Fig. 3. The reading operation of 12T SRAM

The reading operation of 12T SRAM is similar with 10T SRAM. For weight 1, SRAM is turned on. So BL will not be discharged and BLb will be discharged. For weight -1, SRAM also turned on. So BL will be discharged and BLb will not be discharged. When weight 0 is stored, BL and BLb (bit-line-bar) will not be discharged. This phenomenon is crucial in common circuit because BL and BLb store different data from saved data in SRAM after reading operation. However, for this operator, this problem is not crucial because of the method of convolution.

### D. *Convolution [3]*

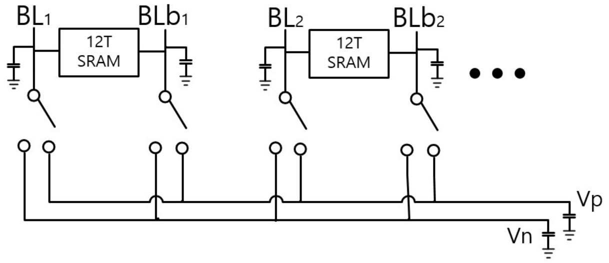


Fig. 4. The circuit of convolution

The circuit of convolution is implemented in [3]. Convolution is multiplication and addition. The multiplication is implemented with SRAM reading operation. And the addition is implemented with charge conservation law. The switch between summing node(Vn, Vp) and each node(BLB1, BL1) is controlled by 'SIGN' signal which represent sign of input value. When negative input comes in, BL is connected to Vn, and BLb is connected to Vp. When positive input comes in, BL is connected to Vp, and BLb is connected to Vn. Voltage of each BL and BLb are averaged since the capacitance of each node is same. The result of convolution is the voltage difference between Vp and Vn. Because of this method, even though BL and BLb stored different data with ideal result after reading operation with weight 0, The charge of BL and BLb is same. i.e. The same voltage is added to Vp and Vn.

## III. The Architecture of Convolution Operator

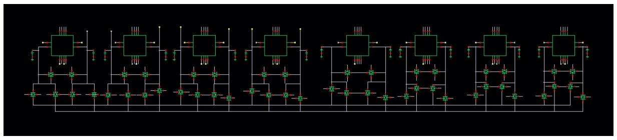


Fig. 5. The schematic of convolution operator

Fig. 5. is the implementation of convolution operator by cadence. It consists of the 1x8 12T SRAM array. Each switch is implemented with T-Gate. This operator stored 8 weights and get 8 input as analog voltage signal. The output is the analog voltage in Vp and Vn nodes. A 100fF capacitor was attached to each BL and BLb. Every MOSFET in T-Gate and 12T SRAM are 2V MOSFET and the HVT MOSFETs (additional MOSFET in 12T SRAM) are implemented as 3V MOSFET.

## IV. Simulation Process

There are two simulations. First, test the convolution operation with specific input for comparing the accuracy and power consumption with conventional operator. The other one is to test the writing margin. The write margin was considered as how little difference voltage could be properly stored. The simulation was run with spectre.

### A. Convolution operation

TABLE I. INPUT DATA AND WEIGHTS FOR SIMULATION

| | Convolutional Operator SRAM cell | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Cell 1** | **Cell 2** | **Cell 3** | **Cell 4** | **Cell 5** | **Cell 6** | **Cell 7** | **Cell 8** |
| Input | 1 | 5 | -4 | 3 | 9 | -8 | 10 | -1 |
| Weight (Real) | 1 | 0.3 | -0.8 | 0.6 | 0.2 | 0.1 | 0.8 | -1 |
| Weight (10T) | 1 | 1 | -1 | 1 | 1 | 1 | 1 | -1 |
| Weight (12T) | 1 | 0 | -1 | 1 | 0 | 0 | 1 | -1 |

In the simulation, it was assumed that the input data already passed through the DAC and converted into analog voltage, and the difference voltage between the Vp and Vn would be converted into digital data through the ADC.

Input data and weight for simulation is shown as TABLE 1. The weight stored in 10T SRAM and 12T SRAM were determined by rounding off the real weight. The ideal output is 17.5. The ideal 10T SRAM output is 22, and the ideal 12T SRAM output is 19. The reference voltage for input is 100mV.

### B. Writing margin

The simulation of writing margin was tested with one cell. For 10T SRAM, $M_2$ MOSFET made larger than $M_1$ MOSFET considering the sizing of SRAM. For 12T SRAM $M_2$ MOSFET made same as $M_1$ MOSFET. While sweeping input voltage, check the writing operation for both SRAM cell. The data stored in SRAM are Q: 1, Qb: 0. Then, change the data to Q: 0, Qb: 1 with sweeping voltage in BL and BLb.

## V. Simulation Result

### A. Convolution operation

#### 1) 10T Convolution operator

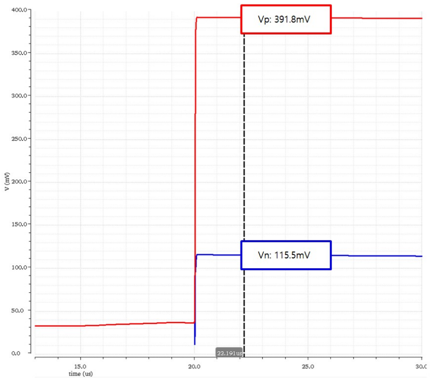


Fig. 6. Simulation result of 10T convolution operator

The result of convolution is shown as Figure 6. The voltage difference between Vp and Vn is 276.3mV. Since the reference voltage of input is 100mV, the reference voltage of output is 12.5mV. The output value of the 10T convolution is 22.1.

#### 2) 12T Convolution operator

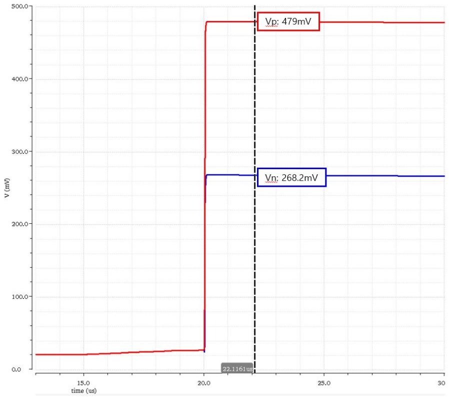


Fig. 7. Simulation result of 12T convolution operator

For 12T convolution, the voltage difference between Vp and Vn is 210.8mV. The output value of the 12T convolution is 16.9.

TABLE II. SIMULATION RESULT OF CONVOLUTION

| | Measured output | Target output | Target error | Ideal output | Real error |
|---|---|---|---|---|---|
| Conventional (10T) [3] | 22.1 | 22 | 0.5% | 17.5 | 26.2% |
| Proposed (12T) | 16.9 | 19 | 11% | 17.5 | 3.4% |

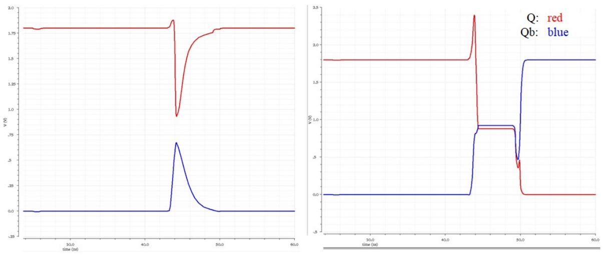
The error between measured output and target output of 12T SRAM is much higher than 10T SRAM. However, actual accuracy of 12T SRAM is much higher than 10T SRAM.

TABLE III. POWER CONSUMPTION

| | Power consumption |
|---|---|
| Conventional (10T) [3] | 275.7μW |
| Proposed (12T) | 172.3μW |

The proposed operator had 103.4μW (37.5%) lower power consumption than the conventional operator during convolution.

### B. Writing margin



(a) 10T SRAM (b) 12T SRAM

Fig. 8. Simulation result of writing margin

The simulation condition is 0.85mV in BL and 0.95mV in BLb. Although the size of $M_1$ and $M_2$ are same in 12T SRAM, the stored value of 10T SRAM did not change under simulation conditions, but the value of 12T SRAM did. This shows that the writing method of 12T SRAM is more stable than the conventional 10T SRAM.

## VI. Conclusion

This paper presents the high accuracy low power convolution operator for CNN. We demonstrated the low power and high accuracy with simple convolution. Accuracy of proposed convolution operator is 23.2% higher than conventional operator(10T). And the power consumption is also 37.5% lower. The actual accuracy has improved, but the measured output is far from the target value. It can be a big problem, so we need to do more research for this result.

The weight distribution of CNN is around a zero-value peak. So, Therefore, it is expected that more power consumption can be reduced when the proposed architecture is grafted onto an actual CNN. This result shows that the proposed convolution operator is suitable as convolution operator in CNN to be mounted on edge devices such as IoT because of low power consumption and high accuracy.

## Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2019-0-00421, Artificial Intelligence Graduate School Program(Sungkyunkwan University))

## References

[1] S. Annareddy and S. Tammina, "A Comparative Study of Deep Learning Methods for Spam Detection," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2019, pp. 66-72, doi: 10.1109/I-SMAC47947.2019.9032627.

[2] M. Qjidaa *et al*., "Early detection of COVID19 by deep learning transfer Model for populations in isolated rural areas," *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, Fez, Morocco, 2020, pp. 1-5, doi: 10.1109/ISCV49265.2020.9204099.

[3] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An Energy-Efficient SRAM With In-Memory Dot-Product Computation for Low-Power Convolutional Neural Networks," in IEEE Journal of Solid-State Circuits, vol. 54, no. 1, pp. 217-230, Jan. 2019, doi: 10.1109/JSSC.2018.2880918.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.