Low-Power Spiking Neural Network with Clock-gating technique

Rui Shiota
School of Computer Science and Engineering, University of Aizu, Japan
Email: s1290033@u-aizu.ac.jp
Supervisor: DANG Nam Khanh

***Abstract*—Spiking Neural Networks (SNNs) have gained significant attention in the field of Artificial Intelligence (AI) due to their ability to recognize and learn complex patterns in various types of data, including audio and images. Applications of SNNs are widespread, ranging from autonomous vehicles to machine translation systems. A critical challenge in adjusting SNNs for real-world applications is improving their energy efficiency. One effective approach to address this challenge is the use of clock gating, a technique designed to reduce power consumption in digital circuits. This study evaluates the impact of clock gating on power efficiency by comparing SNNs designs with and without the implementation of clock gating. Simulations, synthesis, and power estimations were conducted to assess the differences in power consumption. The findings of this study highlight the 24.70% power reduction achieved through the application of clock gating in SNNs designs.**

***Index Terms*—Spiking Neural Networks, IF neuron, power, clock gating, hardware design**

I. Introduction

Spiking Neural Networks (SNNs) are brain-inspired computational models that utilize individual abstract neurons to replicate the communication processes of biological systems. These models mimic the spiking behavior of biological neurons, enabling them to reproduce patterns of neuronal activity[1]. A common approach for implementing SNNs is through the design of hardware architectures, often realized on Application-Specific Integrated Circuits (ASICs) or Field-Programmable Gate Arrays (FPGAs). ASICs are integrated circuits tailored for specific applications, while FPGAs are reprogrammable logic devices. Due to the complexity of SNN models, their hardware implementation can lead to significant power consumption.

This study proposes the use of the Integrate-and-Fire (IF) neuron model, one of the simplest and most widely used spiking neuron models in SNNs. The IF model effectively simulates the firing patterns and information transmission of biological neurons. Additionally, it strikes a balance between computational cost and biological plausibility, making it a practical choice for hardware implementations[2][3]. Membrane potential in IF neuron is calculated based on below equation[4].

$$V_j(t) = V_j(t-1) + \sum_{i=1}^{n} w_{i,j} \times x_i(t-1) - \lambda$$

$$\left(x_j(t) = \begin{cases} 1, & \text{if } V_j(t) > V_{\text{thres}} \\ 0, & \text{otherwise} \end{cases} \right)$$

$V_j(t)$: membrane potential of neuron at time step

$w_{i,j}$: synapse weight between neuron and neuron

$x_i(t-1)$: output of the presynaptic neuron

V_{thres} : threshold value

Energy efficiency is a critical consideration in SNNs. Currently, power consumption in SNNs can be reduced by factors of 100 to 1,000, making them far more energy-efficient compared to traditional non-neuromorphic systems[5]. Achieving energy efficiency is crucial not only for reducing operational costs but also for minimizing the environmental impact of AI systems, which are typically power-hungry and contribute to carbon dioxide emissions[6][7]. While SNNs are not without environmental impact, they offer a more energy-efficient alternative compared to conventional AI systems.

Power consumption in SNNs can be categorized into dynamic and static power. Below equations[8] are the definitions of them.

$$Dynamic\ Power = f_{sw} \times C_L \times V_{cc}^2 + T_{SC} \times I_{peak} \times V_{cc}$$

$$Static\ Power = V_{cc} \times I_{cc}$$

f_{sw} : switching frequency

C_L : dynamic effective capacitance

V_{cc} : voltage applied to a logic IC

T_{SC} : shortcut-circuit time period

I_{peak} : peak current

I_{cc} : static supply current of the IC

A widely used technique to reduce dynamic power consumption is clock gating. Clock gating is a method employed in sequential circuits to save power by turning off the clock[9]. When the clock is disabled, no switching or computational activity occurs, leading to significant power savings.

The methodology section of this paper outlines the processes involved, including coding, simulation, synthesis, and power consumption analysis. The results section provides an estimation of power consumption, focusing on the reduction achieved through clock gating in both individual neurons and neuron networks. The discussion section explores potential future directions for further power reduction. The paper concludes by summarizing the key findings.

II. Methodology

Initially, Verilog HDL code for the Integrate-and-Fire (IF) neuron model without clock gating is developed. Subsequently, a simulation is performed to verify the correctness of the code using ModelSim. Once the simulation results are verified, the design is synthesized using Synopsys Design Compiler. This step generates a Verilog file, which is then utilized in the subsequent simulation and power estimation phase. In the post-synthesis simulation, the results are compared to those obtained in the pre-synthesis simulation to ensure consistency. Once the results are confirmed to be consistent, power estimation is carried out using Synopsys PrimeTime. This phase leverages the Verilog code generated during the synthesis step. Both dynamic and static power consumption are analyzed in this stage. After completing the aforementioned processes, the same steps are performed using the Verilog HDL code for the IF neuron with clock gating. Once the processes involving the clock-gated code are completed, the power consumption results from the two designs are compared to evaluate the impact of clock gating.

A. Design of IF neuron

Figure 1 illustrates the design of the IF neuron network, which consists of five (3:2) neurons. In this configuration, the network has a 3-bit input signal and a 2-bit output signal. Each neuron in the network includes memory, and the weights associated with each memory are determined based on the number of input signals, as shown in Figure 2.

The Verilog HDL code for this design is depicted in Figure 3.

The neuron outputs are determined by the following algorithm, where the values of the weight are randomly assigned using a random number function. “i” means the index of “inspike” and “Weight”. “n” represents the number of bits of them. The threshold voltage is fixed at a value of 43.

B. Clock gating

1) Fundamentals of clock gating: This subsection explains the operation of clock gating using a simple adder circuit shown in Figure 4 as an example.

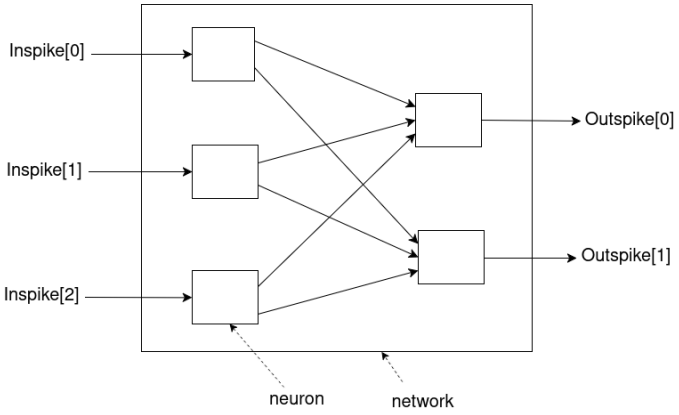

Fig. 1. Overall Design of IF neuron

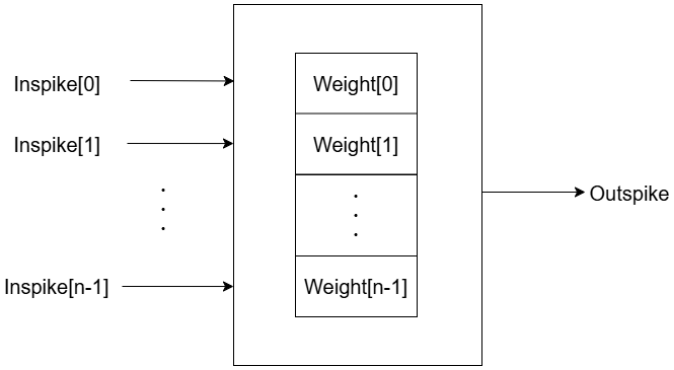

Fig. 2. Neuron Architecture.

As illustrated in Figure 5 and Figure 6, the gated clock signal is generated as the logical product of the clock signal and the clock enable signal. In a clock-gated system, the circuit operates only when the gated clock signal transitions to a high state.

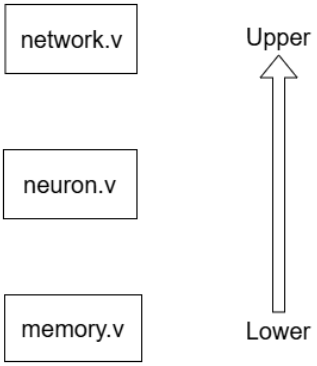

Fig. 3. Architecture of Verilog HDL files

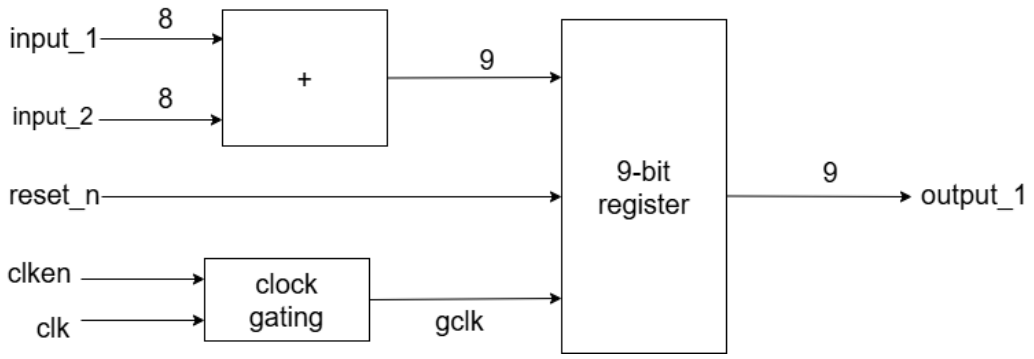

Fig. 4. Clock-gated Adder

Algorithm 1 Calculate Outspike

$V = V_{reset}$
$i = 0$
for $i < n$ **do**
  $V = V + inspike[i] * Weight[i]$
end for
if $V \geq Threshold$ **then**
  $outspike = 1$
  $V = V_{reset}$
else
  $ouspike = 0$
end if

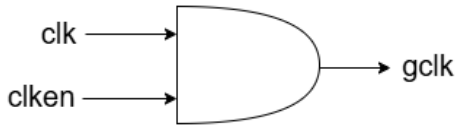

Fig. 5. Module of Clock Gating

Figure 7 and Figure 8 show the result of no clock gating and clock-gated adder respectively. These adders conduct calculation twice. The first calculation is 00100100 plus 10000001 in binary. The result of this calculation became 010100101 in binary. When this calculation is converted to decimal, this means that 36 plus 129 equals 165. The result of the second calculation is correct. The simulation was conducted for a duration of 135 ns. In a conventional adder, the circuit performs computations continuously. In contrast, a clock-gated adder executes computations only during one clock cycle whenever the input signal values change. While the clock-gated signal "clken" is disabled, the adder remains idle, performing no calculations. This reduction in activity leads to decreased power consumption, demonstrating the effectiveness of clock gating in lowering power usage.

The following three tables present the synthesis results. These tables provide the values for area, power, and timing, respectively. Non-combinational area is about sequential area such as flip-flop and latch. Because clock-gated cell is not a sequential area, only the value of No Clock Gating area increased. In the timing table, slack defined as the subtraction of data required time from data arrival time. If slack is above 0, it means that circuits are proper in terms of timing. Therefore, In this case, the timing of this circuit is appropriate. As for power, dynamic power decreased by 71.21% and static power increased by 2.11%, so total power decreased by 59.12%.

TABLE I
AREA COMPARISON OF NO CLOCK GATING AND CLOCK-GATED ADDER

	No Clock Gating Adder	Clock-gated Adder
Combinational Area	42.03 μm^2	43.09 μm^2
Noncombinational Area	40.70 μm^2	40.70 μm^2
Total Area	82.73 μm^2	83.79 μm^2

TABLE II
TIMING COMPARISON OF NO CLOCK GATING AND CLOCK-GATED ADDER

	No Clock Gating Adder	Clock-gated Adder
Data Required Time	1.96 ps	1.96 ps
Data Arrival Time	0.85 ps	0.85 ps
Slack	1.11 ps	1.11 ps

TABLE III
POWER COMPARISON OF NO CLOCK GATING AND CLOCK-GATED ADDER

	No Clock Gating Adder	Clock-gated Adder
Dynamic Power	7.05 $\times 10^{-6}$ W	2.03 $\times 10^{-6}$ W
Static Power	1.39 $\times 10^{-6}$ W	1.42 $\times 10^{-6}$ W
Total Power	8.44 $\times 10^{-6}$ W	3.45 $\times 10^{-6}$ W

2) Applying clock-gating to IF neuron: Clock gating is applied to the network.v and neuron.v modules. To ensure accuracy, the clock enable signal is set to 1 during the period when the neurons in the first layer read weights from memory and compute the output signal. Afterward, the clock enable signal is set to 0 until the neurons in the first layer read the weights again.

III. RESULTS

This section presents two types of results: the first focuses on power consumption at the neuron unit level, while the second addresses power consumption at the network unit level.

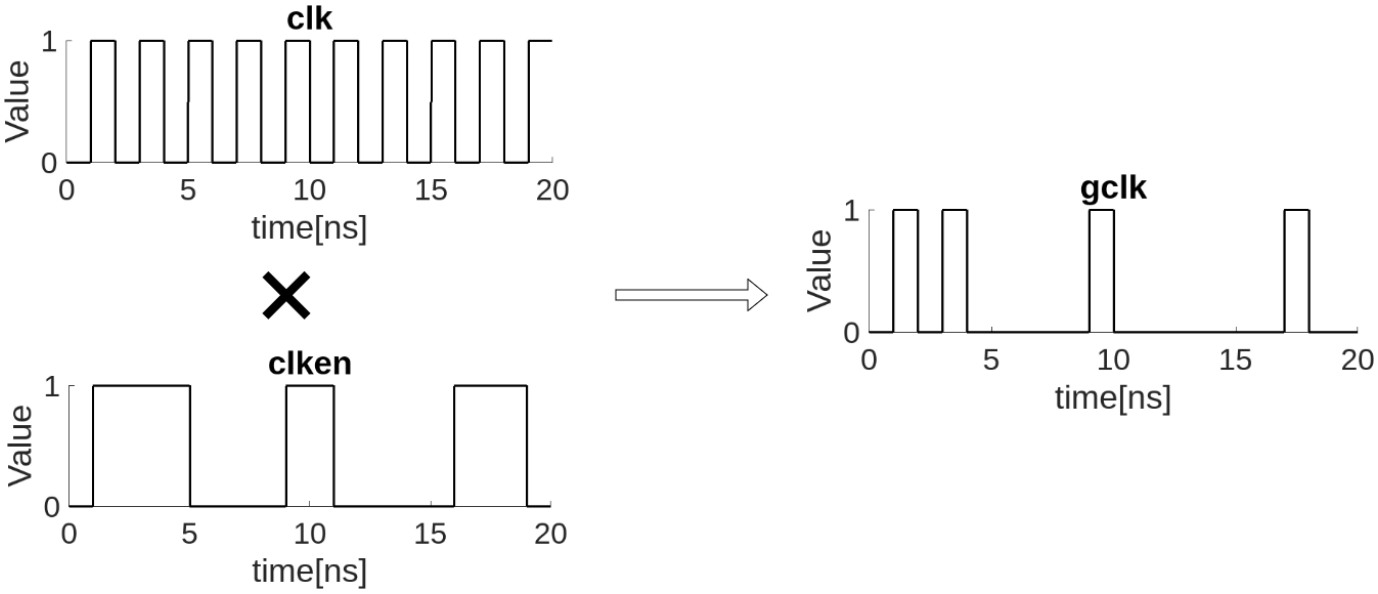

Fig. 6. Clock-gated Signals

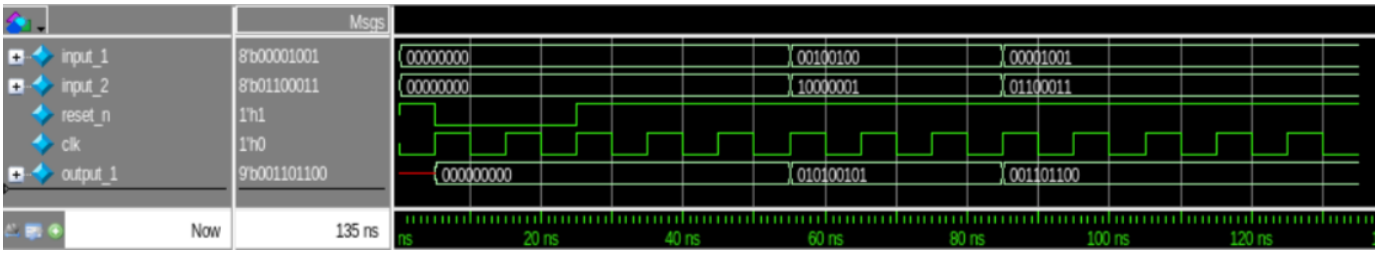

Fig. 7. No Clock Gating Adder

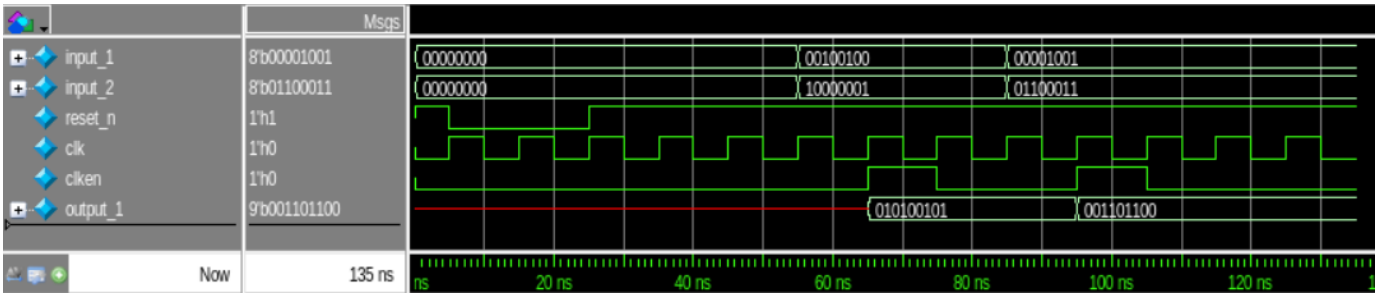

Fig. 8. Clock-gated Adder

A. *Simulation*

The simulation was conducted for a duration of 335 ns in neuron units and 435 ns in network units. As Table IV and Figure 9 shows, “inspike” and “outspike” are input and output signal respectively. “read_en”, “write_en” and “enable” signal are 5-bit in binary. They enable neurons to read values of weight from memory, write them to memory and do computation respectively depending on their values. “d1”,“d2” and “d3” are weight of three neurons in first layer and saved in memory. “d4” and “d5” are those in second layer. “mid” means a set of middle signals connecting first and second layer. The values of “outspike” of neurons in first layer are stored in this signal and work as “inspike” of neurons in second layer.

TABLE IV
RESPECTIVE SIGNALS AND ROLES

Signal	Role
clk	clock signal
clken	clock enable signal
gclk	clock-gated clock signal
reset	reset signal
inspike	input spike signal
read_en	signal to read weight from memory
write_en	signal to write weight to memory
enable	signal to work neurons
d1, d2, d3	weight of neurons in first layer
d4, d5	weight of neurons in Second layer
mid	signal to connect first and second layer
outspike	output signal

B. Power estimation in neuron units

The power consumption results for neuron units are summarized in Table VII. The dynamic power consumption was reduced by 26.23%, while the static power remained unchanged. Consequently, the total power consumption was reduced by 21.77%.

TABLE V
AREA COMPARISON OF NO CLOCK GATING AND CLOCK-GATED NEURON

	No Clock Gating Neuron	Clock-gated Neuron
Combinational Area	209.34 μm^2	211.74 μm^2
Noncombinational Area	316.54 μm^2	316.54 μm^2
Total Area	525.88 μm^2	528.28 μm^2

TABLE VI
TIMING COMPARISON OF NO CLOCK GATING AND CLOCK-GATED NEURON

	No Clock Gating Neuron	Clock-gated Neuron
Data Required Time	1.95 ps	1.95 ps
Data Arrival Time	1.41 ps	1.41 ps
Slack	0.54 ps	0.54 ps

TABLE VII
POWER COMPARISON OF NO CLOCK GATING AND CLOCK-GATED NEURON

	No Clock Gating Neuron	Clock-gated Neuron
Dynamic Power	5.30 $\times 10^{-5}$ W	3.91 $\times 10^{-5}$ W
Static Power	1.04 $\times 10^{-5}$ W	1.04 $\times 10^{-5}$ W
Total Power	6.34 $\times 10^{-5}$ W	4.96 $\times 10^{-5}$ W

C. Power estimation in network units

The power consumption results for network units are shown in Table X. While the static power consumption increased slightly, the dynamic power consumption decreased by 30.23%. As a result, the total power consumption for the network was reduced by 24.70%. In addition, simulation with clock enable signal disabled was done. Comparing with No Clock Gating network, dynamic power reduced by 97.53%. On the other hand, static power increased by 2.95%. As a result, total power decreased by 80.28%.

TABLE VIII
AREA COMPARISON OF NO CLOCK GATING AND CLOCK-GATED NETWORK

	No Clock Gating Network	Clock-gated Network
Combinational Area	898.28 μm^2	910.52 μm^2
Non-combinational Area	1311.38 μm^2	1311.38 μm^2
Total Area	2209.66 μm^2	2221.90 μm^2

TABLE IX
TIMING COMPARISON OF NO CLOCK GATING AND CLOCK-GATED NETWORK

	No Clock Gating Network	Clock-gated Network
Data Required Time	1.95 ps	1.95 ps
Data Arrival Time	1.66 ps	1.66 ps
Slack	0.29 ps	0.29 ps

TABLE X
POWER COMPARISON OF NO CLOCK GATING AND CLOCK-GATED NETWORK

	No Clock Gating Network	Clock-gated Network	Constantly Clock-gated Network
Dynamic Power	2.08 $\times 10^{-4}$ W	1.45 $\times 10^{-4}$ W	5.14 $\times 10^{-6}$ W
Static Power	4.28 $\times 10^{-5}$ W	4.29 $\times 10^{-5}$ W	4.41 $\times 10^{-5}$ W
Total Power	2.51 $\times 10^{-4}$ W	1.89 $\times 10^{-4}$ W	4.95 $\times 10^{-5}$ W

IV. DISCUSSION

Clock gating was introduced into the Verilog code of the Integrate-and-Fire (IF) neuron, and experiments were conducted to evaluate its impact. As a result, power consumption was successfully reduced at both the neuron and network unit levels.

For future work, further reductions in power consumption for Spiking Neural Networks (SNNs) are planned. However, achieving greater energy efficiency may require balancing accuracy and energy savings. To determine the acceptable level of accuracy reduction, a technique called “rate coding” will be explored. Figure 12 shows examples of rate coding. Rate coding is one of the most widely used coding schemes in neural network models. It represents information through spiking rates and has been a dominant paradigm in both neuroscience and artificial neural networks (ANNs) due to its robustness and simplicity[8].

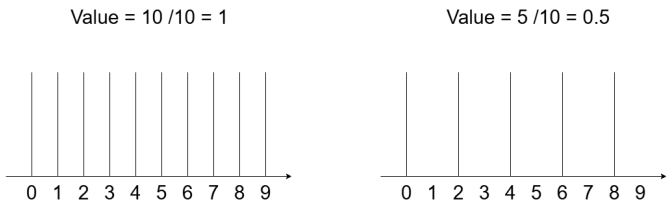

Fig. 12. Rate Coding

V. CONCLUSION

The Integrate-and-Fire (IF) neuron was designed both with and without clock gating. Following simulation and synthesis,

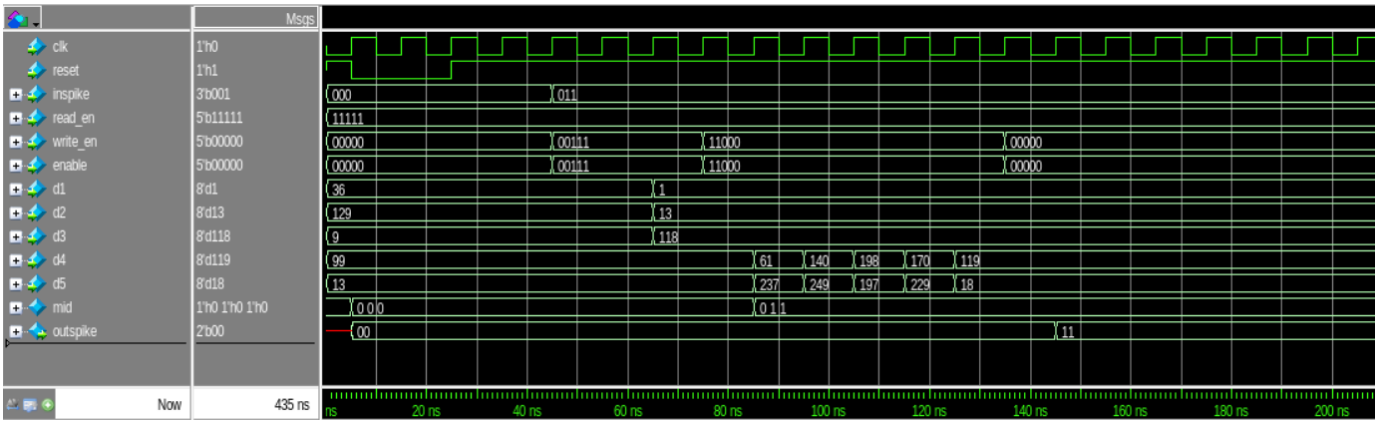

Fig. 9. Simulation of No Clock Gating Network

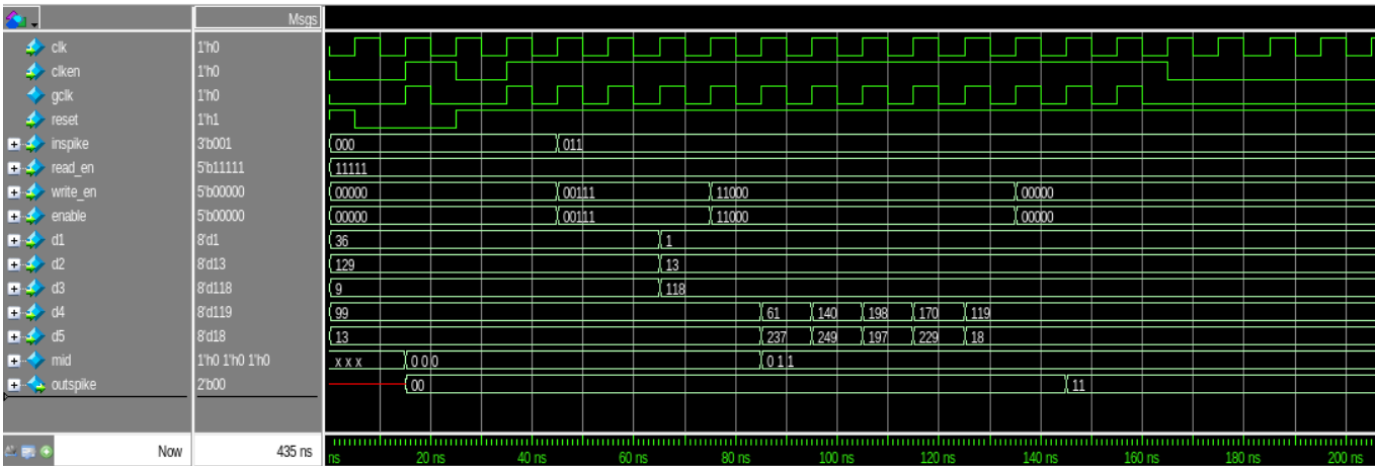

Fig. 10. Simulation of Clock-gated Network

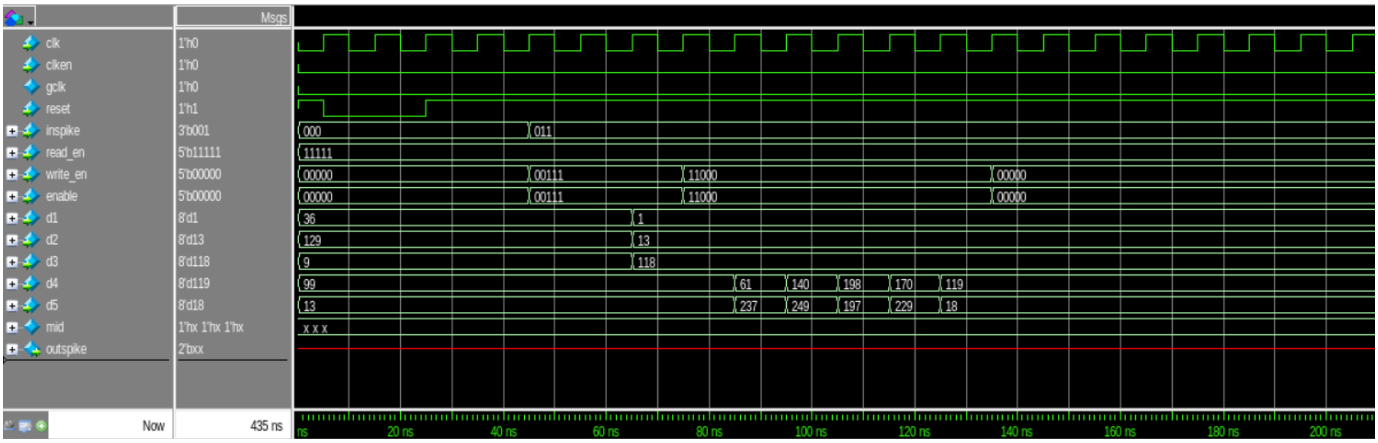

Fig. 11. Simulation of Clock-gated Network whose clock enable signal is always ”0”

power consumption was evaluated at both the neuron and network unit levels. In neuron units, total power consumption was reduced by 21.77%. Similarly, in network units, total power consumption decreased by 24.70%. To further enhance the efficiency of Spiking Neural Networks (SNNs), it may be necessary to balance energy efficiency with accuracy.

REFERENCES

[1] Prithwineel Paul, Petr Sosik, Lucie Ciencialova, "A survey on learning models of spiking neural membrane systems and spiking neural networks", *arXiv*, March 2024.

[2] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothee Masquelier, Tiejun Huang, Yonghong Tian, "Incorporating Learnable Membrane Time Constant to Enhance Learning of Spiking Neural Networks", *arXiv*, August 2021.

[3] Xiaoyan Fang, Derong Liu, Shukai Duan, Lidan Wang, "Memristive LIF Spiking Neuron Model and Its Application in Morse Code", *frontiers*, April 2022.

[4] "Robust Cognitive Brain-inspired Computing System: Architectures and Algorithms", https://u-aizu.ac.jp/~khanh/share/pubs/ETLTC-2022.pdf.

[5] Bennie Mols, "Making AI more energy efficient with neuromorphic computing", *CWI*, March 2024.

[6] David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean, "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink", *IEEE*, July 2024.

[7] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee,Hsien-HsinS.Lee, BugraAkyildiz, MaximilianBalandat, JoeSpisak,Ravi Jain, MikeRabbat,KimHazelwood, "Sustainable AI: Environmental Implications, Challenges and Opportunities", *arXiv*, March 2024.

[8] "CMOS Power Calculation", https://resources.pcb.cadence.com/blog/2023-cmos-power-calculation.

[9] Nandita Srinivasana, Navamitha.S.Prakasha, Shalakha.Da, Sivaranjani.Da, Swetha Sri Lakshmi.Ga, B.Bala Tripura Sundari, "Power Reduction by Clock Gating Technique", *ScienceDirect*, November 2015.

[10] Wenzhe Guo, Mohammed E. Fouda, Ahmed M. Eltawil and Khaled Nabil Salama, " Neural Coding in Spiking Neural Networks: A Comparative Study for Robust Neuromorphic Systems", *frontiers*, March 2021.