Design of a Spiking Neural Network Based on Stochastic Neuron

Hyeyeon Jeon[1], Yoon Kim[1,3], and Minsuk Koo[a,2,3]
[1]Department of Electrical and Computer Engineering, University of Seoul, Seoul 02504, South Korea
[2]School of Advanced Fusion Studies, University of Seoul, Seoul 02504, South Korea
[3]IM Electronics co., Seoul 02505, South Korea
E-mail : janet1999@naver.com

***Abstract* - In this study, we propose a 9×10 neuromorphic architecture that utilizes stochastic spiking neurons and SRAM as synapses. The stochastic spiking neuron features a capacitor-less structure, reducing hardware complexity and enabling low-power operation. Additionally, binary SRAM is employed as synapses to further enhance low-power characteristics. The proposed architecture was fabricated using the TSMC 28nm CMOS process. Simulation results confirm that the proposed design achieves a high energy efficiency of 60.55 TOPS/W. Through this approach, we aim to develop a low-power artificial neural network suitable for edge devices.**

***Keywords*—Spiking neural network, neuromorphic system, stochastic computing, binary SRAM, pattern classification**

I. INTRODUCTION

AI has recently emerged as a core technology in next-generation industries, leading to an increasing demand for large-scale data processing. However, the traditional von Neumann architecture, in which the processing unit and memory unit are separate, faces challenges in large-scale parallel computation due to bottlenecks. To address this von Neumann bottleneck, neuromorphic architectures inspired by the human brain have been proposed.

Spiking neural networks (SNNs) are a type of neuromorphic architecture that mimics the way biological neural networks process information. In general, SNNs operate by receiving spikes from input neurons, performing multiplication with synaptic weights, and summing the computed values in the output neurons. When the membrane potential exceeds a certain threshold, an output spike is generated. Since neural interactions are processed through spikes and operate in an event-driven manner only when signals occur, SNNs offer the advantage of low power consumption. Due to this low-power characteristic, SNNs are well-suited for next-generation AI hardware. [1],[2],[3]

In this paper, we propose the design of a spiking neuron that operates with reduced energy consumption by using stochastic neurons that generate spikes with a certain probability for given inputs. In our previous study [4], we introduced the concept of the "stochastic bit", demonstrating its potential to implement SNNs in a power-efficient manner. The synapses are implemented using 1-bit SRAM to achieve a simple operational mechanism. The fabricated chip was manufactured using TSMC's 28nm process, implementing a 9×10 input array.

a. Corresponding author; koo@uos.ac.kr

Manuscript Received Feb. 13, 2025, Revised May 15, 2025, Accepted May 15, 2025

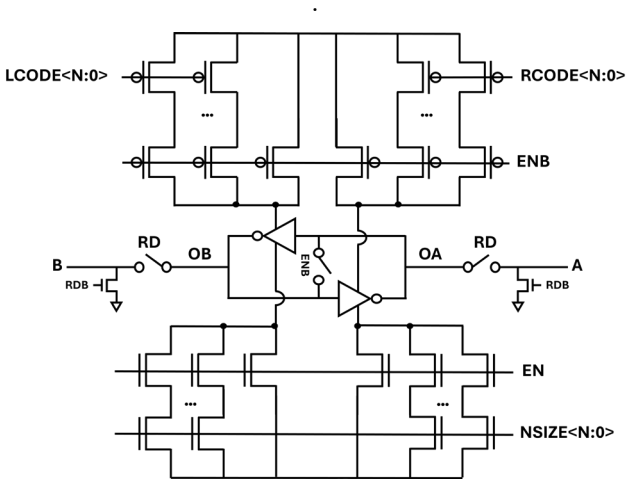
II. CIRCUIT DESIGN

A. Stochastic Spiking Neuron Circuit

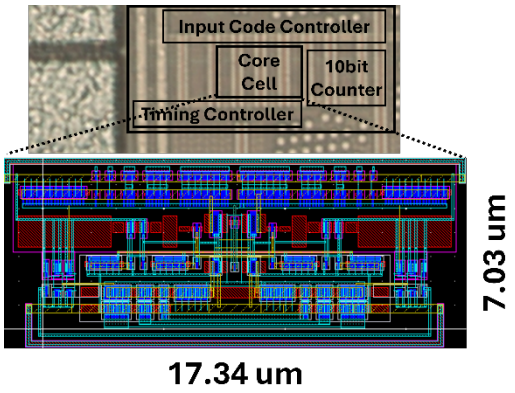
We propose a binary spiking neural network utilizing stochastic bits to design an energy- and memory-efficient edge neuromorphic computing circuit. The key feature of stochastic SNN is its probabilistic bit generation mechanism, where the logic high and low states are determined by a sigmoid probability function based on the input code.

To implement the stochastic core circuit shown in Fig. 1, the input signals were designed in a digital format as RCODE<4:0> and LCODE<4:0>. These input values are applied to the gates of PMOS transistors forming the Left wing and Right wing. This configuration determines the current flowing through the transistors on the Left and Right sides, which is then supplied to the cross-coupled inverter. Consequently, the inverter′s output nodes, Output A (OA) and Output B (OB), are determined based on the current difference between the Left and Right sides, where one node becomes high and the other low.

Using this probabilistic mechanism, the stochastic core circuit was employed as the neuron of the spiking neural network, generating spikes based on the input values. Unlike the commonly used Integration-and-Fire (IF) neurons in other spiking neural networks, this design does not require capacitors for charge storage, enabling low-power operation.



(a)



(b)

Fig. 1. Stochastic core circuit (a) schematic and (b) layout.

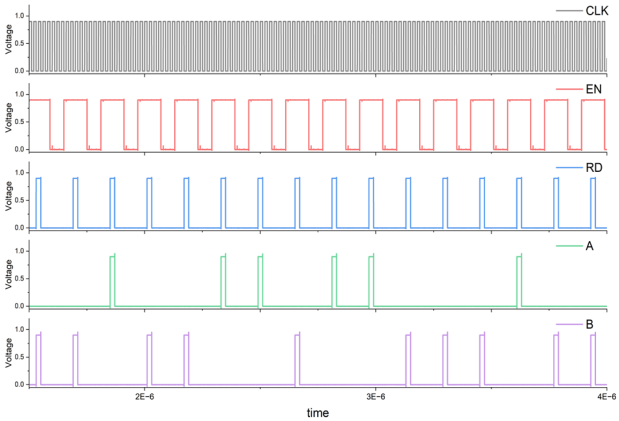
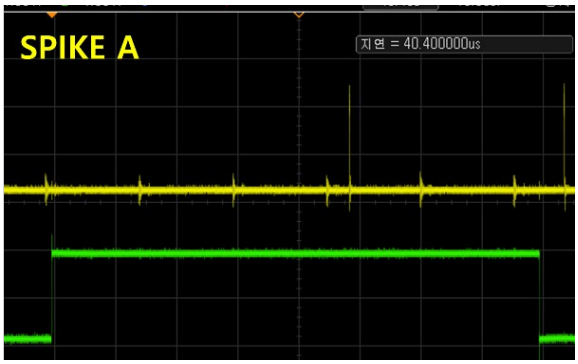


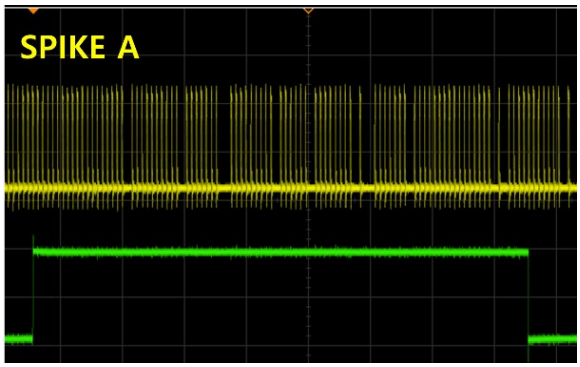
Fig. 2. Timing diagram of output A/B based on input signals EN/RD.

The simulation results show that, based on the Enable (EN) and Read (RD) signals, spikes are generated with a sigmoid probability according to the input code, as illustrated in Fig.2. The input code represents the difference between RCODE and LCODE. When all right-wing transistors corresponding to RCODE are turned on and all left-wing transistors corresponding to LCODE are turned off, the configuration is 11111_2 - 00000_2, which corresponds to an input code of 31. In this case, all 1000 spike events occur at node A, resulting in a spike probability of 100%. Conversely, when the configuration is 00000_2 - 11111_2, corresponding to an input code of -31, all 1000 spike events occur at node B, yielding a spike probability of 0%.

The measurement results, as shown in Fig.3, confirm that the number of generated spikes varies according to the input code. Fig.3 (a) presents an oscilloscope capture of the spike events when the input code is -31, demonstrating a near 0% spike probability. In contrast, Fig.3 (b) shows the oscilloscope capture when the input code is 31, where the output spikes occur with nearly 100% probability. The number of output spikes was measured by sweeping the input code from -31 to 31 while applying 1000 EN signals.



(a)

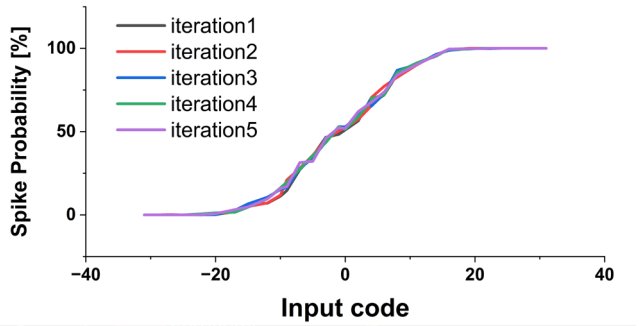


(b)

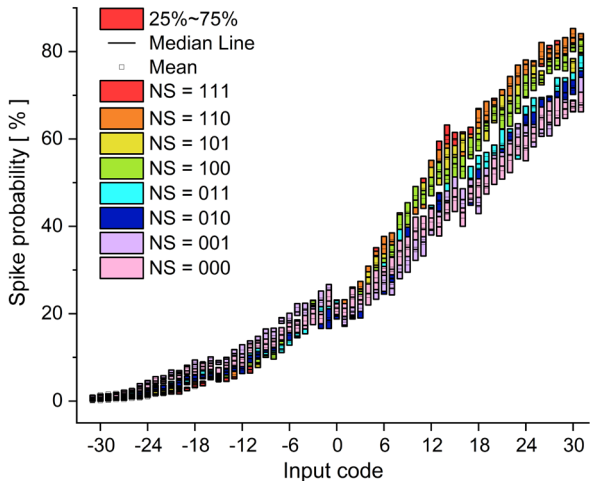
Fig. 3. Measurement of spike generation result according to input code (a) Input code = -31 = 0 % (b) Input code = 31 = 100 %.

In the simulation results shown in Fig. 4(a), spikes are generated based on a sigmoid-shaped probability. The measurement results in Fig. 4(b) similarly demonstrate that spikes are generated with a nonlinear probability. The supply voltage (VDD) was set to 1.15 V, and to compensate for the increased VDD, a high signal was applied to the gate of the lower NMOS transistor using the NSIZE code.

The measurement results indicate that when the NSIZE code is fully high (111_2), the observed spike generation closely matches the simulation results.



(a)



(b)

Fig. 4. Spike generation probability according to the stochastic input neuron's code (a) Cadence simulation result (b) Chip measurement results.

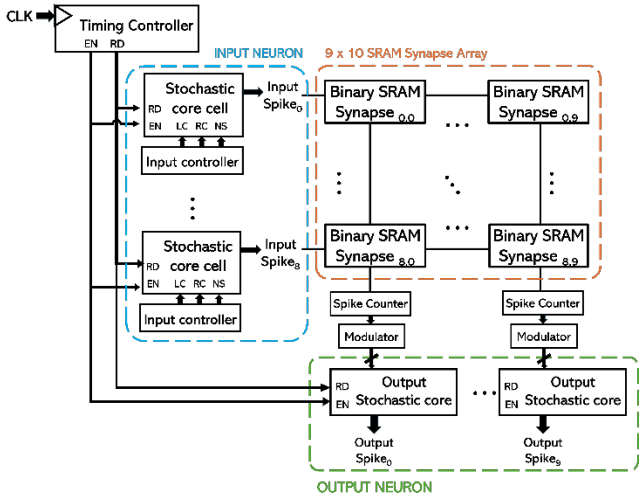
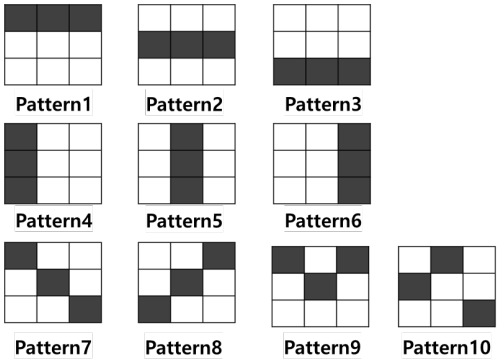


Fig. 5. 9×10 array architecture.

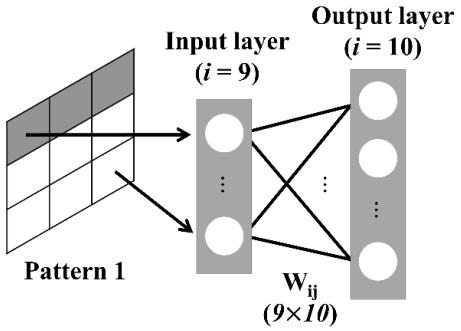
B. 9×10 Array Architecture

For the pattern classification simulation, the array consists of 9 input neurons, 10 output neurons, and 90 one-bit binary SRAM synapses, as shown in Fig.5. Spikes generated by the input neurons are multiplied by the synaptic weights and transmitted to the spike counter. The counter counts the number of spikes occurring in each column and forwards this information to the modulator. The modulator converts the spike count into a format suitable for input to the output neurons.

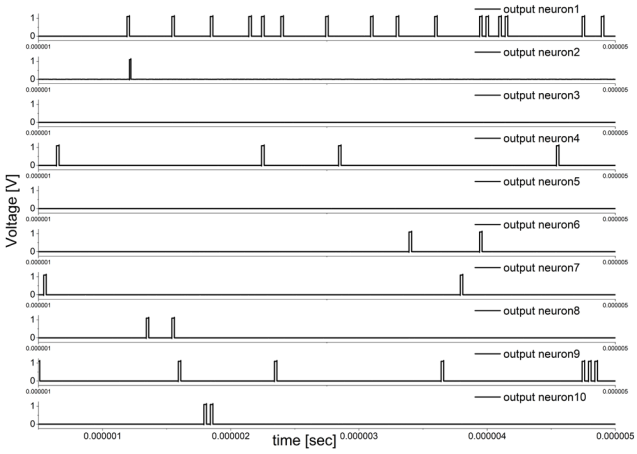
In the output neurons, the stochastic core circuit generates spikes with a certain probability. The classification follows a rate coding approach, where the output neuron that produces the highest number of spikes is considered the correct result.[5],[6],[7]



(a)



(b)



(c)

Fig. 6. (a) 10 types of input patterns, (b) Pattern classification simulation input scheme, and (c) Simulation result of inference pattern 1.

III. EVALUATION

A. Pattern classification simulation

For the array simulation, ten arbitrary patterns were defined, as shown in Fig.6 (a), and each was matched to one of the ten output neurons. The 3×3-sized patterns were

applied to the nine input neurons. The weights were determined through MATLAB simulations, and inference simulations were conducted using the Cadence tool.

The results from the Cadence simulation, as shown in Fig.6 (c), confirm that the output neuron corresponding to pattern 1 generates the highest number of spikes, demonstrating successful pattern matching.

Based on the experimentally measured stochastic core probability graph, the MATLAB simulation results confirm the expected behavior, as shown in Fig.7. The measured spike generation probabilities were incorporated into the MATLAB code to simulate the stochastic spike generation process.

The spike generation probability values used in the simulation correspond to the average measured values when NS = 111.

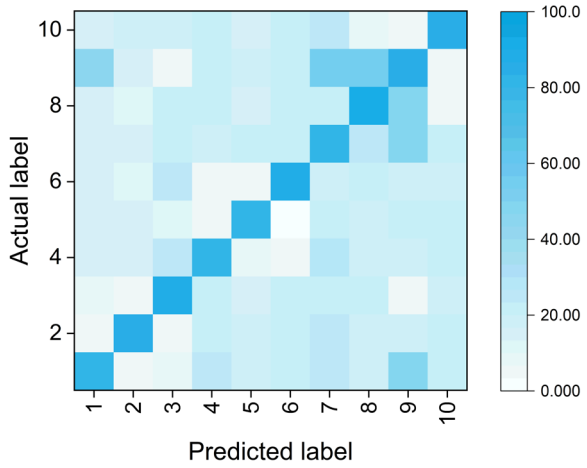


Fig. 7. Pattern classification simulation result.

B. Power Consumption

Power consumption simulations were conducted based on 200 RD spikes, a core VDD of 1.15 V, and a peripheral circuit VDD of 0.9 V. The stochastic core consumes 3.36 µW of power, and since it is included in 9 input neurons and 10 output neurons, the total power consumption amounts to 63.84 µW. The modulator connected to each output neuron consumes a maximum of 2.58 µW per unit, while each pulse counter consumes a maximum of 16.51 µW, resulting in a total power consumption of 190.9 µW ($2.58 \times 10 + 16.51 \times 10$). The timing controller consumes 669.1 nW. In the synapse array, each AND gate operation, which multiplies the stored SRAM value and the input spike, consumes 105.7 nW, leading to a total synapse array power consumption of 9.5 µW. Consequently, the total power consumption during the inference process is 264.24 µW.

During inference, 90 synapses perform multiplication operations, and 10 output neuron pulse counters perform addition operations. As a result, the system achieves an energy efficiency of 60.55 TOPS/W ($\frac{(90+10)\times 160\ MHz}{264.24\ \mu W}$).

As shown in Table I, the comparison with previous studies confirms that using stochastic neurons achieves higher energy efficiency compared to LIF SNN. Additionally, the use of SRAM enables fabrication with standard CMOS processes, allowing for low-cost manufacturing. Despite employing a smaller technology node compared to reference [4], our work exhibits lower energy efficiency (TOPS/W). This is attributed to utilizing a core voltage of 1.15 V rather than 0.9 V, and targeting a smaller-scale 9×10 array for basic pattern classification, as opposed to the 784×400×10 array designed for MNIST classification in reference [4].

TABLE I. Comparison of proposed method with state-of-the-art

	This work	[4]	[8]	[9]	[10]
TOPS/W (Tech node)	60.55 (28 nm)	89.49 (90 nm)	12.19 (65 nm)	18.7 (40 nm)	17.56 (40 nm)
Network size	9×10	784×400×10	784×300×10	X	X
System configuration	Stochastic SNN	Stochastic SNN	Stochastic SNN	LIF	LIF
Synapse	SRAM	SRAM	RRAM	Flip-Flops	SOT-MRAM
Weight resolution	1bit	1bit	3bit	8bit	4bit
CMOS-compatible synapse	○	○	X	○	X

IV. CONCLUSION

This study fabricated an artificial neural network using stochastic spiking neurons and verified its operation through measurements. The stochastic neuron operates based on stochastic behavior, directly generating spikes without accumulating input values, enabling lower power consumption compared to conventional LIF or IF neurons.

Through this approach, a high energy efficiency of 60.55 TOPS/W was achieved. The proposed architecture aims to overcome the energy efficiency limitations caused by the von Neumann bottleneck and serves as a suitable hardware platform for next-generation AI applications.

The current 9×10 single-layer array can be further developed into a more advanced array with higher accuracy and integration, offering the potential for enhanced performance and scalability.

ACKNOWLEDGMENT

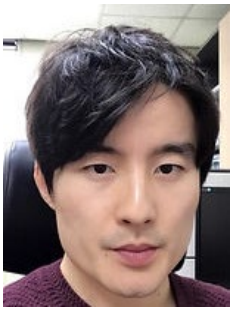
This work was supported by the National R&D Program through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT (RS-2024-00402495) (50%). This work was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0017011, HRD Program for Industrial Innovation) (50%). The chip fabrication and EDA tools were supported by the IC Design Education Center(IDEC), Korea.

REFERENCES

[1] S. R. Kheradpisheh, T. Masquelier, and A. Maida, “Deep learning in spiking neural networks,” Neural Netw., vol. 111, pp. 47–63, Mar. 2019.

[2] Y. Liu, S. Liu, Y. Wang, F. Lombardi, J. Han, "A Survey of Stochastic Computing Neural Networks for Machine Learning Applications," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, pp. 2809-2824, Aug. 2020.
[3] P. Wijesinghe, A. Ankit, A. Sengupta, K. Roy, "An All-Memristor Deep Spiking Neural Computing System: A Step Toward Realizing the Low-Power Stochastic Brain," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, pp. 345-358, Sep. 2018.
[4] M.Koo, G.Srinivasan, Y.Shim, and K.Roy, "SBSNN:Stochastic-bits enabled binary spiking neural network with on-chip learning for energy efficient neuromorphic computing at the edge," IEEE Transactions on Circuits and Systems I, vol. 67, pp. 2546-2555, Aug. 2020.
[5] O. Krestinskaya, L. Zhang, K. N. Salama, "Towards Efficient In-Memory Computing Hardware for Quantized Neural Networks: State-of-the-Art, Open Challenges and Perspectives," IEEE Transactions on Nanotechnology, vol. 22, Jul. 2023.
[6] Y. Qi, Y. Feng, H. Wang, C. Wang, M. Bai, J. Liu, X. Zhan, J. Wu, Q. Wang, and J. Chen, "Flash-Based Computing-in-Memory Architecture to Implement High-Precision Sparse Coding," Micromachines 2023, Nov. 2023.
[7] M. Ali, S. Roy, U. Saxena, T. Sharma, A. Raghunathan, K. Roy, "Compute-in-Memory Technologies and Architectures for Deep Learning Workloads," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 30, Nov. 2022.
[8] H.Kim, Y.An, M.Kim, G.C.Heo, and Y.Shim, "All Stochastic-Spiking Neural Network (AS-SNN): Noise Induced Spike Pulse Generator for Input and Output Neurons With Resistive Synaptic Array," IEEE Transactions on Circuits and Systems II, vol. 72, pp. 78-82, Jan. 2025.
[9] S.Uenohara, K.Aihara, "A 18.7 TOPS/W Mixed-Signal Spiking Neural Network Processor With 8-bit Synaptic Weight On-Chip Learning That Operates in the Continuous-Time Domain," IEEE Access, vol. 10, pp. 48338-48348, Apr. 2022.
[10] H.Fu, Y.Huang, T.Chen, C.Fu, H.Ren, and Y.Zhou, "DS-CIM: A 40nm Asynchronous Dual-Spike Driven, MRAM Compute-In-Memory Macro for Spiking Neural Network," IEEE Transactions on Circuits and Systems I, vol. 71, pp. 1638-1650, Jan. 2024.

Hyeyeon Jeon received the B.S. degree in electrical and computer engineering from the University of Seoul, Seoul, South Korea, in 2023, where she is currently pursuing the M.S. degree in electrical and computer engineering.



Yoon Kim (Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Seoul National University, Seoul, South Korea, in 2006 and 2012, respectively. From 2012 to 2015, he was a Senior Engineer with Samsung Electronics Company Ltd., South Korea. He was an assistant professor at Pusan National University, Busan, South Korea, from 2015 to 2019. In 2018, he joined the University of Seoul and became an Associate Professor in 2020.

Minsuk Koo (Member, IEEE) received the B.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2007, and the M.S. degree in electrical engineering from Seoul National University, Seoul, South Korea, in 2009. He received a Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA. From 2009 to 2012, he was a Senior Engineer with RadioPulse Inc., Seoul, South Korea. He was an assistant professor at Incheon National University, Incheon, South Korea, from 2020 to 2024 and has been an associate professor at the University of Seoul since 2024.