LETTER

6T-8T Hybrid SRAM for Lower-Power Neural-Network Processing by Lowering Operating Voltage

Ji WU[†a)], **Ruoxi YU**[†], ***Nonmembers*, *and*** **Kazuteru NAMBA**[††], ***Member***

SUMMARY This letter introduces an innovation for the heterogeneous storage architecture of AI chips, specifically focusing on the integration of six transistors(6T) and eight transistors(8T) hybrid SRAM. Traditional approaches to reducing SRAM power consumption typically involve lowering the operating voltage, a method that often substantially diminishes the recognition rate of neural networks. However, the innovative design detailed in this letter amalgamates the strengths of both SRAM types. It operates at a voltage lower than conventional SRAM, thereby significantly reducing the power consumption in neural networks without compromising performance.

key words: *neural network, SRAM, bit error rate (BER), low power consumption, AI chips*

1. Introduction

In recent years, the potent learning and portability of deep learning have led to the deployment of deep learning applications in numerous cloud-centric edge devices and IoT devices [1]. Particularly, convolutional neural networks (CNNs), a flagship of deep learning, have significantly improved accuracy in various classification domains such as voice recognition and image classification [2]. However, the deployment of deep learning algorithms in embedded systems and mobile devices is restricted due to high memory power consumption, driven by network storage requirements and intense memory access. Voltage scaling can significantly reduce power consumption, but there is a limit imposed by the stability of read and write operations in SRAM. In traditional SRAM, as the supply voltage decreases, the Bit Error Rate (BER) increases. To address this, a transition from conventional 6T cells to more reliable structures like 8T cells [3], further voltage scaling, transitioning from traditional CMOS technology to FINFET for enhanced reliability [4], or protecting certain critical parts from voltage scaling are some of the proposed methods.

Recent advancements in edge computing have integrated neural network computation into various devices, prompting interest in novel SRAM technologies and their bit error rates (BER) under low voltage, such as the 8T SRAM architecture [3]. Given the inherent fault tolerance of neural networks, which permits certain memory data errors, more aggressive voltage scaling becomes viable. This fault tolerance has been harnessed to devise power conservation strategies within neural networks. Previous research [5] used a multi-voltage approach, applying high voltage to critical bits and low voltage to other bits, using only one type of SRAM. This approach led to challenges in voltage control and required additional, complex voltage control circuits to manage two different voltages. In contrast, our study employs two types of SRAM, which operate at a consistent voltage level, thereby addressing the shortcomings identified in prior research.

A myriad of scholars have made notable contributions to low-power designs within hybrid architectures. For instance, Nemati and associates adeptly merged SRAM with RRAM to craft a low-power, in-memory computing architecture [6]. However, our research predominantly concentrates on the recognition processes inherent in neural networks of conventional, broadly deployed AI chips, thereby offering more immediately applicable scenarios at this juncture. Similarly, research conducted by Srinivasan et al. [7] adopted a hybrid 8T and 6T SRAM architecture to curtail power consumption in memory units specifically tailored for neural network applications. Nevertheless, our inquiry delves deeper, rigorously analyzing the influence of individual bits on the overall recognition accuracy. We have meticulously designed the bit allocation to optimize recognition rates, voltage efficiency, and the consequent increase in area.

Our research introduces a hybrid approximate SRAM, specifically optimized for the storage of neural network weight data, enabling broader voltage scaling while preserving the integrity of critical data. The SRAM design is bifurcated into a High Significance Part (HSP) utilizing robust 8T SRAM and a Low Significance Part (LSP) employing less stable 6T SRAM. This arrangement ensures minimized bit errors in the HSP at equivalent supply voltages, consequently reducing power consumption in deep learning tasks without sacrificing classification accuracy.

This letter introduces 6T SRAM and 8T SRAM in Sect. 2, explains the proposed method in Sect. 3, performs simulation and evaluation in Sect. 4, and concludes in Sect. 5.

2. Comparison between 6T SRAM and 8T SRAM

When the voltage of SRAM's power supply is lowered, bit errors become more likely during data write and read access, particularly due to noise. This issue is more acute in read

Manuscript received March 13, 2024.
Manuscript revised April 25, 2024.
Manuscript publicized May 20, 2024.

[†] Graduate School of Science and Engineering, Chiba University, Chiba-shi, 263–8522 Japan.
[††] Graduate School of Engineering, Chiba University, Chiba-shi, 263–8522 Japan.
a) E-mail: wuji@ieee.org

DOI: 10.1587/transinf.2024EDL8022

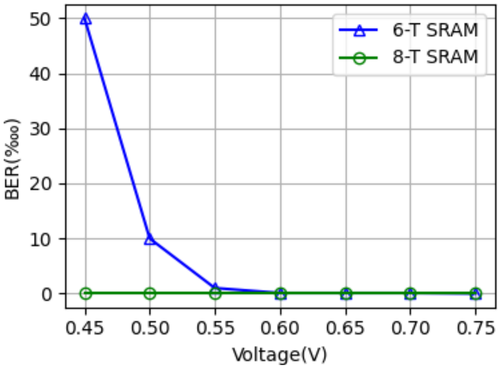

Fig. 1 Comparison of BER between 6T SRAM and 8T SRAM

access as voltage decreases. In a 6T cell, reading involves measuring the potential difference between the Bit Line and Bit Line Bar, which becomes challenging at lower voltages. However, in an 8T cell, the read Static Noise Margin is significantly improved by separating the Bit Line. Read access in an 8T cell only measures the Read Bit Line's potential, avoiding any impact on the cell's contents.

According to Previous studies [3], [8]–[10], as depicted in Fig. 1, highlight a marked contrast in the BER of 8T and 6T SRAM at the same voltages. Notably, 6T SRAM exhibits increased bit errors when voltages fall below 0.6 volts, while 8T SRAM maintains stable performance. However, the production of 8T SRAM comes at a higher cost and occupies a larger area of the circuit. Hence, to attain lower operating voltages, particularly under 0.6 volts, without a substantial increase in circuit area, the adoption of a 6T-8T hybrid architecture emerges as a feasible strategy, balancing various considerations.

3. Proposed Method

As shown in Fig. 2, This study proposes a low-voltage 6T-8T hybrid SRAM that preserves CNN weight parameters. Weight parameters are stored in IEEE754 standard single precision floating point number format. As explained in Sect. 2, lowering the SRAM voltage increases the BER. Since lowering the memory operating voltage is a method to reduce power consumption, the purpose of this research is to suppress the impact on CNN's classification accuracy even if BER occurs due to lowering the operating voltage.

In typical W-bit data encoding, ranging from 0 to W-1, the lower bits of the memory address are usually aligned rightward, with the LSB minimally impacting the overall value upon alteration. In floating-point number calculations, the lower bits of the mantissa have the least influence among all bits. Conversely, errors in the exponent bit significantly multiply the effect on the actual value, and a reversal in the sign bit completely inverts this value. Thus, the exponent and sign bits are crucial for the integrity of a floating-point number. In order to prevent the CNN's classification accuracy from being affected even when the operating voltage decreases, it is assumed that BER will occur, and if important bits are protected from errors, errors will only occur in non-important bits. Nevertheless, conventional methods have demonstrated that such errors typically result in only minor changes in the actual data value.

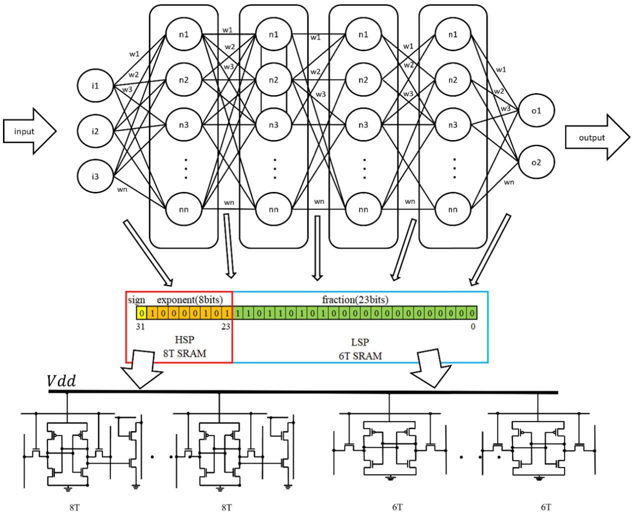

Fig. 2 proposed 6T-8T hybrid SRAM model

4. Simulation

4.1 Simulation Conditions

In the simulations of this research, we use LeNet-5 as the neural network module and MNIST as the dataset. The training was performed 1000 times to generate the network model, and the training accuracy was 98.28%. This research assumes using already trained CNNs on edge computing devices and adds errors only during inference.

In this letter, 99% of the accuracy during training is used as the evaluation standard during inference. In other words, if the accuracy during inference exceeds 97.30%, the inference is considered successful. In the simulation, by setting the BER from 0 to 10^{-8} or the number of upper protection bits from 0 bits to 9 bits, we investigated the relationship between the number of protection bits and inference accuracy under a constant BER, We investigate the relationship between BER and inference accuracy under the number of protection bits.

4.2 Result and Evaluation

Figure 3 illustrates the correlation between BER and inference accuracy, contingent upon a constant count of protected bits. As inferred from Fig. 3, safeguarding 9 bits results in an accuracy surpassing 97.3%, thereby limiting the BER of the residual bits to a maximum of $10^{-2.6}$. Hence, protecting the top 9 bits ensures successful inference provided the BER of the lower 23 bits remains below $10^{-2.6}$. Additionally, with zero protection bits, i.e., without any bit protection, successful inference is achievable if the BER stays under 10^{-5}.

Table 1 evaluates operating voltage and area implications. This study achieves a reduction in operating voltage by approximately 70% and in power consumption by about 45%

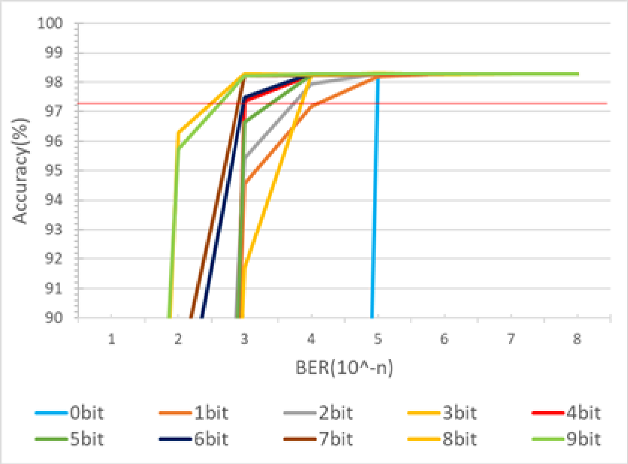

Fig. 3 Relationship between BER and inference accuracy

Table 1 Evaluation of operating voltage and area of the proposed SRAM

Number of protection bits	5	6	7	8	9
Maximum BER(10ˆn)	-3.3	-3	-2.9	-2.5	-2.6
Operating voltage(V)	0.46	0.46	0.46	0.46	0.45
Operating voltage ratio(%)	70.33	75.41	74.53	74.69	73.18
Area ratio (%)	103.13	103.75	104.38	105	105.63

compared to [5]. Particularly, with 9 bits under protection, the operating voltage can be decreased by roughly 53.23%, and the area cost rises by 5.63% relative to traditional 6T SRAM operating at 1V.

5. Conclusion

This letter proposes a hybrid SRAM, utilizing the varying reliability of 8T and 6T SRAM cells, for storing weights in convolutional neural networks. Its structure includes HSP and LSP, ensuring crucial bits are safeguarded against errors under a unified operating voltage. With all 9 bits of the HSP protected, equivalent inference accuracy is attainable with a Bit Error Rate of LSP under $10^{-2.6}$. At a reduced operating voltage of 0.45V, the HSP remains error-free with an LSP BER of 10^{-3}. Despite the area of proposed SRAM being up to 105.63% larger than conventional 6T SRAM, it enables a substantial reduction in operating voltage by about 47% and power consumption by up to 55%.

References

[1] I. Ahmed, G. Jeon, and F. Piccialli, “From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where,” IEEE Trans. Ind. Informat., vol.18, no.8, pp.5031–5042, 2022.

[2] B. Bahmei, E. Birmingham, and S. Arzanpour, “Cnn-rnn and data augmentation using deep convolutional generative adversarial network for environmental sound classification,” IEEE Signal Process. Lett., vol.29, pp.682–686, 2022.

[3] F. Samiullah, M.L. Gan, S. Akleylek, and Y. Aun, “Group key management in internet of things: A systematic literature review,” IEEE Access, vol.11, pp.77464–77491, 2023.

[4] V. Kumar, R.k. Shrivatava, and M.M. Padaliya, “A temperature compensated read assist for low vmin and high performance high density 6t sram in finfet technology,” 2018 31st International Conference on VLSI Design and 2018 17th International Conference on Embedded Systems (VLSID), pp.447–448, 2018.

[5] K. Kozu, Y. Tanabe, M. Kitakami, and K. Namba, “Low power neural network by reducing sram operating voltage,” IEEE Access, vol.10, pp.116982–116986, 2022.

[6] S.H.H. Nemati, N. Eslami, and M.H. Moaiyeri, “A hybrid sram/rram in-memory computing architecture based on a reconfigurable sram sense amplifier,” IEEE Access, vol.11, pp.72159–72171, 2023.

[7] G. Srinivasan, P. Wijesinghe, S.S. Sarwar, A. Jaiswal, and K. Roy, “Significance driven hybrid 8t-6t sram for energy-efficient synaptic storage in artificial neural networks,” 2016 Design, Automation and Test in Europe Conference and Exhibition (DATE), pp.151–156, 2016.

[8] L. Yang and B. Murmann, “Sram voltage scaling for energy-efficient convolutional neural networks,” 2017 18th International Symposium on Quality Electronic Design (ISQED), pp.7–12, 2017.

[9] L. Yang, D. Bankman, B. Moons, M. Verhelst, and B. Murmann, “Bit error tolerance of a cifar-10 binarized convolutional neural network processor,” 2018 IEEE International Symposium on Circuits and Systems (ISCAS), pp.1–5, 2018.

[10] L. Chang, R.K. Montoye, Y. Nakamura, K.A. Batson, R.J. Eickemeyer, R.H. Dennard, W. Haensch, and D. Jamsek, “An 8t-sram for variability tolerance and low-voltage operation in high-performance caches,” IEEE J. Solid-State Circuits, vol.43, no.4, pp.956–963, 2008.