Input-Layer Neuron Design for Spiking Neural Network Application

Seyed Amirhossein Nasrollahi

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Electrical Engineering) at

Concordia University

Montréal, Québec, Canada

January 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Seyed Amirhossein Nasrollahi**

Entitled: **Input-Layer Neuron Design for Spiking Neural Network Application**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Electrical Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Sébastien Le Beux

_____ External Examiner
Dr. Andrew Delong

_____ Examiner
Dr. Sébastien Le Beux

_____ Supervisor
Dr. Glenn Cowan

Approved by _____
Dr. Yousef R. Shayan, Chair
Department of Electrical and Computer Engineering

_____ 2023

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Input-Layer Neuron Design for Spiking Neural Network Application

Seyed Amirhossein Nasrollahi

Artificial Intelligence and Machine Learning algorithms help humans in various applications. Neural Networks systems are one of this area's most important research topics, inspired by the human brain. In this field, Spiking Neural Networks (SNN) use spikes to communicate between neurons mimicking the brain's algorithm. The input data produced by sensors has to be converted to spikes for training and testing these systems. Rate encoding is a popular method that encodes the input signal into the spiking frequency.

This work presents two methods to design an analog input encoder that receives the information and converts them to spiking output. Both ways use a $\Delta\Sigma$ modulator to create a digital output from the input signal. The first input encoder, called synchronous $\Delta\Sigma$ analog to spike converter, reads the digital output of the $\Delta\Sigma$ and produces a spike for every '1' bit. The second design is called neuromorphic $\Delta\Sigma$ analog to spike converter, which uses a synapse and neuron model to produce the rate-encoded spiking output. The synapse converts the $\Delta\Sigma$ output to a current, and the neuron receives this current at its input.

This thesis is the first design to build a general input encoder that can be used in most SNN systems. A clock signal can change the firing frequency of both encoders. The synchronous $\Delta\Sigma$ A-S converter can perform for clock signals between 1 kHz and 4 MHz while the neuromorphic one can perform between 1 kHz and 2 MHz. The optimized clock frequency is 50 kHz for both of them. With this clock, the synchronous one's accuracy is 99.2% encoding a DC input, and its input can have a maximum bandwidth of 120 Hz to achieve an SNR higher than 50 dB. It consumes 13.4 μW average power with 500 μm^2 area. The neuromorphic one's accuracy for DC inputs is 97.3%, and its maximum bandwidth is 65 Hz. It consumes 12.7 μW average power with 0.011 mm^2 area.

Acknowledgments

I would like to thank the following people and organizations for helping with this research project:

My supervisor, Glenn Cowan for guidance and support during my MASc program.

Concordia, FRQNT, and NSERC for research opportunities, facilities, and financial support.

Ted Obuchowicz for CAD tool support, and Terry Stewart for general advice.

and finally, my family, friends, and everyone in the Research Lab for their help in different circumstances.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Since the early stages of Artificial Intelligence (AI), the human brain has inspired many engineers. A fascinating ability of the brain is that it can handle data processing, calculating, control, classification, and cognition tasks with 86 billion neurons and trillions of synapses, consuming an average power of 20W [11, 12]. An artificial Neural Network (ANN) is a graph in which each node represents a neuron, and each connection is only a static number (weight). There are input, hidden, and output layers of neurons in the system. Each neuron receives weighted inputs from its previous layer. Then after applying a function to the sum of these inputs, it produces an output received by the next layer of neurons. Although the ANN method is significantly different from the brain's functionality, both are based on computations of neurons and their interconnectivity [13].

Our brain's energy-efficient processing is in the analog domain. An analog neural network, like the brain, processes, stores, and reads data in the same unit to save energy [14]. On the other hand, ANNs are usually built digitally using the processing power of a CPU or GPU with an external memory unit. Von Neumann's digital architectures in neural networks consume a lot of energy [15]. The challenging part is to process and train an analog system with data and update the existing data in the memory, which is much easier in digital systems [16].

Deep learning is evolved to a point where low-latency energy-efficient computation is needed for complex applications and large datasets. Therefore, traditional Von Neumann's architectures face a bottleneck in these areas [17]. Some solutions in AI have been proposed for this problem. One of these solutions is Optical Neural Networks which can be implemented on-chip with silicon or

with fiber-based optical circuitry [18]. These Optical NNs support a large bandwidth and calculate with high-speed and low power. On the other hand, there are limitations in both silicon and fiber-based Optical NNs. For instance, silicon-based ones suffer from the limitations of Optical-Electrical conversion, and fiber-based ones need a massive system occupying a large area [18].

IBM proposes another solution to the conventional Von-Neumann architectures' problem. Their TrueNorth Spiking Neural Network (SNN) chip functions with a pre-trained network's weights and consumes ultra-low power [17]. Digital synaptic weights are stored in static random access memory (SRAM) arrays on TrueNorth, which faces obstacles in training [19]. They resolved this problem using analog CMOS devices that encode the synaptic weights in their conductances [19]. Although this method's accuracy is smaller than conventional DNN systems, it speeds up the computations in training and forward path while consuming much lower power [19].

One of the most promising solutions is analog Spiking Neural Network systems that are beneficial compared to Artificial ones in terms of power efficiency, and biological plausibility [20]. They can replace the existing ANN systems with almost the same computation power in some applications while consuming much less energy. [21–24] prove that SNNs trained by MNIST dataset classify different digits with an accuracy comparable to ANN systems while consuming fewer resources. In addition, [25] shows that by using proper techniques, the SNN accuracy reaches the ANN counterparts in more sophisticated VGG and Residual Network (ResNet) architectures. Moreover, studying the learning and decision-making process in the brain's biological neural networks is possible with SNN systems [26]. These benefits of SNNs motivate this research and study, in addition to their feasibility of implementation with analog/mixed-signal silicon circuitry.

In the brain, neurons communicate by sending action potentials through the synapses. An action potential, also called a spike, is a pulse signal with narrow width. Spiking Neural Networks (SNN) are neural networks that communicate using spikes, a dynamic signal, instead of static signals used in ANNs. Building blocks of this system are neurons and synapses like our brain. Neurons receive the information and create a spiking output sent to connected neurons using synapses.

There are two popular methods of information encoding in SNN systems [21]. The first one is rate-encoding, where the information is coded in the firing rate of the neurons. In this method, the output signal's average spiking frequency is proportional to the signal's level. Figure 1.1 shows

how the input layer neurons use rate-encoding to convert the analog input information into spiking signals. The other encoding method is time encoding, where the exact timing of spike arrivals encodes the data. There are different models of temporal encoding. In most of them, the period is divided into discrete times, and the neurons can extract patterns in the arrival time of spikes, which usually applies in pattern recognition algorithms [27]. Although time encoding is utilized more in biological neurons, analyzing and implementing it on silicon is more complex than rate encoding.

The Leaky Integrate-and-Fire (LIF) model describes the mathematics of a simplified biologically plausible neuron [4]. In this model, the neuron receives its input in a current form and integrates this current. Furthermore, a leakage slowly decreases the integrated value. The output of this neuron is in the voltage domain creating a spike train that encodes the information in its firing rate.

Input layer neurons function differently than other layers because they receive analog voltages produced by sensors, whereas other layers receive spiking signals. As shown in Figure 1.1, the input layer neurons must be able to encode the sensory data into spikes with an acceptable resolution and then communicate the spikes to other layers. Since the received data are analog, the neurons' input range must match the sensors' output range. Furthermore, these neurons' spiking frequencies must change linearly with respect to input signals. Finally, the encoders' outputs must follow the changes in the input signals at an acceptable speed. Therefore, an architecture that satisfies the mentioned characteristics must be designed for the input layer neurons, which differ from other layers' structures.

This work proposes two approaches to designing input-layer analog to spike (A-S) converter neurons using a $\Delta\Sigma$ modulator. Each design is simulated and tested with different test benches in Cadence Virtuoso environment using the TSMC 65nm technology.

1.1 Input Neuron Architecture

In this work, two different input neurons are proposed. Both consist of a $\Delta\Sigma$ modulator and a spike-generating mechanism. The $\Delta\Sigma$ modulator is a popular choice in designing various types of Analog-to-Digital-Converters (ADC). They encode the analog input information in densities of '1's and '0's. The larger the input size, the higher the density of '1's. As a result, the mean value of the

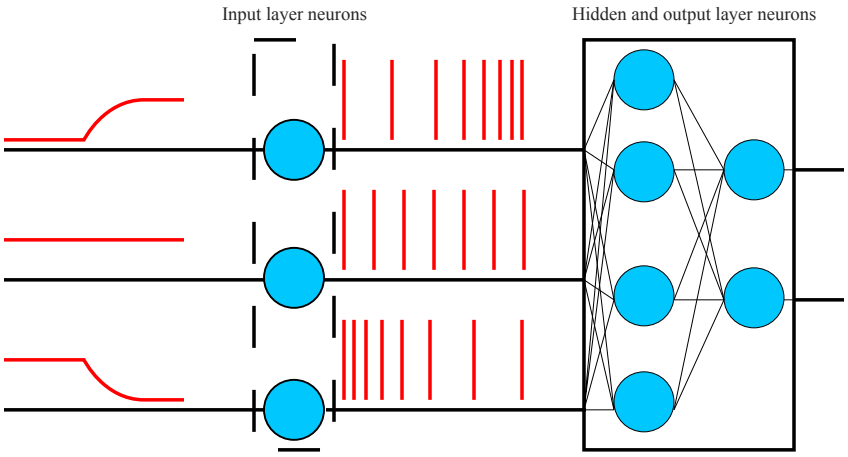

Figure 1.1: Input-layer neurons interfacing sensors and SNN

output represents the input signal. This behavior is similar to what we want as our input neuron, and the remaining part of the design is to convert this one-bit density of '1's to a density of spikes. So, a spike-generating mechanism is needed for this conversion. The difference between the two proposed architectures is the spike generator.

The first design produces spikes directly from the output signal of the $\Delta\Sigma$ encoder. It converts each '1' bit on the digital output signal of the $\Delta\Sigma$ into a spike. As a result, the density of spikes represents the density of '1's in which the input signal is encoded.

The second design produces spikes from the output signal of the $\Delta\Sigma$ using a neuromorphic spike generating circuit. This system uses the same neuron and synapse design as other layers of neurons and includes them after a $\Delta\Sigma$ encoder. The synapse is responsible for receiving the digital output signal of the $\Delta\Sigma$, then creating a current that encodes this digital signal. The neuron receives the encoded current produced by the synapse and fires spikes based on this current. Hence, having a higher '1' density at the $\Delta\Sigma$ output, which encodes the input signal, results in a larger spiking frequency at the neuron's output.

1.2 Thesis Organization

Chapter 2 provides the background and literature review, which discusses the theory of spiking neuromorphic models and $\Delta\Sigma$ converters.

Chapter 3 describes the design methodology and the architectures designed in this work.

Chapter 4 shows the simulation results of two designed neurons and compares them.

Chapter 5 concludes the work and presents further improvements that can be made to the design.

1.3 Contribution

In neuromorphic circuits, it is better if sensory interface, computation, and memory are implemented on a single chip [28]. The input interface is a crucial part of every SNN system. There are limitations in the designed circuits that are supposed to receive the data from the sensory system and convert it to spikes. The neurons' input data is a current. A standard method is to convert the data into spikes outside the chip and feed those spikes into the SNN system [24]. Another common way is that each system has a specific encoding circuit that can only be used on that particular SNN system like a high-frequency oscillation detection system in [10] that only encodes the changes in input. This work's main contribution is to introduce a new architecture for input layers of neurons. It utilizes the prior designs of the $\Delta\Sigma$ encoder, neuron, and synapse and combines them with some modifications to operate as an input encoder. It can be used in every SNN system to encode the input sensory data into spiking patterns.

A benefit of using the $\Delta\Sigma$ encoder is that it comes with a clock frequency set by the designer to fulfill the system's requirements. The average firing frequency in an SNN system can vary from biological frequency ($\sim$ 100 Hz) up to tens of MHz. The maximum firing frequency required by the system is determined by the clock frequency of the $\Delta\Sigma$ encoder. If the clock frequency changes, nothing needs to be changed in the circuit except for the capacitor values to adapt to the new firing frequency. Therefore, this design can be used in different frequency applications.

This work can replace any voltage-to-current (V-I) converters used in the input of the SNN system to read the sensory data. The current feeding into the neuron is very low to keep the transistors of the neuron at sub-threshold. A V-I converter that creates a sub-threshold current must consume a large area because of the large resistors and transistors or lose conversion accuracy. This input-layer neuron uses the one-bit output of the $\Delta\Sigma$ to create a spiking output. Although its building blocks are not the same as other layers, it does not consume a larger area than other neurons and can provide

much better accuracy than any V-C converters.

$\Delta\Sigma$ encoder is well-known in ADC design with great conversion accuracy. The resolution of the $\Delta\Sigma$ can be preserved while creating spiking voltage from its output. Both input neurons encode the input signal with accuracy higher than 97% in their optimized frequency using a $\Delta\Sigma$ encoder. We can use this ADC in a wider variety of applications related to neuromorphic encoding and achieve the same level of data encoding accuracy as ADCs.

This work is the first analog/mixed-signal architecture in designing encoders for the input data of any SNN systems by using the basic structures of the $\Delta\Sigma$ encoder and the Integrate-and-Fire (IF) neuron model. The result of this work is shown with different types of input signals in accuracy, linearity error, power, area, input signal bandwidth, and Signal-to-Noise Ratio (SNR). In the future, researchers can investigate what improvements can be made in each criterion by changing the loop-order (number of integrators) or design of the $\Delta\Sigma$ encoder. Moreover, all neuron models, such as adaptive LIF [29], Izhekevich [30], AdExpIF [31], ... can be tested in the framework of the second proposed architecture, and their pros and cons be evaluated.

The results of this work are also published at the IEEE NEWCAS 2022 conference [32].

Chapter 2

Literature Review

This work presents two architectures that encode the input signal for SNN systems. The building blocks are $\Delta\Sigma$ modulator, silicon neuron, and synapse.

2.1 Silicon Neurons

2.1.1 Biological Neuron Model

In ANN systems, neurons have a static processing model, whereas, in SNNs, the neurons communicate with each other using spike signals. The brain inspires this method of communication. In the brain, a neuron's inside is separated from the outside of it by a membrane. A sudden increase in the membrane potential and resetting to the initial value after a short period is called an action potential or a spike. A voltage over a capacitor in the neuron can model the membrane potential. Hodgkin and Huxley's neuron model is the first attempt to calculate neurons' dynamic behavior by measuring the flow of ions between the inside and outside of a membrane [1, 33]. As shown in Figure 2.1, voltage-dependent resistors model the ion channels, and an independent resistor models the leakage of the neuron. A voltage source is set for the resting potential of each channel.

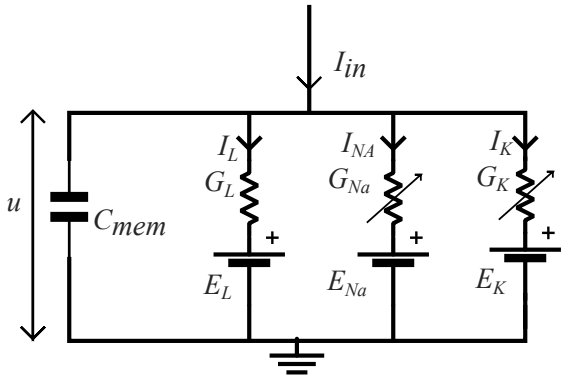

Figure 2.1: Circuit model of Hodgkin-Huxley neuron model [1].

Mathematical equations of (2.1) describe the behavior of the electronic circuit shown in Figure 2.1:

$$\begin{aligned} I_{in}(t) &= I_{C_{mem}}(t) + \sum_j I_j(t) \\ C_{mem}\frac{du}{dt} &= -\sum_j G_j(u - E_j) + I_{in}(t) \end{aligned} \tag{2.1}$$

where I_j, G_j, and E_j represent each branch's current, conductance, and voltage source values.

The objective of Hodgkin and Huxley was to measure the changes of channel resistors with respect to time and the membrane voltage [34]. In their work, sodium (Na) and potassium (K) conductances are voltage-dependent, where $G_{Na} = g_{Na}m^3h$ and $G_K = g_K n^4$, whereas G_L is a constant. The changes of these resistors to time and voltage are modeled in constant variables g_{Na} and g_K, and probability variables m, n, and h that are between 0 and 1 and alter the channel resistors in time. (2.2) describes the dynamics of these probability variables:

$$\frac{dx}{dt} = -\frac{1}{\tau_x(u)}[x - x_0(u)] \tag{2.2}$$

where x has a value between 0 and 1 and can be replaced by m, n, and h. τ_x is its time constant, and x_0 is its steady-state value which are shown in Figure 2.2. The Hodgkin and Huxley neuron model is described in 4 dimensions with 4 differential equations calculating $\dot{u}, \dot{m}, \dot{h}$, and $\dot{n}$ shown in (2.1) and (2.2). Detailed analysis of these equations is in [1].

The simple description of this 4-dimensional equation is that the input current is integrated into

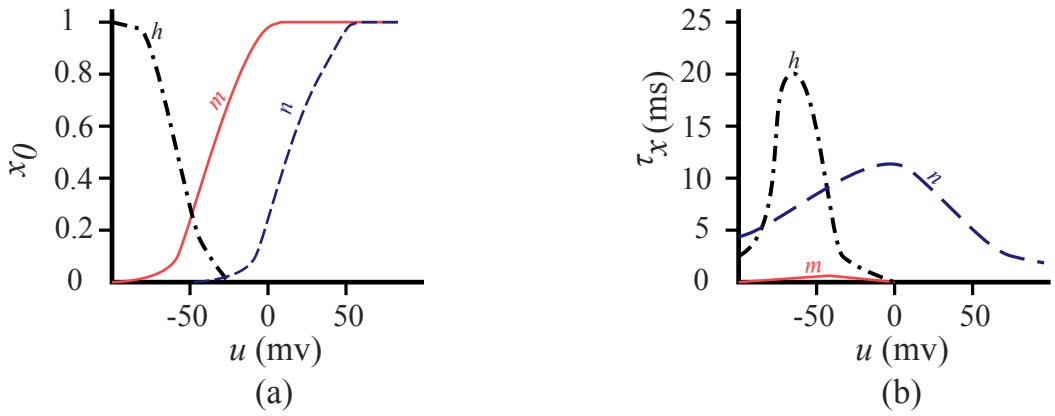

Figure 2.2: Steady-state value and time constant of m, n, and h with respect to the membrane potential in the Hodgkin-Huxley neuron model calculated in [2]

the membrane capacitor. The sodium channel conductance is $G_{Na} = g_{Na}m^3h$, which changes with the membrane potential, u. As shown in Figure 2.2b, the time constant of m is tiny. Therefore, we can assume that the change in m is instantaneous. When membrane potential increases, m rapidly changes and increases G_{Na}; therefore, I_{Na} also rises. Since $u < E_{Na} = 55$ mv [33], the rise of I_{Na} increases the membrane potential known as the depolarization phase in Figure 2.3. Afterward, with a rise of u, h decreases and starts to close the Na channel slowly. Then, the potassium channel, $G_K = g_K n^4$, opens when n increases. Since $\tau_n > \tau_m$ and $u > E_K = -77$ mv, the K channel's effect appears later than the Na channel, and its current decreases the membrane potential. This effect is known as the repolarization phase, shown in Figure 2.3. When $u < E_L = 65$ mV, known as the hyperpolarization phase, $m = n = 0$ and therefore both Na and K channel are closed, and G_L resets the membrane potential to its resting voltage. Figure 2.3 shows this entire process of producing an action potential or spike resulting from the mentioned differential equations.

Implementing Hodgkin and Huxley neuron model on silicon for computational purposes is complex and inefficient. There is a trade-off between biological plausibility and computation cost in SNN systems. Integrate-and-Fire (IF) and Izhekevich [30] neurons describe the spiking model of neurons in an uncomplicated way with a fixed threshold, which can help us for a simpler and cheaper circuit simulation of neurons [33].

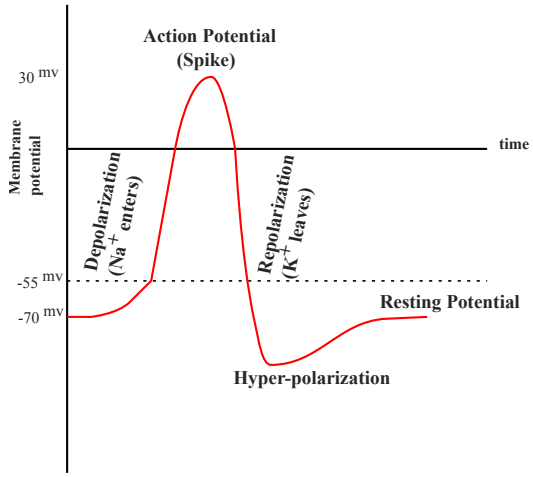

Figure 2.3: The entire process of creating a spike.

2.1.2 Integrate-and-Fire Neuron Model

In the Integrate-and-Fire (IF) model, a biological neuron's four-dimensional dynamic equations are reduced to one equation and a resetting mechanism. In general, the IF neuron is described as equation (2.3) [33]:

$$\tau \frac{du}{dt} = R(u)I_{in} + f(u)$$
$$\text{when } u \geq u_\theta \Rightarrow \text{spike and } u \to u_{rest} \tag{2.3}$$

where τ is the time constant, $R(u)$ is the leakage resistor model, $f(u)$ is the neuron model's function, u_θ is the threshold voltage and u_{rest} is the membrane's resting potential. In the linear IF model, known as Leaky IF (LIF), $R(u)$ is a constant resistor, and $f(u)$ is linear as (2.4):

$$\tau \frac{du}{dt} = RI_{in} - (u - u_{rest})$$
$$\text{when } u \geq u_\theta \Rightarrow \text{spike and } u \to u_{rest} \tag{2.4}$$

where $R(u)$ and $f(u)$ are replaced by a constant resistor R and the linear function $-(u - u_{rest})$.

In electronic circuits, we usually want to imitate the behavior of biological neurons for computational purposes. We have to scale the biological values for better computational accuracy on electronic circuits. In biological neurons, the threshold and resting potential are around -55 mV and -70 mV and the spike peaks at 30 mV as shown in Figure 2.3. First, we defined an output

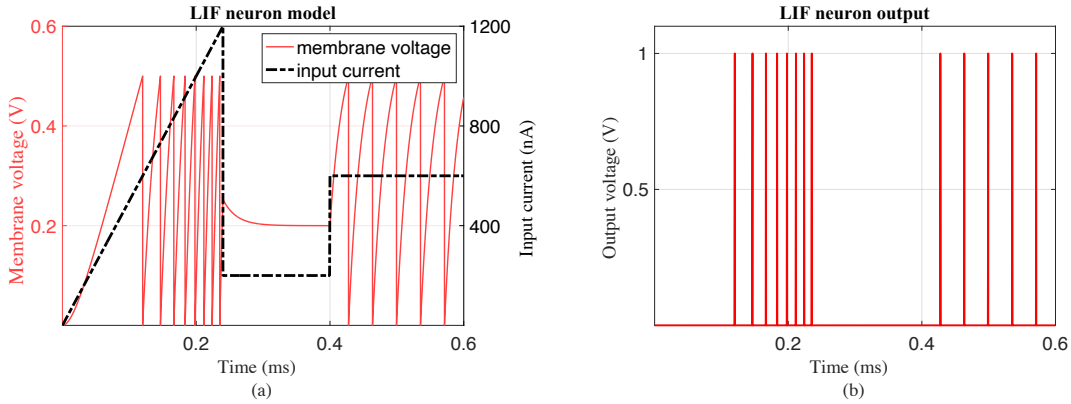

Figure 2.4: **a)** The input current and the membrane potential in the LIF differential equation simulated in MATLAB **b)** The resulting output spiking voltage.

voltage of V_{out} that changes from 0 V to V_{DD} at each spike time and resets after a delay. Then, the resting potential is set to 0 V, and the threshold is changed to $V_{DD}/2$. The time constant and the leakage resistor is set to 20 μs and 1 MΩ, respectively resulting in a membrane capacitance of 20 pF and achieving kHz range spiking frequency. With this setup, the membrane potential and output voltage are shown in Figure 2.4.

When we increase the input current, the neuron fires more rapidly, as expected from the differential equations of (2.4). If the input current is below I_θ, the membrane voltage settles at a voltage below the threshold, and the neuron does not fire spikes. As a result, a threshold current is calculated in (2.5):

$$\tau \frac{du}{dt} = 0 \Rightarrow RI_{in} = u - u_{rest} \Rightarrow I_\theta = \frac{u_\theta - u_{rest}}{R} \tag{2.5}$$

where I_θ is the firing threshold current which is 500 nA for the setup of Figure 2.5. The firing frequency of the neuron rises linearly with respect to the input current increase if the current is above its threshold. The output firing frequency with respect to the input current is shown in Figure 2.5.

2.1.3 IF Neuron Implementation

The IF neuron is popular in silicon neurons because of its simplicity and efficiency [3]. We need an integrator, a threshold comparator, and a reset circuit to design an IF neuron. One of the simplest linear IF neuron circuits is shown in Figure 2.6 with a non-linear resetting loop [35]. First,

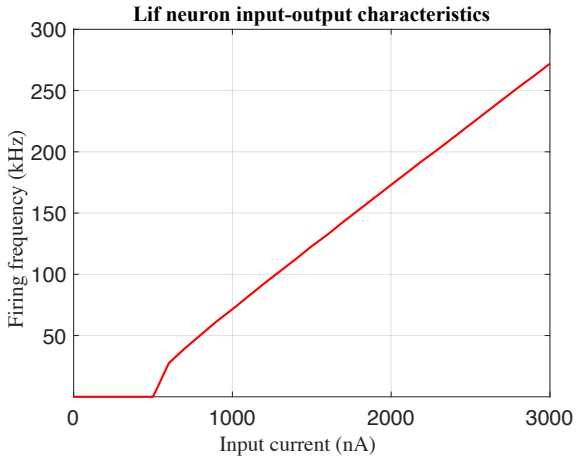

Figure 2.5: Output firing frequency of an LIF neuron to DC input current. (I_θ=500 nA)

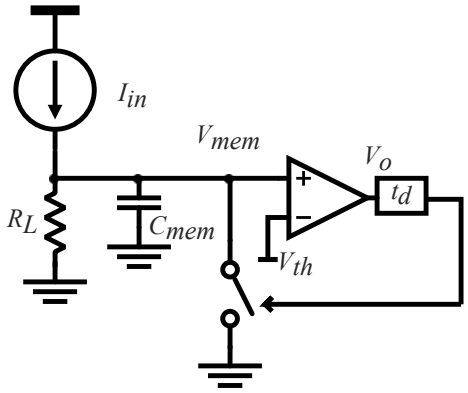

Figure 2.6: LIF neuron circuit, implementing equation (2.4) [3].

a capacitor, C_{mem}, integrates the input current, I_{in}, producing the membrane voltage, V_{mem}. Then, a comparator turns on its output whenever the membrane voltage hits a threshold, V_{th}. Then after a delay of t_d, which is the duration of a spike, a switch resets the membrane voltage to its resting value. The differential equation and resetting mechanism describing this process are as below:

$$\begin{gathered} C_{mem}\frac{dV_{mem}}{dt} + \frac{V_{mem}}{R_L} = I_{in} \xrightarrow{R_L \times C_{mem} = \tau} \tau\frac{dV_{mem}}{dt} = R_L I_{in} - V_{mem} \\ \text{If } V_{mem} \geq V_{th} \Rightarrow V_o = V_{DD} \xrightarrow{\text{after a delay of } (t_d)} V_{mem} = 0 \Rightarrow V_o = 0 \end{gathered} \tag{2.6}$$

where all of the variables are shown in Figure 2.6.

The IF neuron model can be designed in sub-threshold to save power and improve biological

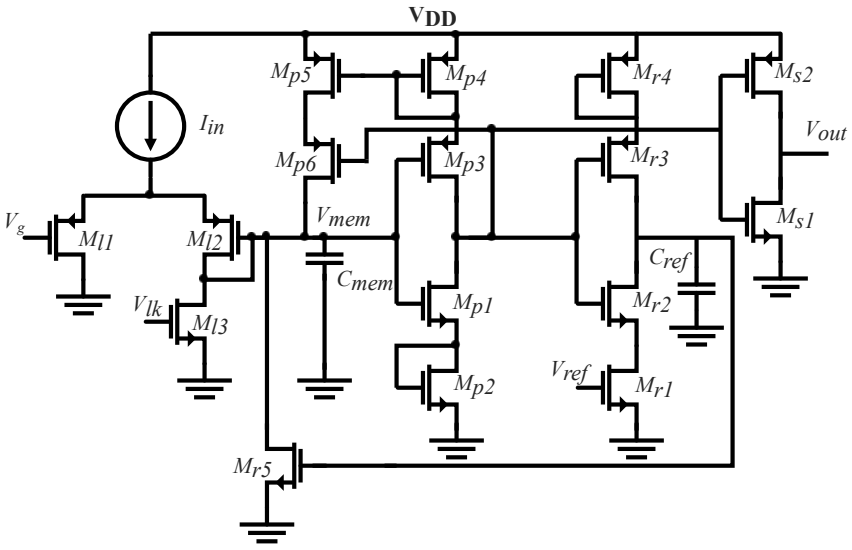

Figure 2.7: Complete LIF neuron circuit with positive feedback and reset current from [4].

plausibility [4] as shown in Figure 2.7. A differential-pair integrator (DPI) with leakage receives the input current in this LIF model. Then, positive feedback is applied to decrease the rise time of the membrane potential, by M_p transistors, which the simple IF neuron of Figure 2.6 does not include it. Furthermore, a current-starved inverter, implemented by M_r transistors, replaces the reset circuit to implement the delay and activates a switch to reset the membrane voltage. Therefore, the LIF neuron of Figure 2.7 is more biologically plausible because it resembles the sodium and potassium current of Hodgkin-Huxley [1] model with the positive feedback current and the resetting circuit, respectively.

There are four different parts in Figure 2.7. The differential pair integrator transistors are labeled M_l. In this integrator, V_g, V_{lk}, and C_{mem} control the gain, leakage, and time constant, respectively. After the integration, M_p transistors compare V_{mem} to the threshold of the inverter, $V_{DD}/2$. When V_{mem} increases, the positive feedback current of M_{p6} charges the membrane voltage faster. Then, M_r transistors reset the membrane voltage. In this setup, C_{ref} and V_{ref} set the time of the refractory period, which controls the minimum Inter-Spike Interval (ISI) duration. In the end, M_s MOSFETs produce a spike from 0 V to V_{DD} whenever the membrane hits the threshold. The detailed analysis is in [29].

By assuming that the LIF circuit is not in the reset state and $I(M_{r5}) = 0$, the differential

equation in the membrane node of this neuron's circuit is as below:

$$C_{mem}\frac{dV_{mem}}{dt} = I_{dpi} - I_{lk} + I_{fb} \tag{2.7}$$

where $I_{dpi} = I(M_{l2})$, $I_{lk} = I(M_{l3})$, and $I_{fb} = I(M_{p6})$. The leak current, I_{lk}, is constant and controlled by V_{lk}. The DPI's current is as below [29]:

$$I_{dpi} = I_{in}\frac{e^{\frac{\kappa(V_{DD}-V_{mem})}{U_T}}}{e^{\frac{\kappa(V_{DD}-V_{mem})}{U_T}} + e^{\frac{\kappa(V_{DD}-V_g)}{U_T}}} = \frac{I_{in}}{1+e^{\frac{\kappa(V_{mem}-V_g)}{U_T}}} \tag{2.8}$$

where κ, U_T, V_{mem}, and V_g are the sub-threshold slope factor, the thermal voltage, the membrane potential, and the integration gain voltage. (2.8) shows that the DPI's current is a fraction of I_{in} depending on the membrane voltage and the gain voltage. When V_{mem} is increased to values larger than V_g, I_{dpi} decreases. By defining I_{mem}, the positive feedback current is as below [29]:

$$I_{mem} = I_0 e^{\frac{\kappa V_{mem}}{U_T}} \Rightarrow I_{fb} = I_0^{\frac{1}{\kappa+1}} I_{mem}^{\frac{\kappa}{\kappa+1}} \frac{1}{1+e^{-\alpha(I_{mem}-I_{th})}} \tag{2.9}$$

where α and I_{th} are related to the gain and switching point of the inverter controlled by the layout and process technology. When V_{mem} is small, $I_{mem} \approx 0$, resulting in $I_{fb} \approx 0$ and $I_{dpi} \approx I_{in}$. After replacing them in (2.7) we have:

$$C_{mem}\frac{dV_{mem}}{dt} = I_{in} - I_{lk} \tag{2.10}$$

where C_{mem}, I_{in}, and I_{lk} are the membrane capacitor, input current, and the leakage of the DPI, respectively. Since I_{lk} is the leakage current and an equivalent of $\frac{V_{mem}}{R_L}$ in (2.6), (2.10) resembles the LIF's original differential equation. When V_{mem} increases, I_{fb} becomes dominant, and it charges the membrane capacitor until it reaches the threshold of the inverter, which makes the neuron spike and then reset.

The circuit of Figure 2.7 is simulated in the TSMC 65nm technology, and Figure 2.8 shows the result with the default parameters set in Table 2.1. First, we change the input current, I_{in}, from 0 to 300 nA as a ramp. Then after resetting it to 0 A, we apply a step current to 200 nA. The

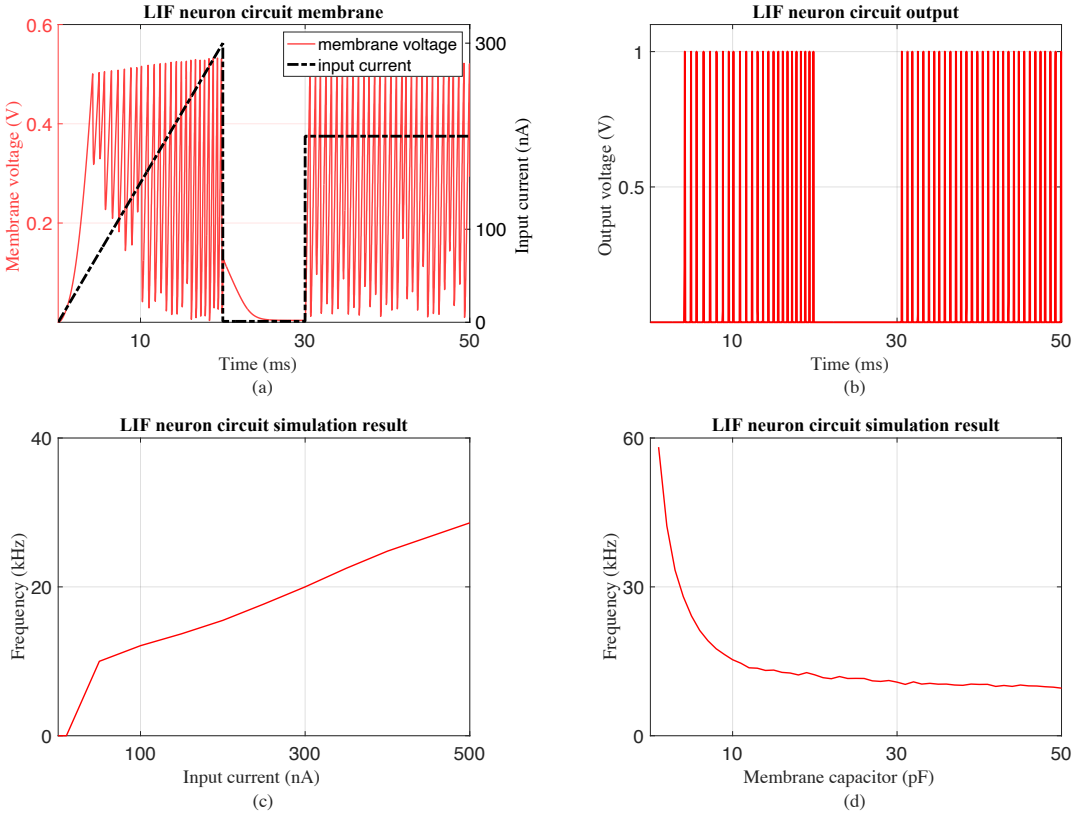

Figure 2.8: **a)** LIF neuron's circuit membrane voltage response to the ramp and step input **b)** LIF neuron's spiking output to the same input **c)** Output firing frequency of an LIF neuron's circuit to the DC input current ($I_\theta = 10$ nA) **d)** Output firing frequency to the membrane capacitor's value.

membrane potential and spiking output are plotted in Figures 2.8a and 2.8b. We can realize that the firing frequency changes with the input current as expected. We expect firing frequency to increase linearly after the current threshold with input current increase. The result is depicted in Figure 2.8c. There is a 6% linearity error in this figure caused by the DPI circuit. If the input current changes between 50 nA and 500 nA, the firing frequency increases from 10 kHz to 30 kHz. The threshold current in this setup is around 10 nA. Moreover, Figure 2.8d shows how the frequency changes with respect to the membrane capacitor when other parameters are set as their default given in Table 2.1.

Table 2.1: Control parameters of the differential-pair integrator LIF neuron

Parameter	I_{in}	V_g	V_{lk}	V_{ref}	C_{mem}	C_{ref}	V_{DD}
Default Value	100 nA	700 mV	250 mV	250 mV	20 pF	1 pF	1 V

There are some differences between simple LIF circuit of Figure 2.6 and the DPI LIF circuit of

Figure 2.7. First, the leakage of the DPI circuit is built with low-current sub-threshold transistors that can save area significantly compared to R_L. Secondly, in the DPI, V_g controls the gain of the integration. Finally, with positive feedback, the DPI circuit resembles the biological neurons better. On the other hand, there is a small non-linearity in the DPI circuit's rate-encoding, which is not seen in the circuit of Figure 2.6. If we assume that the comparator of Figure 2.6 is an inverter as Figure 2.7, the DPI IF neuron consumes a little more power because of the current-starved inverter and the addition of positive feedback current. Therefore, the DPI neuron's benefits cost dissipating more power and losing linearity.

2.2 Silicon Synapses

Synapses are responsible for connecting neurons and communicating spike signals between them. A synapse receives spike voltages from the output of its pre-synaptic neuron. It filters the incoming spiking signal and produces a current based on a weight value. Then, the synapse feeds this weighted current to its post-synaptic neuron's membrane. Moreover, synapses play an essential role in learning. While implementing a vast number of synapses on software and applying learning rules can be expensive in terms of power and time, using hardware for these methods and doing parallel computations can reduce costs [6].

Low-pass filtering of the incoming spikes produces an average value related to the firing frequency of its pre-synaptic neuron. The low-pass filtering exists in the synapse to moderate the sudden changes in the spiking current received from the pre-synaptic neuron.

Simple synapses consist of two transistors to create a current based on an adjusted weight, such as Figure 2.9a [5]. This synapse model creates a spiking current based on the spiking input voltage, and the post-synaptic neuron integrates this current. In this circuit, V_w controls the magnitude of the output current and V_{in} is an active low spiking input. Since V_{in}'s low value is 0 V, the maximum V_{GS} is applied to this transistor, forcing it to operate as a switch. Hence, it is assumed that approximately $V_x = V_{DD}$, when a spike is received. The equation below calculates the output

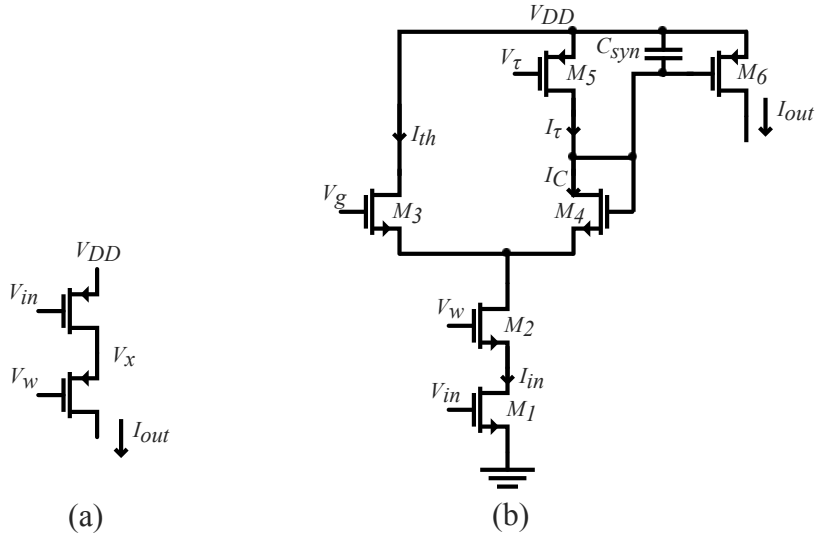

Figure 2.9: **a)** Simple two transistor synapse circuit [5] **b)** Differential-pair integrator synapse circuit [6].

current when a spike arrives at V_{in}:

$$\text{when } V_{in} = 0 \Rightarrow V_x = V_{DD} \Rightarrow I_{out} = I_0 e^{-\frac{\kappa}{U_T}(V_w - V_{DD})} \tag{2.11}$$

where V_x is the voltage node between the two transistors, I_{out} is the produced current, and I_0, κ, and U_T are determined by the process technology. V_w controls the output current's magnitude, and learning in SNNs is due to changing V_w.

The differential-pair integrator (DPI) synapse is shown in Figure 2.9b, and designed in [6] to create the post-synaptic current after low-pass filtering of the input spikes. We can use current-mode analysis, as we did in (2.7) for the DPI neuron, to analyze the behavior of these synapses:

$$\begin{gathered} I_{in} = I_D(M_1), I_{th} = I_D(M_3), I_C = I_D(M_4), I_\tau = I_D(M_5) \Rightarrow I_{th} + I_C = I_{in} \\ \text{and } I_C = I_{in} \times \frac{I_C}{I_{in}} = I_{in} \times \frac{I_C}{I_C + I_{th}} = \frac{I_{in}}{1 + \frac{I_{th}}{I_C}} \end{gathered} \tag{2.12}$$

where I_{th} and I_C are sub-threshold currents controlled by the gate voltage of M_3, V_g, and the gate voltage of M_4, V_C. The following equations replace these currents with their sub-threshold

equations:

$$I_C = I_0 e^{\frac{\kappa V_C}{U_T}}, I_{th} = I_0 e^{\frac{\kappa V_g}{U_T}} \Rightarrow \frac{I_{th}}{I_C} = e^{\frac{\kappa(V_g - V_C)}{U_T}} \tag{2.13}$$

By replacing the result of (2.13) in (2.12) we have:

$$\rightarrow I_C = \frac{I_{in}}{1 + e^{\frac{\kappa(V_g - V_C)}{U_T}}} \tag{2.14}$$

where we can define the output current, I_{out}, and I_g as below and rewrite I_C as a function of I_{out} and I_g:

$$I_{out} = I_0 e^{-\frac{\kappa(V_C - V_{DD})}{U_T}} \text{ and } I_g = I_0 e^{-\frac{\kappa(V_g - V_{DD})}{U_T}} \Rightarrow I_C = \frac{I_{in}}{1 + \frac{I_{out}}{I_g}} \tag{2.15}$$

where I_g is the drain current of a hypothetical PMOS transistor with the gate voltage of V_g, and I_{out} is shown in Figure 2.9b.

We can write the derivative of the output current to define the differential equation as (2.16):

$$\begin{aligned} I_{out} = I_0 e^{-\frac{\kappa(V_C - V_{DD})}{U_T}} \Rightarrow \frac{d}{dt} I_{out} &= I_0 e^{-\frac{\kappa(V_C - V_{DD})}{U_T}} \times \frac{d}{dt}\left(-\frac{\kappa(V_C - V_{DD})}{U_T}\right) \\ \frac{d}{dt} I_{out} &= I_{out} \times \left(-\frac{\kappa}{U_T}\right) \frac{d}{dt} V_C \end{aligned} \tag{2.16}$$

where $\frac{d}{dt} V_C$ can be calculated from the current of C_{syn} in Figure 2.9b:

$$C_{syn} \frac{d}{dt} V_C = I_\tau - I_C \Rightarrow \frac{d}{dt} V_C = \frac{I_C - I_\tau}{C_{syn}} \tag{2.17}$$

where I_C and I_τ are shown in (2.12). I_C is calculated in (2.15), and $\frac{d}{dt} V_C$ is calculated in (2.17). The derivative of the output current calculated in (2.16), and the resulting differential equation are as the following:

$$\frac{d}{dt} I_{out} = \left(\frac{\kappa I_\tau}{U_T C_{syn}}\right)\left(\frac{\frac{I_{in}}{I_\tau}}{1 + \frac{I_{out}}{I_g}} - 1\right) \times I_{out} \xrightarrow{\text{if } \tau = \frac{U_T C_{syn}}{\kappa I_\tau}} \tau \frac{d}{dt} I_{out} + I_{out} = \frac{\frac{I_{in}}{I_\tau}}{\frac{1}{I_{out}} + \frac{1}{I_g}} \tag{2.18}$$

where τ is the time constant of the filtering. If $V_g > V_C$, the differential equation of the filter can be simplified as below:

$$\text{If } I_{out} >> I_g \Rightarrow \tau \frac{d}{dt} I_{out} + I_{out} = \frac{I_{in} I_g}{I_\tau} \tag{2.19}$$

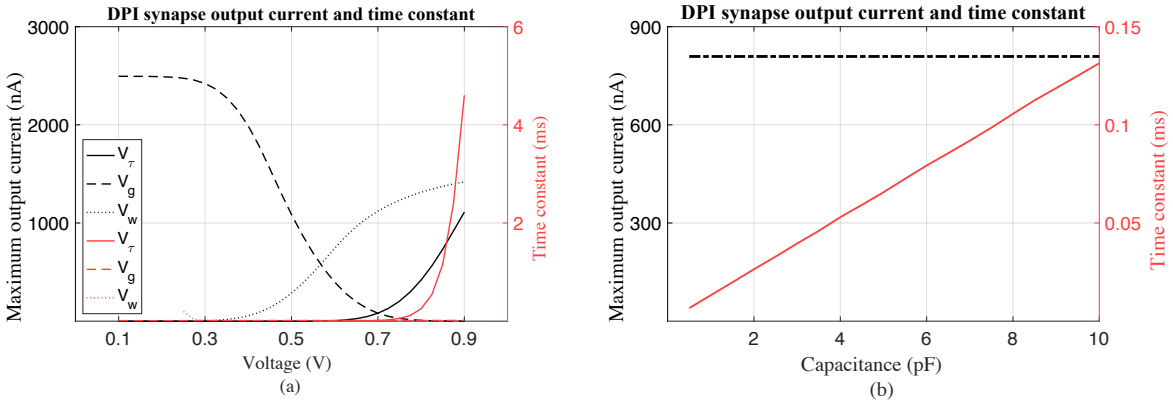

Figure 2.10: The maximum output current and time constant of the filter for different values of: **a)** V_τ, V_g, and V_w **b)** C_{syn} (In both figures, the black lines show the change in the maximum output current, and the red lines show the change in the time constant of the filter, which is $1/f_{3dB}$).

where it is the differential equation of a first-order low-pass filter.

If we assume that input spikes come at t^+, the final solution of (2.19) for the output current is as below [6]:

$$\begin{aligned} \text{rise } &\Rightarrow I_{out}(t) = \frac{I_{in} I_g}{I_\tau}(1 - e^{-\frac{t-t^+}{\tau}}) + I_{out}(t^+)e^{-\frac{t-t^+}{\tau}} \\ \text{decay } &\Rightarrow I_{out}(t) = I_{out}(t^+)e^{-\frac{t-t^+}{\tau}} \end{aligned} \tag{2.20}$$

where $I_{out}(t^+)$ is the output current's value in spike arrival time. The time constant of this filter depends on the value of the capacitor and V_τ. In this filter's circuit in Figure 2.9b, the values of V_g and V_w control the filter gain. The maximum output current and time constant of the filter are shown for default values of Table 2.2, and sweeping values of V_τ, V_g, V_w, and C_{syn} in Figure 2.10. As shown in the figure, V_g and V_w have the most effect on the filter's gain, as expected from the differential equation. Moreover, V_τ, and C_{syn} dominantly affect the filter's time constant.

Table 2.2: Control parameters of the DPI synapse

Parameter	V_W	V_g	V_τ	V_{DD}	C_{syn}
Default Value	420 mV	700 mV	700 mV	1 V	1 pF

The default values in Table 2.2 are set for the synapse circuit. Figure 2.11a shows the filter's frequency response. The output current's magnitude is plotted to an ac input current source that replaces I_{in} in Figure 2.9b. The low-pass filtering of DPI synapse is shown in Figure 2.11 with 3 dB loss frequency of 83 kHz. In Figure 2.11b, the applied input is spiking signals in different

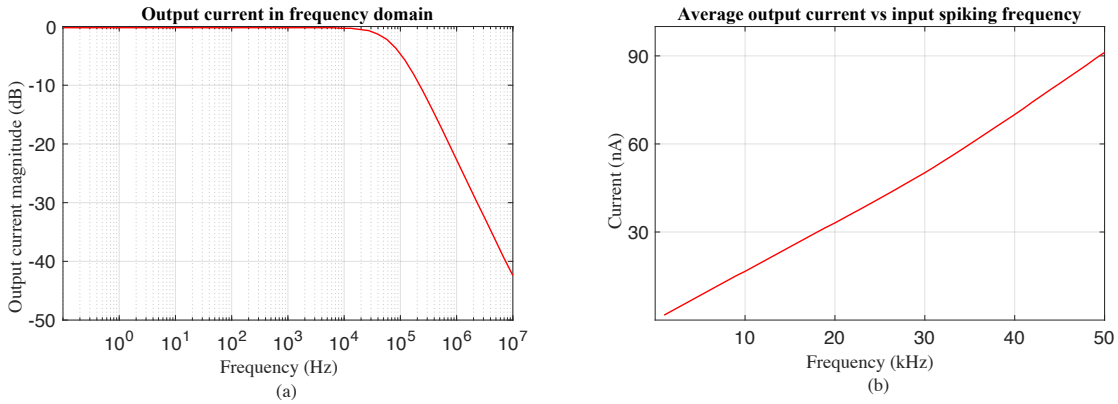

Figure 2.11: **a)** The magnitude of the output current with AC input in different frequencies **b)** The average of output current with spiking input in different firing frequencies.

frequencies from 1 kHz to 50 kHz. The mean value of the output spiking current is measured with the mentioned input signal. The average value of the output current and input voltage's spiking frequency have a linear relationship. I_{out} of this synapse is received by the post-synaptic neuron's membrane capacitor, which converts the current to a membrane voltage by integrating it.

2.3 Spiking Neural Network System

Significant progress has been made in the Artificial and Deep Neural Networks field in the last decade. Although ANN systems achieved excellent results, they are resource-intensive in terms of power consumption, data requirement, and computational cost. In applications where real-time data monitoring and controlling is needed, especially battery-powered ones, biologically plausible Spiking Neural Networks benefit from the power and computational cost efficiency [20].

An issue in SNN systems is training models. A well-known model for training Artificial and Deep NNs is the stochastic gradient descent method with back-propagation [36]. On the other hand, spike operations are non-differentiable, and neurons have complex dynamics. Therefore, training SNNs are more difficult compared to ANN systems [20]. Because of this training problem, SNN applications have been limited to simple issues such as digit classifications [37]. Promising results are shown in [25], where with the proper training method that maps trained weights of ANN to

SNN, the error of the SNN system becomes almost as small as the counterpart ANN in more complex applications of VGG and Residual Networks architectures while consuming fewer resources. Hence, it is possible to use SNN in a wider variety of applications and achieve comparable results to ANN [25].

Another issue that SNNs are not widely used in industrial applications is that they lack a good common input encoding architecture [38]. One method of encoding is time encoding. Since the neurons' output have temporal dynamics, the neurons can extract pattern from the spike arrival time [27]. The other encoding method, which is more popular, is rate encoding, where the density of spikes in a time window represents the input information [38]. In one type of rate encoding, the average spiking frequency fired by a neuron is proportional to the input as follows:

$$\text{input} \propto f_{spike}^{avg} = \frac{N_{spike}}{t} \tag{2.21}$$

where f_{spike}^{avg}, N_{spike}, and t are the average spiking frequency, the number of spikes, and the time window, respectively. IF neurons inherently convert the input current into a spiking output using rate encoding. On the other hand, the input data received from sensors are usually in voltage form in SNN systems. This voltage value must convert to a spiking signal and be received by the hidden layer of the NN system. Thus, the design of the input layer neurons is different from other layers.

[23] evaluates different input encoding methods in SNN systems for digit classification applications. The MNIST dataset used in [23] consists of digits' images in black and white, and non-linear encoding seems sufficient for this dataset [21]. Conversely, in [23], it was shown that 8-bit variable-rate linear input encoding improves the network's accuracy up to 9% compared to 1-bit fixed-rate encoding in digit classification. As a result, we can assume that increasing the resolution of input encoding improves the accuracy of tests significantly with more complex inputs compared to the MNIST dataset. This work's goal is to design and simulate input neurons that can support an 8-bit resolution of rate encoding which can be used in different SNN systems and improve their accuracy. The rest of this chapter covers input encoding methods for SNN systems in the following sections and evaluates their usefulness.

2.4 $\Delta\Sigma$ Encoder

A $\Delta\Sigma$ encoder is an architecture usually used in ADCs to convert an analog input signal into a bit-stream, as shown in Figure 2.12. It uses a clock signal that samples the input at a higher frequency than the Nyquist rate. This method is called oversampling. High converting accuracy despite having a one-bit output in $\Delta\Sigma$ ADCs is because of the oversampling technique. Oversampling ratio is the ratio of sampling frequency to the Nyquist frequency. It allows the quantization noise to spread across higher frequencies. If the oversampling ratio is M, the output signal after filtering has M times less in-band noise [7]. The complete analysis of the $\Delta\Sigma$ system and its circuit implementation is discussed in this section.

2.4.1 $\Delta\Sigma$ System

The block diagram of a simple first-order $\Delta\Sigma$ modulator is provided in Figure 2.13. The encoder consists of an integrator, a quantizer, and a DAC [7, 39]. First, the input signal $u(n)$ gets integrated. The quantizer receives the integrated signal and converts it into a single-bit output by comparing it to a threshold voltage. At this step, the quantizer produces a quantization noise, $q(n)$. The quantization noise is the difference between the digital voltage level at the output of the quantizer, $y(n)$, and its input analog voltage. This output, $y(n)$, is converted to analog and subtracted from the original input signal. $y(n)$ is the single-bit digital output sampled at f_s that is shown in Figure 2.13. When the sampling frequency is much higher than the signal's bandwidth, the difference between two consecutive samples of $u(n)$ and $u(n-1)$ is very small. Therefore, the produced output quantization noise is close to its previous sample's noise and will be subtracted from the input. As a result of the oversampling technique, the noise at $y(n)$ is small in low frequencies and shaped at much higher frequencies than the original input bandwidth. (2.22) shows the signal (STF) and noise

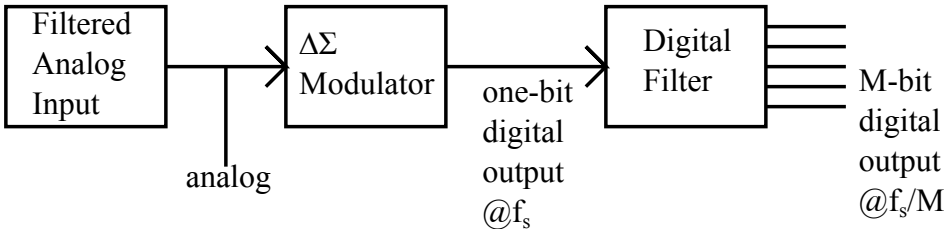

Figure 2.12: $\Delta\Sigma$ system converting an analog signal to digital [7].

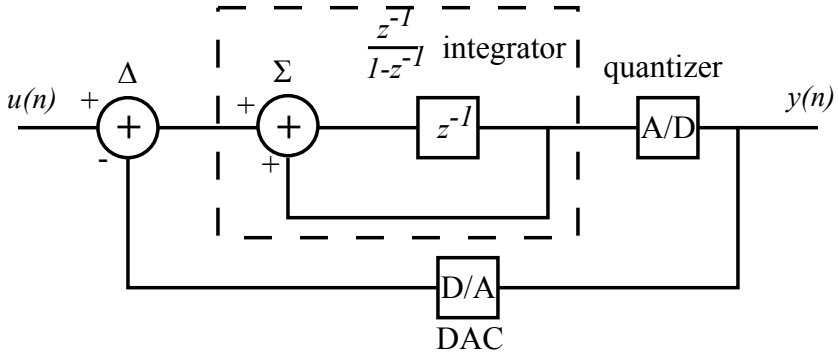

Figure 2.13: $\Delta\Sigma$ encoder block diagram [7].

(NTF) transfer functions:

$$
\begin{aligned}
Y(z) &= (U(z) - Y(z)) \times \frac{z^{-1}}{1 - z^{-1}} + Q(z) \\
\text{STF} &= \frac{Y(z)}{U(z)} = \frac{\frac{z^{-1}}{1-z^{-1}}}{1 + \frac{z^{-1}}{1-z^{-1}}} = z^{-1} \\
\text{NTF} &= \frac{Y(z)}{Q(z)} = \frac{1}{1 + \frac{z^{-1}}{1-z^{-1}}} = 1 - z^{-1}
\end{aligned}
\tag{2.22}
$$

where $U(z), Q(z)$ and $Y(z)$ are the z transformed signals of the $u(n), q(n)$ and $y(n)$. The signal is only delayed by a sample, z^{-1}. On the other hand, the noise transfer function is given by $1 - z^{-1}$. It shows that the noise is filtered to higher frequencies while keeping the same total noise power compared to a Nyquist rate sampling technique. Therefore, most noise power is filtered in the signal's bandwidth.

Figure 2.14 shows the signal and noise transfer functions calculated in equations (2.23) and (2.24):

$$
\text{STF} = z^{-1} = e^{-j2\pi \frac{f}{f_s}} \rightarrow |STF| = 1 \tag{2.23}
$$

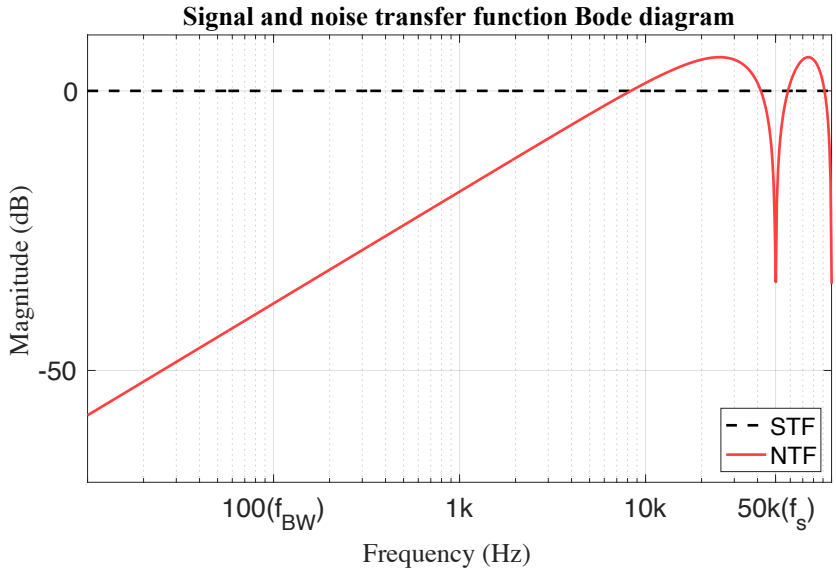

Figure 2.14: Bode diagram magnitude of signal and noise transfer functions.
(signal bandwidth = 100 Hz, sampling frequency = 50 kHz).

where f_s is the sampling frequency. (2.23) shows that signal's magnitude is unchanged after the $\Delta\Sigma$ modulation. The noise transfer function is calculated as below:

$$\begin{aligned} \text{NTF} = 1 - z^{-1} = 1 - e^{-j2\pi\frac{f}{f_s}} = 1 - \cos\left(2\pi\frac{f}{f_s}\right) - j\sin\left(2\pi\frac{f}{f_s}\right) \\ \xrightarrow{f<<f_s} \text{NTF} \simeq -j2\pi\frac{f}{f_s} \end{aligned} \tag{2.24}$$

where f_s is the sampling frequency which is much higher than the signal's bandwidth. As expected, the noise's magnitude is small when the frequency is low. In the bandwidth, the signal power is dominant compared to the noise, and most high-power noise gets filtered, resulting in a high Signal-to-Noise Ratio (SNR).

The output signal of a $\Delta\Sigma$ encoder results from an unchanged low-frequency input and a high-frequency noise shaping. Figure 2.15a shows us that the ratio of '1' to '0' bits in a time window encodes the input value. As a result, the density of '1' bits at the output digital signal increases when the input gets larger. In Figure 2.15b, the input signal is recovered after a first-order low-pass filter. Figure 2.15c depicts the noise shaping of the $\Delta\Sigma$ modulation in the output's power spectral density (PSD). As expected, the dominant in-band power is the signal's power at 100 Hz. Also,

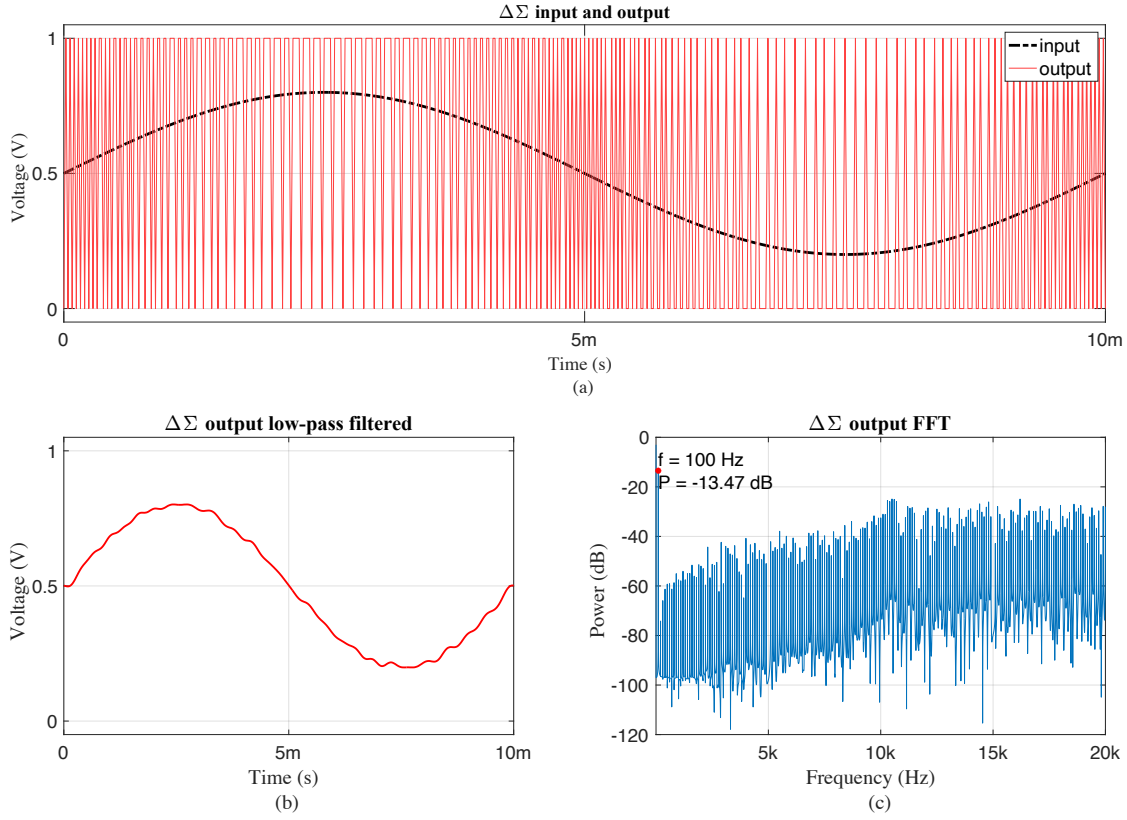

Figure 2.15: **a)** Input and output signal of the ΔΣ encoder **b)** First-order low-pass filtered output **c)** Power spectral density of the ΔΣ output signal.

most of the noise power is moved to the higher frequencies, which makes filtering the noise easier.

The in-band SNR is defined as the sum of all noise power up to the Nyquist frequency of the input's bandwidth:

$$\text{SNR} = \frac{P_{signal}}{\sum P_{noise}} = \frac{P_{signal}}{\sum_{f \neq f_{signal}}^{2 \times f_{signal}} P_{PSD}} \tag{2.25}$$

where P_{PSD} is the power of the signal at each frequency and P_{signal} is its power at f_{signal}, defined as the input signal's bandwidth; twice this bandwidth is the Nyquist frequency in which the noise power is calculated. In this PSD figure, the in-band SNR is 51.23 dB. From SNR, we can measure the equivalent number of bits resolution (N_{bit}) with (2.26) [7].

$$\text{N}_{\text{bit}} = \frac{\text{SNR} - 1.77}{6.02} \tag{2.26}$$

where SNR is calculated in (2.25). In the simulated system, 50 kHz oversampling of a 100 Hz input

signal results in 51.23 dB SNR and 8 bits of resolution.

2.4.2 $\Delta\Sigma$ Common Circuit Implementation

A first-order $\Delta\Sigma$ encoder consists of a discrete-time integrator, a quantizer, and a DAC, as shown in Figure 2.16. Φ_1 and Φ_2 are two non-overlapping clock signals at f_s. The standard choice for the discrete-time integrator is the Switch-Capacitor (SC) circuit [8]. As shown in Figure 2.16, the SC integrator samples the analog input over C_s at the clock signal Φ_2 with a sampling frequency of f_s. By replacing the DAC in the figure with 0 V and opening the loop, the following equations calculate the output voltage of the SC integrator in discrete time:

$$\begin{aligned} 0 - (-C_s V_{in}((n-\frac{1}{2})T)) &= C_f[V_o(nT) - V_o((n-\frac{1}{2})T)] \\ \Rightarrow C_s z^{\frac{-1}{2}} V_{in}^{(2)} &= C_f V_o^{(1)} - C_f z^{\frac{-1}{2}} V_o^{(2)} \end{aligned} \tag{2.27}$$

where $V_x^{(1)}$ is the z-transform of $V_x(nT)$ for the clock signal Φ_1, and $V_x^{(2)}$ is the z-transform of $V_x(\frac{2n-1}{2}T)$, sampled in the clock signal Φ_2. Also, since there is a phase shift of 180° between the two clock signals, we know that $z^{\frac{-1}{2}} V_x^{(2)} = z^{-1} V_x^{(1)}$. By replacing it in (2.27) we have:

$$\begin{aligned} C_s z^{\frac{-1}{2}} V_{in}^{(2)} = C_f V_o^{(1)} - C_f z^{-1} V_o^{(1)} \Rightarrow V_o^{(1)} &= \frac{C_s}{C_f} \frac{z^{-\frac{1}{2}}}{1 - z^{-1}} V_{in}^{(2)} \\ V_o^{(2)} = z^{-\frac{1}{2}} V_o^{(1)} \Rightarrow H(z) = \frac{V_o^{(2)}}{V_{in}^{(2)}} &= \frac{C_s}{C_f} \frac{z^{-1}}{1 - z^{-1}} \end{aligned} \tag{2.28}$$

where C_s and C_f are shown in Figure 2.16, and $H(z)$ is the transfer function of the SC filter and supports the mathematical model of Figure 2.13 with a gain of $\frac{C_s}{C_f}$. Therefore, the voltage V_m is the sampled V_{in}, accumulated over time in the open-loop circuit [40].

The 1-bit ADC is just a clocked-comparator producing '1' and '0' bits at the output from V_m changing at f_s. Then, a single-bit DAC converts the output bits to V_{ref}^+ or V_{ref}^-. The chosen V_{ref} is subtracted from the input voltage by C_s completing the feedback loop of the $\Delta\Sigma$ modulator circuit.

The output produced by a $\Delta\Sigma$ modulator is a single-bit encoding of the input voltage. If we swap '0's and '1's at the output, the result is the negative-processed bits in which the density of '0's

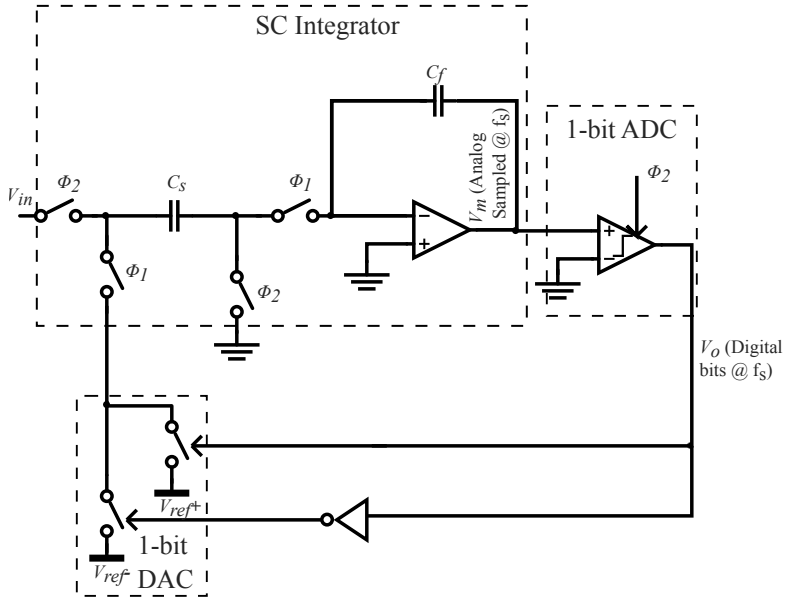

Figure 2.16: Circuit implementation of the $\Delta\Sigma$ system with an SC integrator, a comparator, and single-bit switching DAC [8].

encodes the input signal's voltage level. Hence, the regular $\Delta\Sigma$ modulated bits are called positive-processed, which are the inverse of the negative-processed bits. We have the following equations for these outputs:

$$V_o^+(x) = 1 - V_o^-(x) \tag{2.29}$$

where $V_o^+(x)$ and $V_o^-(x)$ are the positive and negative-processed outputs to input signal x. If two signals of x and x' complement each other, it means $x + x' = 1$. $\Delta\Sigma$ modulating x and x' produces two complements output. Therefore, the number of '1's at the output of one of them is equal to the number of '0's at the other one's output. By producing positive and negative-processed outputs, the following equations describe the relationship between the positive and negative-processed outputs encoding a signal and its complement:

$$V_o^-(x) = 1 - V_o^+(x) = V_o^+(x') \text{ and } V_o^+(x) = V_o^-(x') \tag{2.30}$$

where $V_o^+(x')$ and $V_o^-(x')$ are the positive and negative-processed outputs to input signal x', which is the complement of input x.

2.5 Analog Input to Spike Converters

The input signal encoder is an essential part of the SNN system. The required accuracy of this encoding can vary based on the needed input resolution in the application. In most cases, the input signal is a sensor's output voltage. Since the input of neurons is a current signal, the sensor's output must be converted to a current. Thus, a solution is to convert the voltage input into a current with a Voltage-to-Current (V-I) converter and feed that current into the neuron. The other solution is to design different input-layer neuron circuits that receive voltage signals at their input and produce spiking output. In many SNN systems, the information is converted to spiking signals on a separate system, which is then fed to the synapses of the SNN circuitry [21, 24].

2.5.1 Voltage to Current (V-I) Conversion

A neuron's input signal is a current injected into its membrane voltage. An analog input voltage encoding method is to convert this voltage into a current received by the neuron. A voltage-to-current (V-I) converter can be used for this purpose. This V-I converter has to be linear and support a full input range. Figure 2.17 shows an example of a linear V-I converter [9]. This circuit converts the input voltage into an output current as below:

$$\begin{gathered} I_{in} = \frac{V_s}{R_s}, V_s = V_{in} \Rightarrow I_{in} = \frac{V_s}{R_s} \text{ and } n = \frac{\frac{W}{L}(M_2)}{\frac{W}{L}(M_1)} \\ I_{out} = n \times I_{in} \Rightarrow I_{out} = \frac{nV_s}{R_s} \end{gathered} \tag{2.31}$$

where V_s and R_s are the voltage and resistor of the input transistor's source, and n is the current mirror ratio between M_1 and M_2. Moreover, it is assumed that $V_s = V_{in}$ remains smaller than $V_m = V_{GS}(M_1)$, and if V_{in} is bigger than this value, it saturates I_{out} and the converter loses its linearity.

The V-I converter cannot provide acceptable linearity when the full-range input signal is needed. The input current of the LIF neuron, shown in Figure 2.17, can only be in the sub-threshold region because an LIF neuron is receiving this current. Therefore, when the voltage changes from 0 V to V_{DD}, the corresponding current can only be in the subthreshold current range of MOSFETs, which

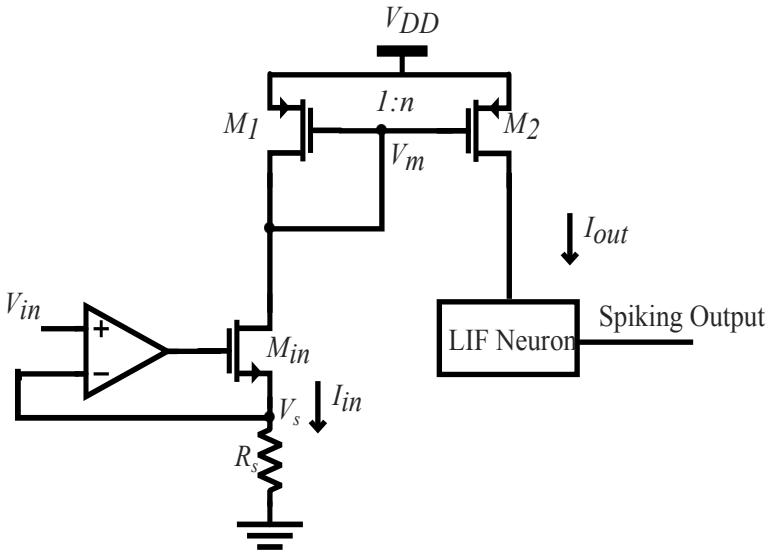

Figure 2.17: The linear V-I converter used in the input interface of an LIF neuron introduced in [9].

is very small. As a result, providing a high current mirror ratio, n, requires large transistors. The other way to produce a current in the sub-threshold range is to increase R_s. This solution still does not solve the large area problem because it needs $M\Omega$ range resistors to produce a sub-threshold range current [9].

This circuit is designed with $R_s = 1$ MΩ, and $n = \frac{1}{2}$. Then, it is simulated with an ideal Op-Amp, and the result is plotted in Figure 2.18. As shown in the figure, the V-I converter of Figure 2.17 is only linear in 70% of the V_{DD} range. When the input voltage reaches $V_{DD} - V_{SG}(M1) = 0.7$ V, the output current becomes 0 A because M_{in} enters triode region. Moreover, 1 MΩ resistor occupies a large area of 0.05 mm^2 inside the chip.

Every input neuron needs an analog input encoder. Since there are multiple input neurons in a system, the area is an important factor in designing the input encoder, and large-area circuits are unacceptable. Since most of the linear V-I converters use current mirror transistors or resistors, the area problem exists in all of them to create a sub-threshold range current.

In applications where lower accuracy is needed, simple V-I converters are helpful. The already-designed synapse can convert the input voltage into a current in these architectures. In [5], a simple V-I converter is designed in Figure 2.19a. An opamp shifts the input signal to 0 V-V_{DD} range. Then the modulated output is converted to a current with an assigned weight by the transistor M_{in}. The

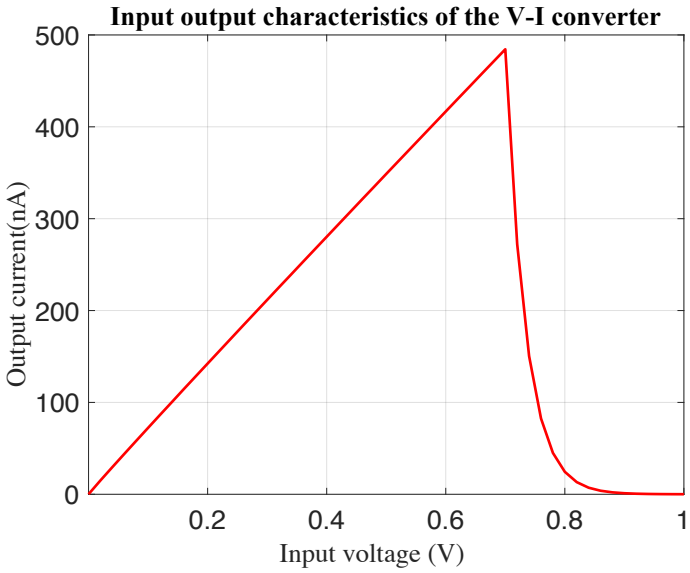

Figure 2.18: Input-output characteristics of the linear V-I converter of Figure 2.17.

neuron receives this current and fires spikes.

Figure 2.19b shows the firing frequency of this neuron to the original input. As shown, the linear range of this conversion is very narrow. In lower input voltages, the output does not fire. Also, the firing frequency gets saturated in the higher values of input. The maximum firing frequency depends on the dynamics of the designed neuron.

In some cases, only 1-2 bits of information from the input signal are sufficient for training and using the SNN system like MNIST classification [37]. In these applications utilizing the synapse as the V-I converter is practical, as depicted in Figure 2.19. In other systems where high resolution of linear encoding is necessary, designed V-I converters consume a large area because of the current mirror transistors or $M\Omega$ range resistors. Replicating large numbers of these converters in an SNN chip to feed the input layer neurons can cost a lot, and encoding the signal outside the chip is more efficient.

2.5.2 Δ Modulation

In applications where detecting change in the input is more important than receiving the input's value, Δ modulators can encode the input signal's change. For instance, [10] introduces a Δ encoder to detect high-frequency oscillation in EEG signals. Figure 2.20 shows the circuit of an input layer

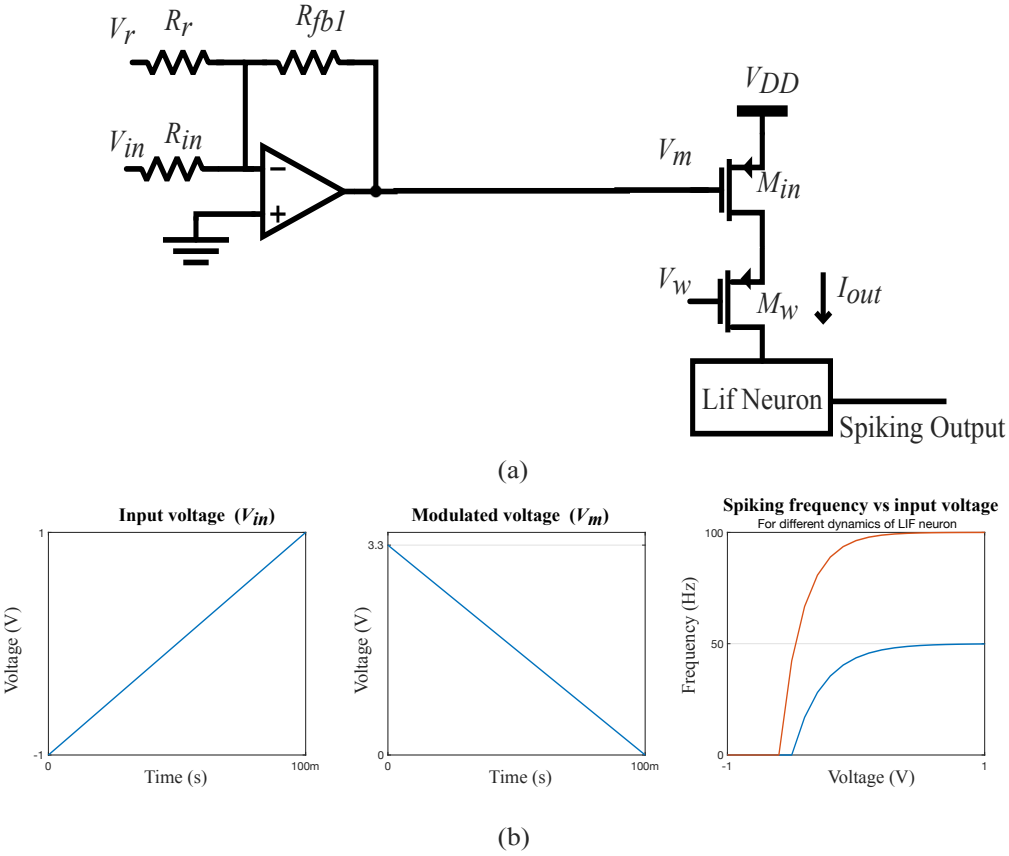

Figure 2.19: **a)** The input encoder designed in [5] **b)** Input, modulated signal and the corresponding spiking frequency of the circuit [5].

neuron with Δ modulation. In this circuit, the changes in band-passed input are compared to two thresholds. Suppose its increase is bigger than the upper threshold, V_{tu}; the Up signal fires after receiving the acknowledgment signal from the SNN. If the decrease in the input signal is bigger than the lower threshold, V_{td}, the Dn signal fires. Afterward, the resetting circuit is the same as the Up signal.

The result of simulating the Δ modulator circuit is shown in Figure 2.21 [10]. This figure shows spikes fire at the Up signal whenever the EEG signal increases. On the other hand, the Dn signal fires when the input decreases. The number of consecutive spikes depends on the rising or decaying slope of the EEG signal, where higher slope results in a larger firing frequency.

Δ encoding of the input is a powerful method when the input amplitude is very low, and no encoder can have that much accuracy. For instance, in [10], the band-passed input signal only varies

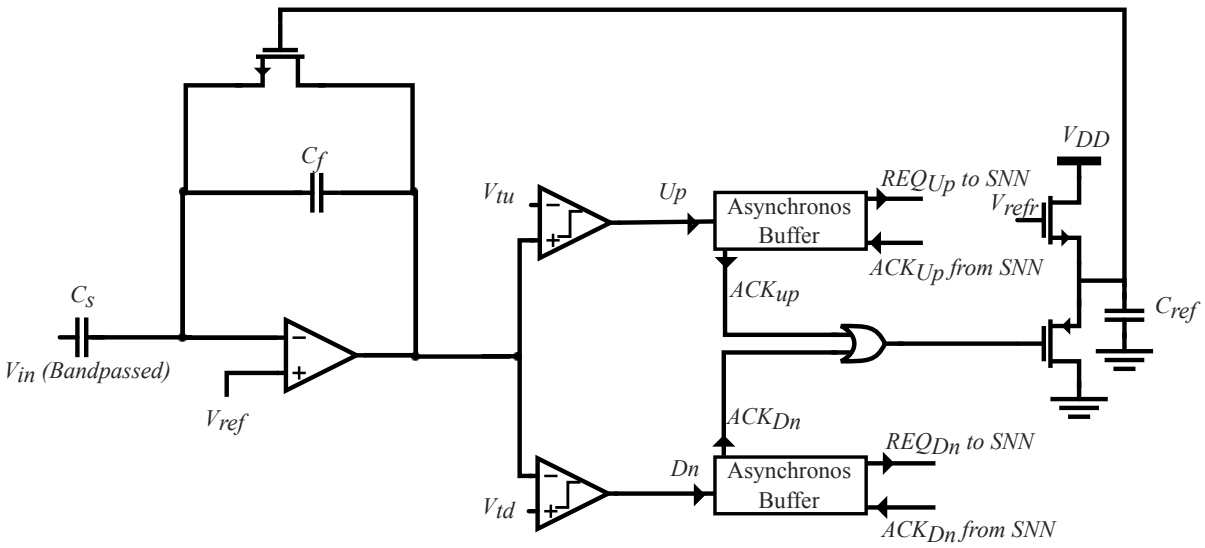

Figure 2.20: Encoding the changes in input in UP and DOWN signals by Δ modulation corresponding to increase or decrease of the input value [10]

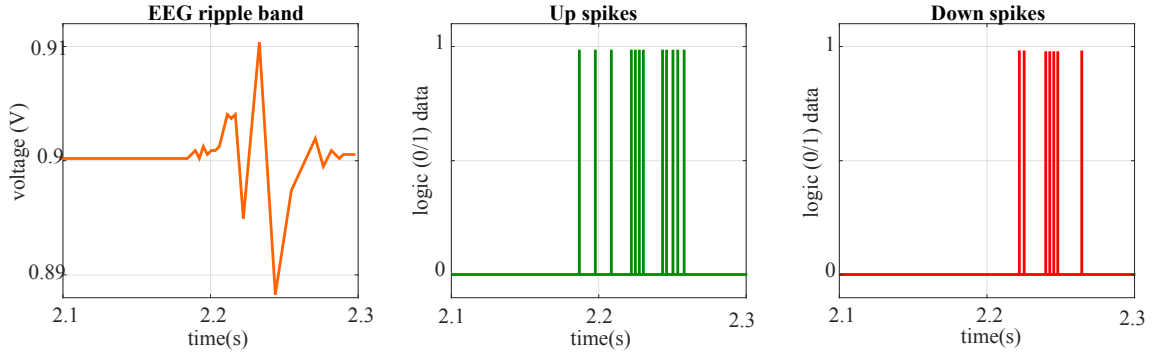

Figure 2.21: Input EEG signal, Up and down spikes encoding change slope of the input [10].

by 0.02 V as shown in Figure 2.21. Encoding the derivative of this signal to time can provide us with the needed information about the input to train and use the SNN system. In case of low-frequency changes in the input, we have to find other solutions for encoding the sensor signal.

2.6 Literature Conclusion

This work aims to encode the input signal of any SNN system and produce spiking outputs. As discussed, the input of neurons is in a current form, and synapses, designed for receiving spikes, cannot convert the analog sensory information into a current with reasonable accuracy. Since few

sensors produce current signals at their output, there is a need to encode the sensory voltage data into a current or spiking output. Encoding the input with a higher resolution results in higher accuracy. For these applications, V-I converters are not practical because sub-threshold neurons receive very low currents, and linear V-I converters consume a large area providing that current. In the following chapters, two methods will be discussed to produce a spiking output of an input voltage using the $\Delta\Sigma$ modulation.

Chapter 3

Design Methodology

This chapter introduces two types of input-layer neurons using $\Delta\Sigma$ modulation. Both neurons consist of a $\Delta\Sigma$ modulator and a spike firing circuit for encoding the analog input signal into spikes. All circuits are designed using the TSMC 65 nm technology kit in Cadence Virtuoso Environment.

3.1 Synchronous $\Delta\Sigma$ Analog to Spike (A-S) Converter

After $\Delta\Sigma$ modulating the input voltage into digital bits, the synchronous A-S converter produces spiking output from every '1' bit. Its high-level and circuit-level design is discussed in this section.

3.1.1 Block Diagram

The density of '1' bits at the output of the $\Delta\Sigma$ encodes the input signal's value. The same mechanism is appropriate for spiking neurons with rate encoding, where the frequency of spikes represents the neuron's input [24]. The Synchronous $\Delta\Sigma$ produces a bit stream of '1's and '0's by $\Delta\Sigma$ modulation as shown in the high-level architecture of Figure 3.1. In this neuron, the oversampling clock frequency controls the timing of the bits, as we have one bit per clock signal's rising edge. Next, the spike generator in Figure 3.1 monitors these bits with the same clock frequency. Whenever the spike converter reads a '1', it produces a spike, and its output resets after a delay. Then it rests until the next bit. As a result, we have a spiking signal at the output of this system.

The synchronous input neuron is simulated in Simulink with the clock frequency of 50 kHz.

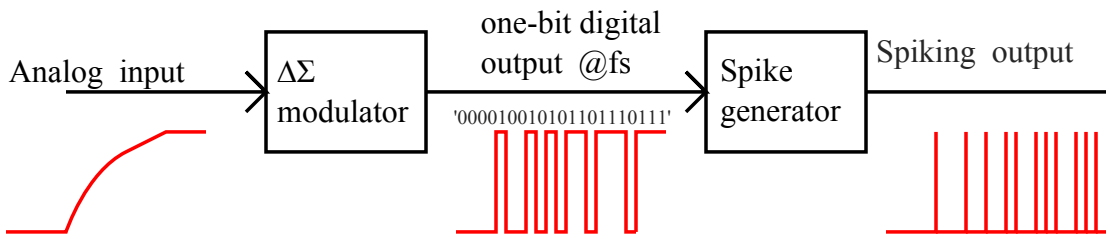

Figure 3.1: Block diagram of the synchronous ΔΣ A-S converter, and example signals at each node.

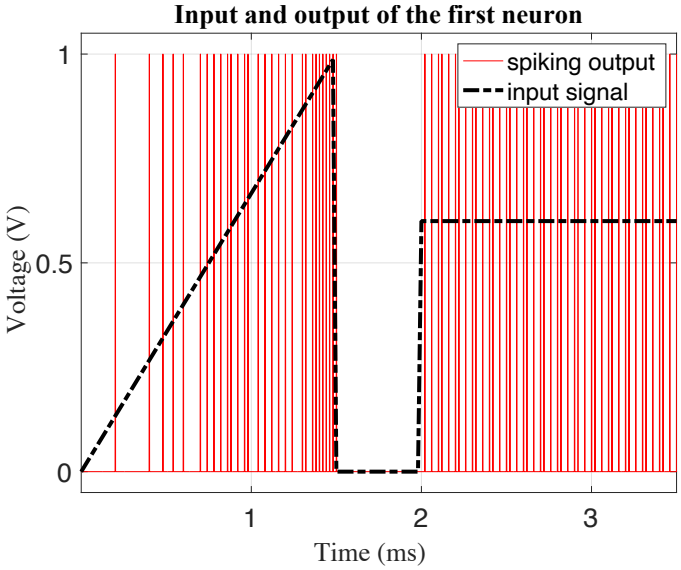

Figure 3.2: The output signal of the Synchronous ΔΣ A-S converter with respect to a ramp and step input (f_{clk} = 50 kHz).

The first part of the input signal in this simulation is a ramp from 0 to 1 V. Then after resetting to 0 V, a 0.6 V step signal is applied to the input. The output signal with this input is depicted in Figure 3.2. As shown, when the input is not changing, the time duration between spikes is not constant, whereas, in an IF neuron, the timing between spikes is uniform with a DC input. The output of the ΔΣ encoder for 0.6 V input, in a 1 V power supply system, is the repeating bit sequence of '11010'. Therefore, the output is '1' for 60% of the time. As a result, when this digital output is converted to spikes, we see the same behavior.

The number of spikes in a time window equals the number of '1's representing the original input value. The maximum firing frequency happens when the input is V_{DD} and all ΔΣ modulated bits are '1'. In this case, the input neuron spikes at the clock frequency of the ΔΣ system.

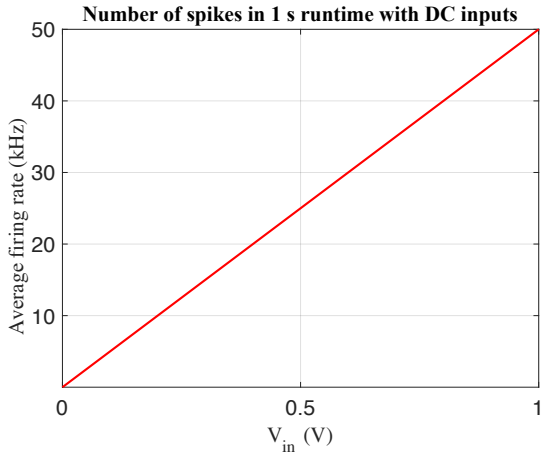

Figure 3.3: The number of spikes in 1 s runtime with respect to different DC input values (f_{clk} = 50 kHz).

The average firing rate in the synchronous $\Delta\Sigma$ input neuron is as follows:

$$f_{avg}^{spikes} = \frac{\bar{V_{in}}}{V_{DD}} \times f_{clk} \tag{3.1}$$

where $\bar{V_{in}}$ is the average of the input voltage. (3.1) shows that the spiking frequency is proportional to the clock frequency with the input voltage ratio to V_{DD}. As an example, in Figure 3.2, the average firing frequency after the 0.6 V step input is 30 kHz. In this case, the Inter-Spike Intervals (ISI) are not constant and vary between 20 and 40 microseconds. The average ISI is 33.3 μs. The following simulation of this neuron sweeps V_{in} from 0 to V_{DD} as shown in Figure 3.3. As expected from (3.1), the firing rate increases linearly with the DC input value with 100% accuracy in Figure 3.3. In this figure, the number of spikes in a 1 s time window is the average firing frequency.

3.1.2 Circuit Design

The circuit design of this neuron is provided step by step.

3.1.2.1 $\Delta\Sigma$ Modulator Circuit

The $\Delta\Sigma$ circuit is the first stage of the A-S converters. The neuron's firing frequency can vary between biological rates of around 100 Hz to accelerated rates below 10 MHz [31]. Accelerated

neurons are better when the input signal changes at high speed. This system receives an analog input signal between 0 and 1 V and samples it at a rate much faster than the input's bandwidth. The input's bandwidth is the maximum rate of change in the input signal received from a sensor. The maximum input bandwidth for this input neuron is 4 kHz. Thus, the maximum oversampling ratio (OSR) for the mentioned bandwidth is set to 1000, resulting in a maximum clock signal at 4 MHz to support accelerated firing. It results in a higher than 50 dB SNR at the output digital signal when the input's bandwidth is below 4 kHz [7].

This work chooses the 1st order $\Delta\Sigma$ topology, as discussed in the previous chapter. Figure 3.4a shows this work's design of the $\Delta\Sigma$ encoder based on Figure 2.16. The power supply for all circuits is $V_{DD} = 1$ V, and an analog ground replaces the ground reference voltage, A_{GND}, which is $V_{DD}/2 = 0.5$ V. Moreover, the reference voltages of the DAC are 0 and 1 V. In the figure, the 1-bit ADC consists of an analog comparator and a DFF. The rest of Figure 3.4 shows the OTA, comparator, DFF, and switches design.

An important part of this design is that the DFF has two outputs that are complements of each other. These outputs make it possible to produce both positive and negative-processed outputs. The negative-processed output's firing rate decreases when the input voltage increases because the density of '0's gets bigger than the density of '1's. The positive-processed output is the standard output of a $\Delta\Sigma$ modulator. We can write the equation below for the negative-processed output of the $\Delta\Sigma$ encoder:

$$V_o^-(V_{in}) = V_o^+(V_{DD} - V_{in}) \tag{3.2}$$

where $V_o^+(x)$ and $V_o^-(x)$ are the positive and negative-processed output to the input voltage x. (3.2) shows that every analysis for the negative-processed output of the $\Delta\Sigma$ encoder to the input voltage of V_{in} is the same as the positive-processed output to the input voltage of $1 - V_{in}$.

V_m is the discrete-time integrator's output signal that depends on the last bit at the output of the $\Delta\Sigma$. It is calculated as below:

$$\begin{aligned} \text{If } V_o(n-1) = 0 \Rightarrow V_m(n) &= V_m(n-1) + \frac{C_s}{C_f} \times V_{in}(n) \\ \text{If } V_o(n-1) = 1 \Rightarrow V_m(n) &= V_m(n-1) + \frac{C_s}{C_f} \times (V_{in}(n) - 1) \end{aligned} \tag{3.3}$$

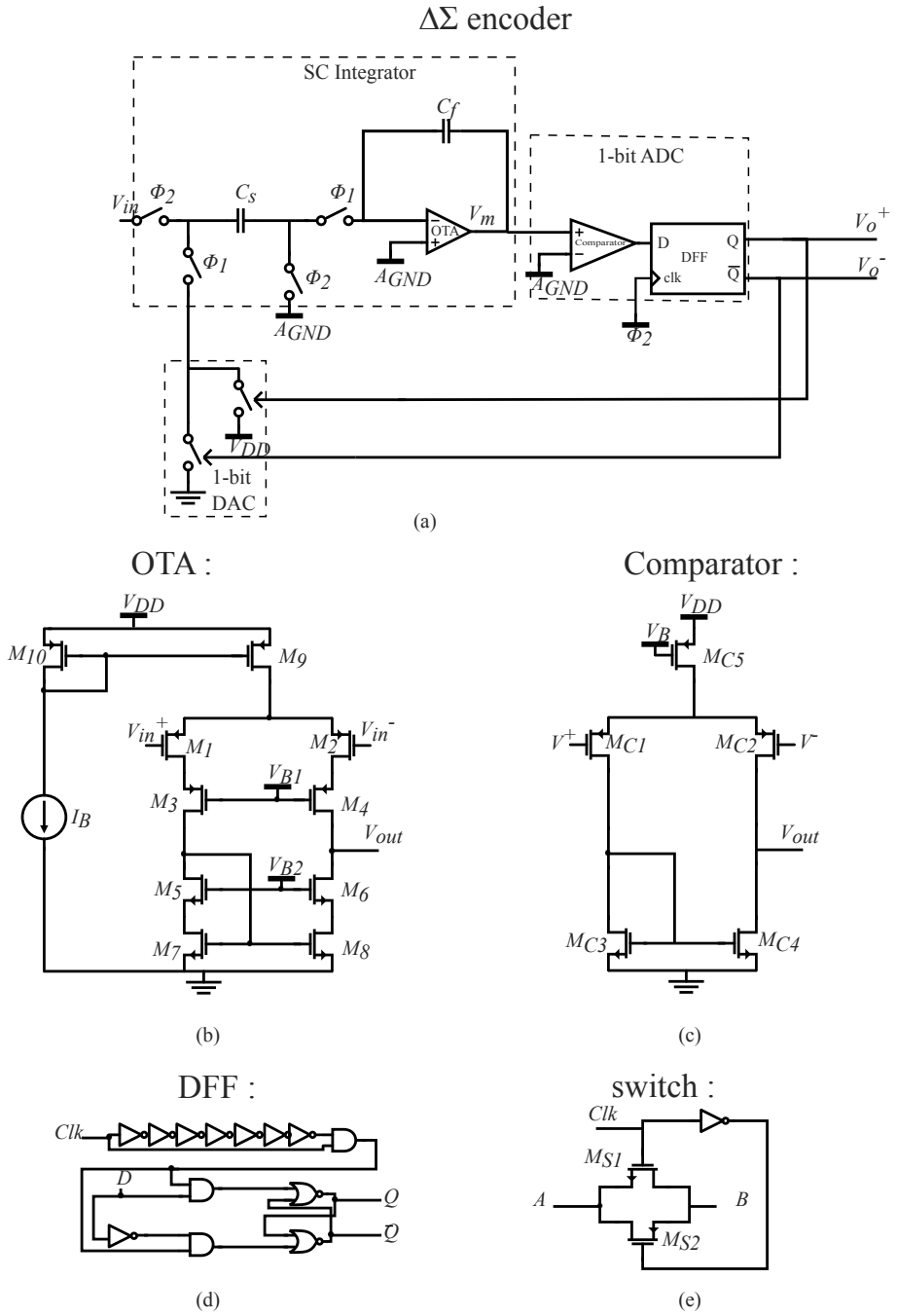

Figure 3.4: **a)** The $\Delta\Sigma$ circuit **b)** The single-stage differential-input OTA **c)** The single-stage analog comparator with differential input **d)** The DFF designed with AND, NOR, and inverter gates **e)** Transmission gate switch design.

where $V_o(n-1)$ is the previous bit at the output of the $\Delta\Sigma$ that is subtracted from the input, V_{in}, in the feedback loop. In an extreme case, when V_{in} is 0.99 V, one bit is '0' in every 100 bits. When the previous output bit is '0', V_m suddenly increases to $0.5+\frac{C_s}{C_f}\times 0.99$V. In the other extreme case, when V_{in} is 0.01 V, one bit is '1' in every 100 bits. When the last output bit is '1', V_m suddenly decreases to $0.5-\frac{C_s}{C_f}\times 0.99$V. Because the power supply is only 1 V and this work wants to encode full-range input, the gain of the closed-loop OTA is set as below:

$$0.5-\frac{C_s}{C_f}\times 0.\bar{9}\geq 0 \text{ and } 0.5+\frac{C_s}{C_f}\times 0.\bar{9}\leq 1 \Rightarrow \frac{C_s}{C_f}\leq 0.5 \tag{3.4}$$

where $\frac{C_s}{C_f}$ is the gain of the discrete-time integrator. From (3.4), we conclude that a gain larger than 0.5 results in clipping V_m to 0 or 1 and producing non-linearity. Therefore, the values of the capacitors are set in the table below.

Table 3.1: Capacitor values in the $\Delta\Sigma$ system

Parameter	C_s	C_f
Value	250 fF	500 fF

The OTA is a single-stage differential-input amplifier. Because the power supply is only 1 V, the low voltage cascode stage design is chosen [7]. This work avoids the double-stage design because of its higher area, power consumption, and frequency compensation. The design variables of the OTA are provided in Table 3.2.

Table 3.2: Design parameters of the differential-pair input OTA in Figure 3.4b

Parameter	L_{1-10}	W_{1-4}	W_{5-8}	W_9	W_{10}	I_B	V_{B1}	V_{B2}
Value	200 nm	4 μm	1 μm	3 μm	1 μm	2.4 μA	300 mV	700 mV

The OTA in the discrete-time integrator receives the input signal with a maximum bandwidth of 4 kHz sampled at 4 MHz. The specifications for this OTA to perform with this input are more than 50 dB gain and larger than 10 MHz gain-bandwidth product with 500 fF capacitive load. Also, we expect that the common-mode voltage gets attenuated, so the common-mode rejection ratio (CMRR) has to be greater than 50 dB. Moreover, the OTA should not consume more than 10 μA while satisfying the mentioned requirements. The simulation results of this OTA are shown in Table 3.3 with a 500 fF capacitive load, which verifies the OTA's specifications.

Table 3.3: OTA simulated results

Parameter	gain	3dB BW	gain-BW	phase-margin	DC power	CMRR
Value	52.6 dB	51.2 kHz	20.2 MHz	89°	9.8 μA	55 dB

The analog comparator is a simple one-stage differential-pair OTA. The $\Delta\Sigma$ utilizes this OTA in an open-loop setup. Since it is an open-loop OTA, the output is either saturated to its maximum or minimum value, which can be used as a comparator between its two inputs. In this setup, V^{-} is connected to $V_{DD}/2 = 0.5$ V, and V^{+} is compared to 0.5 V threshold. If it is larger than that, the output of the comparator rises to its maximum, a voltage close to V_{DD}. If not, the output is equal to its minimum, near 0 V. The design parameters are shown in Table 3.4. This comparator dissipates 2.2 μW power providing the open-loop gain of 27 dB. Furthermore, its 3-dB bandwidth and gain-bandwidth product with a DFF loading are 7 MHz and 165 MHz, which are sufficient for the requirements of the circuit.

Table 3.4: Design parameters of the analog comparator in Figure 3.4c

Parameter	L_{C1-C5}	W_{C1-C2}	W_{C3-C4}	W_{C5}	V_B	$I_D(M_{C5})$
Value	200 nm	2 μm	1 μm	2 μm	0.5 V	2.2 μA

Figure 3.4d shows the design of the D-Flip-Flop (DFF) that receives the analog output of the comparator and produces two digital signals sampled at the clock frequency. These signals complement each other and generate positive and negative-processed outputs. Figure 3.5 shows the design of digital gates for this DFF. In digital MOSFETs (M_{Dn}, M_{Dp}), the length is 65 nm. Digital PMOSs' width is 5.6 μm while for NMOSs, it is 1.3 μm. In this circuit, at each rising edge of the clock signal, Q gets the value of D, and $\bar{\text{Q}}$ gets its complement. In general, whenever V_m hits the threshold of the converter, the positive and negative-processed outputs of the $\Delta\Sigma$ get '1' and '0', respectively, in the next rising edge of the clock signal.

The DAC is designed using only two switches shown in Figure 3.4e. In this figure, both M_s transistors have a $\frac{W}{L}$ of $\frac{4\ \mu\text{m}}{100\ n\text{m}}$. It is a transmission-gate switch feeding the value of A to B whenever the clock signal gets '1'. In the DAC, the gates are controlled by Q and $\bar{\text{Q}}$, and the switches feed back '1' or '0' to be subtracted from the input. Moreover, this $\Delta\Sigma$ modulator uses the same switch design for the switch-cap integrator.

The topology of the SC circuit used in Figure 3.4a is the parasitic-insensitive design [7]. On

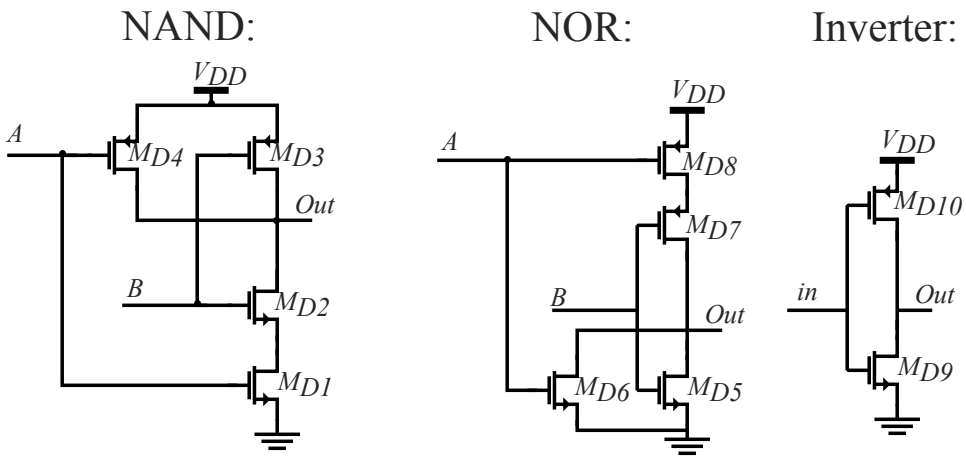

Figure 3.5: Digital gates design used in every digital circuit.

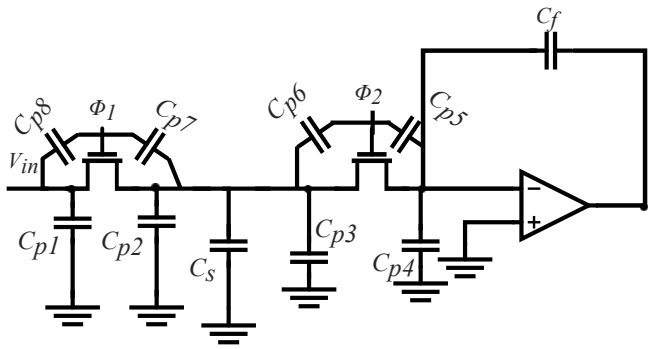

Figure 3.6: The simplest switch-capacitor discrete-time integrator with parasitic capacitors [7].

the other hand, Figure 3.6 shows the simplest SC integrator design with the parasitic capacitors. Because of the gate-source and gate-drain capacitors ($C_{p5} - C_{p8}$), the clock signal is coupled to the input signal's path. This effect is called clock feedthrough, causing sudden changes in the signal path whenever the clock voltage switches. Another effect of the parasitic capacitors is charge injection. When the gate signal of a switch turns off and the switch is not connected to a voltage source or ground, a charge is on the switch's drain and source capacitors. It distributes its channel charge to the source and drain, causing non-linearity. The design of Figure 3.4a eliminates these effects by using two vertical switches that send the charge on the switches to the ground.

The clock generator circuit is the remaining part of the $\Delta\Sigma$ encoder. There should be only one reference clock in the system. From the reference clock, the $\Delta\Sigma$ needs two non-overlapping clock signals. The circuit of Figure 3.7 produces two clock signals 180° apart. It receives the reference

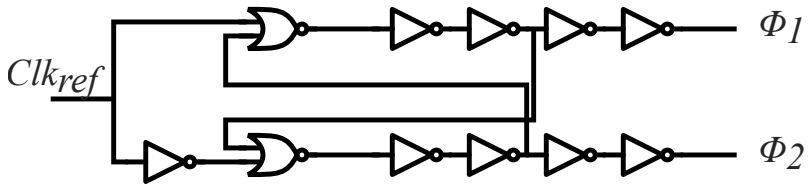

Figure 3.7: Circuit design of creating two non-overlapping clock signals from a clock reference.

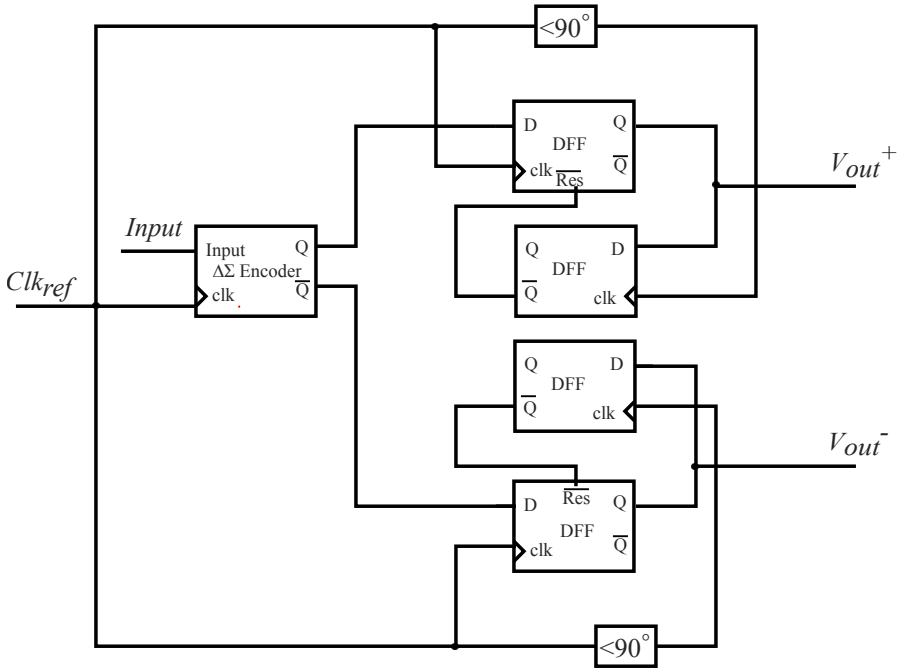

Figure 3.8: Synchronous $\Delta\Sigma$ A-S converter complete circuit design.

clock. Then, whenever one of the clock signals gets 'HIGH', it shuts down the other clock to 'LOW', ensuring the clocks are not 'HIGH' simultaneously.

3.1.2.2 Spike Producing Circuit

Since the Synchronous $\Delta\Sigma$ A-S converter monitors the output of the $\Delta\Sigma$ encoder at each clock cycle, a DFF with the same clock frequency is chosen. Moreover, the reset system must be able to reset this DFF asynchronously. Figure 3.8 shows the complete circuit of this input neuron designed from its block diagram in Figure 3.1.

In Figure 3.8, two positive and negative-processed outputs are produced from the output signal of the $\Delta\Sigma$ encoder. The spike-producing and reset systems are the same for both of these outputs.

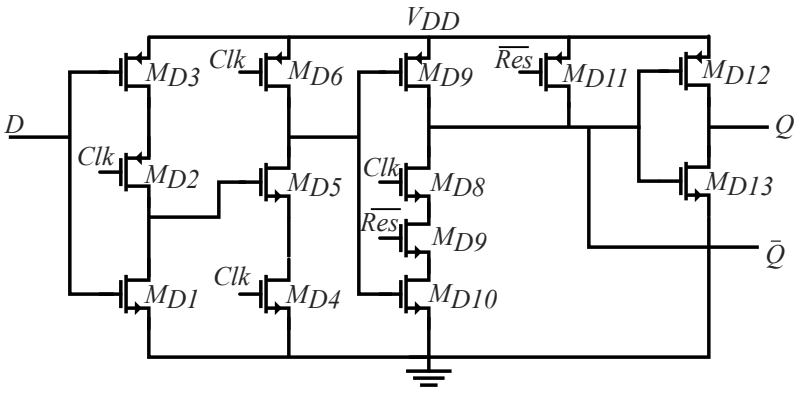

Figure 3.9: Clock-edge triggered DFF circuit with asynchronous active-low reset.

First, a DFF with active-low asynchronous reset reads the output of the $\Delta\Sigma$. The circuit design of this DFF is shown in Figure 3.9 with the same digital MOSFETs as the designed digital gates. If the output is '1', it resets it to '0' after a quarter-phase delay of the clock signal. Hence, the width of spikes at the output is a quarter of the clock period, and it can be lowered by decreasing the clock phase shift value.

The asynchronous reset in the circuit of Figure 3.9 consumes 28 pJ energy every time it resets. Figure 3.10 shows the $\Delta\Sigma$ and spiking output with their energy consumption. As shown in every clock period of 20 μs, if the $\Delta\Sigma$'s output is '1', a spike is generated. In this process, the $\Delta\Sigma$ encoder consumes static power of 12 μW. On the other hand, the DFF burns dynamic energy of 28 pJ at every spike. As a result, after 370 μs run time, the DFF consumed 420 pJ after 15 spikes; while $\Delta\Sigma$ dissipated 4 nJ.

3.1.2.3 High-Frequency Low-Power Spike Generating Circuit

When the clock frequency is in the MHz range, the delay of the spike generating DFFs affects the circuit's performance. Thus, in the high-frequency range, another method to create spikes from the output of the $\Delta\Sigma$ is to AND the clock edge with the output of the $\Delta\Sigma$ as shown in Figure 3.11. An AND gate, designed with digital MOSFETs as other parts of the circuit in Figure 3.5, produces spikes with a narrow width around 0.5 ns. This method produces narrow spikes that are not detectable when neurons fire in low to mid-range frequencies up to around 1 MHz. The negative-processed spiking output with 0.2 V DC input and the circuit's energy consumption is shown in

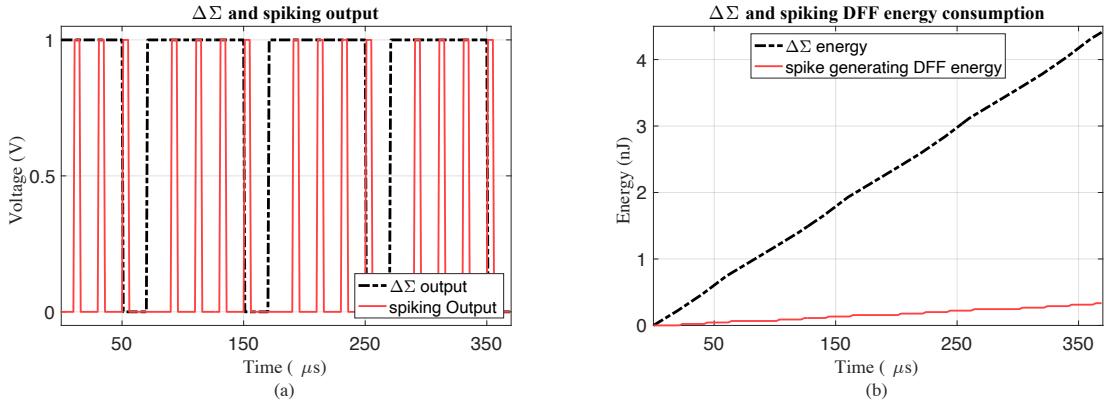

Figure 3.10: **a)** Positive-processed $\Delta\Sigma$ and spiking output with 0.8 V DC input. **b)** $\Delta\Sigma$ encoder and spike generating DFF's energy consumption in 370 μs with the same input

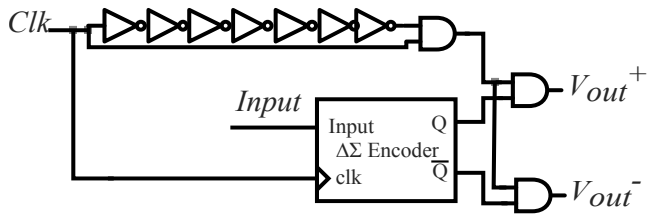

Figure 3.11: Low energy consumption spike generating circuit for high-frequency applications.

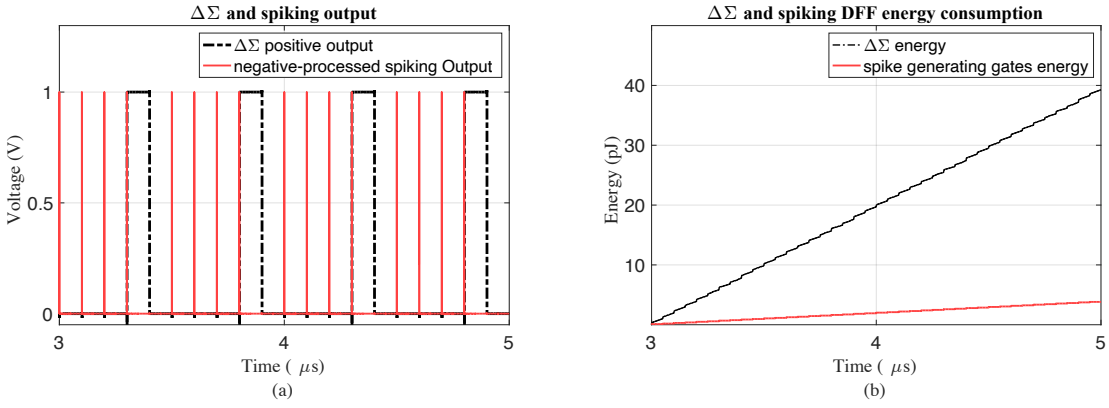

Figure 3.12: **a)** $\Delta\Sigma$ output signal and negative-processed spiking output with 0.2 V DC input **b)** $\Delta\Sigma$ encoder and spike generating AND gate's energy consumption in 2 μs with 0.2 V DC input

Figure 3.12 from 3 μs to 5 μs.

The delay value for the clock edge detection can get bigger to widen the spikes, but it diminishes the benefits of using this high-speed method because the delay circuit needs power and area. As shown in Figure 3.12.b, the gates' power consumption is 250 fJ per each spike. Therefore, this simple method consumes very low power and is beneficial when spikes with lower widths are detectable, such as accelerated spiking neural network circuits.

3.2 Neuromorphic $\Delta\Sigma$ Analog to Spike Converter

The neuromorphic $\Delta\Sigma$ A-S converter produces spiking voltage from $\Delta\Sigma$ modulating the input voltage using an actual synapse and neuron model. The synchronous input neuron encodes the analog voltage into spikes by almost perfect rate-encoding linearity, based on MATLAB simulation of Figure 4.3. The spiking pattern in the synchronous input neuron differs from other layer neurons. When the input is above 0.5 V, the $\Delta\Sigma$ modulated output produces consecutive '1' bits. Therefore, the spike generator of the synchronous A-S converter, which produces spikes from these bits, generates consecutive spikes. As a result, the ISIs are not constant with DC input, as shown in Figure 3.10a. Moreover, it cannot provide different properties of an IF neuron when needed in an application, like the membrane voltage, acknowledgment signal, stimulus current injection, etc. For example, if the SNN is designed with adaptation, the neuron reduces spiking frequency after receiving the input. When the input is unchanged, a feedback current is drained from the membrane reducing the spiking frequency. The synchronous input neuron cannot match other neurons' adaptation since it does not have a membrane node. Because the neuromorphic input neuron model introduced in this section utilizes an actual neuron and synapse model, it can compensate for the mentioned weaknesses of the synchronous one.

3.2.1 Block Diagram

The Neuromorphic $\Delta\Sigma$ A-S converter is designed to match other neurons of the system and consists of a neuron and synapse. The synapses cannot receive analog values at their input since they are designed to accept a spike train. If a synapse receives an analog input voltage, the current

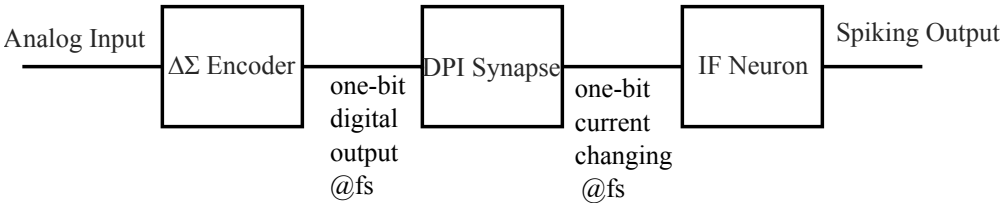

Figure 3.13: Block diagram of the neuromorphic $\Delta\Sigma$ A-S converter

generated is not a linear conversion of the input. Since spikes only have two values of '1' and '0', the single-bit input voltage of the synapse is produced by a $\Delta\Sigma$ encoder in this input-layer neuron model. Figure 3.13 shows the block diagram and neuron models chosen for this work.

In this input neuron model, the $\Delta\Sigma$ encodes the input into the density of '1's like the first one. This single-bit oversampled output controls the input of a DPI synapse. Then the synapse low-pass filters the digital output with a time constant smaller than the sampling frequency of the $\Delta\Sigma$. As a result, the synapse outputs a current whenever it receives a '1' and turns the output current off whenever it receives a '0'. Afterward, an IF neuron receives this current and integrates it. The membrane voltage of this neuron increases by integrating the input current. The rate of the membrane voltage hitting the threshold and creating a spike is proportionate to the neuron's input current. Hence the spiking rate encodes the original analog input received by the $\Delta\Sigma$ encoder.

The neuron's dynamics, like the membrane capacitor and refractory period, control the output's firing rate. In this neuron, assuming a perfect single-bit voltage-to-current conversion in the synapse, the equations defining this input neuron are as below.

$$\begin{gathered}
\Delta\Sigma \text{ conversion } \rightarrow W(z) = z^{-1}V_{in}(z) + (1 - z^{-1})Q(z) \\
\text{Synapse} \rightarrow \text{if } W(nT) = 1 \Rightarrow I_{in}(nT) = I_1, \text{ else } \Rightarrow I_{in}(nT) = 0 \\
\text{IF Neuron } \Rightarrow \tau\frac{du_m}{dt} = RI_{in} - (u - u_{rest}) \\
\text{when } u_m > u_\theta \Rightarrow \text{ spike and } u_m \rightarrow u_{rest} = 0
\end{gathered} \tag{3.5}$$

where $W(nT)$ is the digital output of the $\Delta\Sigma$ sampled at $f_s = 1/T$, and $W(z)$ is its z-transformed signal. I_1 is the constant current generated by the synapse whenever '1' is received. u_m, and u_θ are the membrane potential and the threshold voltage of the IF neuron. Because in the input stage, the

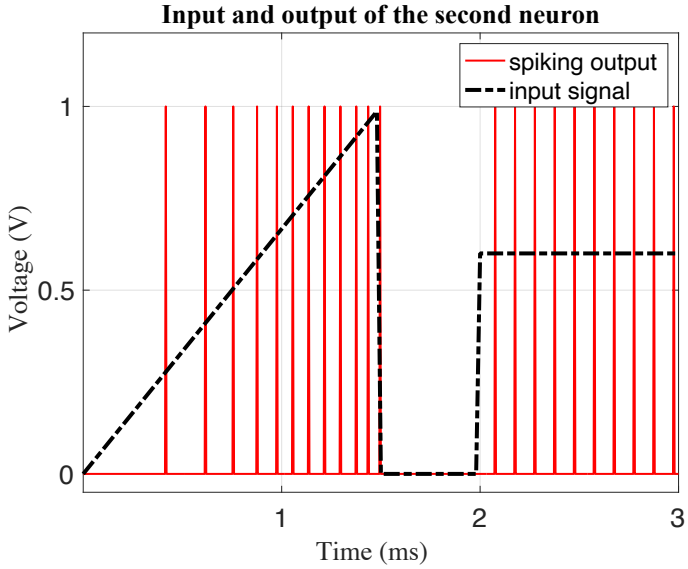

Figure 3.14: The output signal of the neuromorphic $\Delta\Sigma$ A-S converter with respect to a ramp and step input (f_{clk} = 50 kHz).

maximum possible resolution is needed, the leakage of the IF neuron is set to the minimum possible value for better linearity in low-end input voltages.

Equations of (3.5) are simulated in MATLAB Simulink, and the results are shown in Figure 3.14 and 3.15. The integration gain is the value of $\frac{R}{\tau}$, which is equal to 50 $\frac{\text{G}\Omega}{\text{s}}$. The threshold voltage is half the power supply, 0.5 V. The I_1 current generated by the synapse is 500 nA.

Figure 3.14 shows the spiking output to a ramp and 0.6 V step input. Unlike the synchronous one, the output fires with a constant frequency of 10 kHz with the step input, and the distance between spikes is almost constant. In this setup, the maximum frequency is $\frac{f_{clk}}{3} = 16.67$ kHz.

The next simulation of this neuron sweeps V_{in} from 0 to V_{DD}, as shown in Figure 3.15. In this figure, the average firing frequency is plotted with the input voltage proving they have a linear relationship. Thus we can use the following equation to calculate the average firing frequency of any input:

$$f_{avg}^{spikes} = \frac{\bar{V_{in}}}{V_{DD}} \times \frac{f_{clk}}{n} \text{ (n=3 for the tests of Figure 3.15)} \tag{3.6}$$

where n is the number of '1' bits at the $\Delta\Sigma$ modulator's output that the membrane needs to hit the threshold. By altering the integration gain and time constant in (3.5), the value of n changes. If $n = 1$, the output produced by the neuromorphic input neuron equals the synchronous one. n is

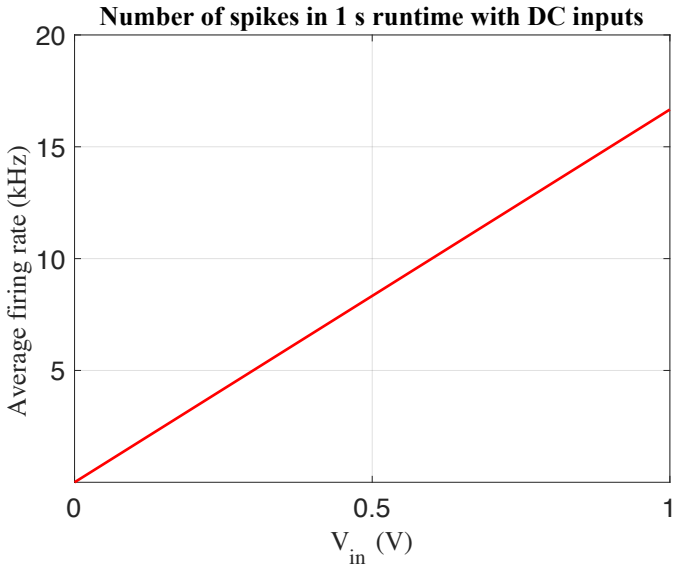

Figure 3.15: The average spiking frequency produced by the neuromorphic $\Delta\Sigma$ A-S converter in 1 s runtime with respect to different DC input values (f_{clk} = 50 kHz).

selected 3 to minimize the changes in ISIs with a DC input.

3.2.2 Circuit Design

The neuromorphic $\Delta\Sigma$ A-S converter consists of a $\Delta\Sigma$, a synapse and IF neuron. The $\Delta\Sigma$ circuit is the same as the synchronous one. This section designs the circuits of the synapse and IF neuron.

3.2.2.1 Synaptic Circuit

This work chooses the differential-pair integrator synaptic circuit introduced in Chapter 2. Figure 3.16 shows the circuit design of the synapse [6]. In this circuit, M_1 receives the one-bit digital output of the $\Delta\Sigma$. Then it converts the output to a current I_{in} based on the assigned weight, V_w. Then C_{syn} integrates I_{in} and produce the gate voltage of M_6. V_τ sets the time constant, and V_g controls the gain of the integration as calculated in (2.20).

The MOSFETs' sizing and control voltages are set as Table 3.5. Since all transistors operate in the sub-threshold region, they consume very low power. This synapse consumes 400 nW on average when its input is '1'. This synapse can be designed with small-sized transistors given in Table 3.5

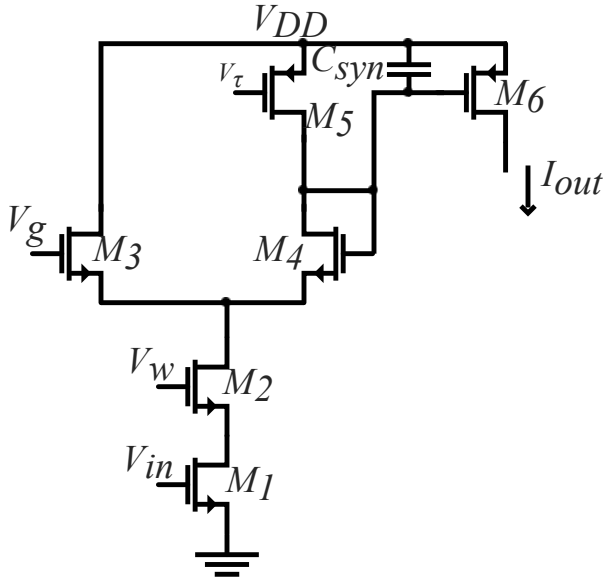

Figure 3.16: Differential-pair integrator synapse circuit.

because it dissipates only 400 nA. This synapse is simulated receiving the positive-processed output voltage of the $\Delta\Sigma$ produced by 0.8 V input and 50 kHz clock signal. The output current is shown in Figure 3.17 with different weight values. As shown, the time constant of the low-pass filtering is smaller than a clock period.

Table 3.5: Design parameters of the DPI synapse

Parameter	L_{PMOS}	L_{NMOS}	W_{PMOS}	W_{NMOS}	V_w	V_g	V_τ	C_{syn}
Value	240 nm	240 nm	1 μm	1 μm	420 mV	700 mV	700 mV	1 pF

As shown in Figure 2.11, and 3.17, increasing the weight voltage does not result in a linear output current increase. For instance in the simulation setup of Figure 3.17, changing the weight voltage from 350 mV to 500 mV, results in increasing the output current from 170 nA to 2.8 μA in a non-linear fashion.

The synaptic current is produced by M_6 and can be altered by changing this MOSFET's size. Although M_6's size is the minimum possible size in TSMC 65 nm technology, multiple MOSFETs can be in series with connected gates to reduce the synaptic current, I_{out}. Furthermore, the number of in-series transistors to create M_6 can be programmable to control the synaptic current by software resulting in a larger degree of freedom in the system.

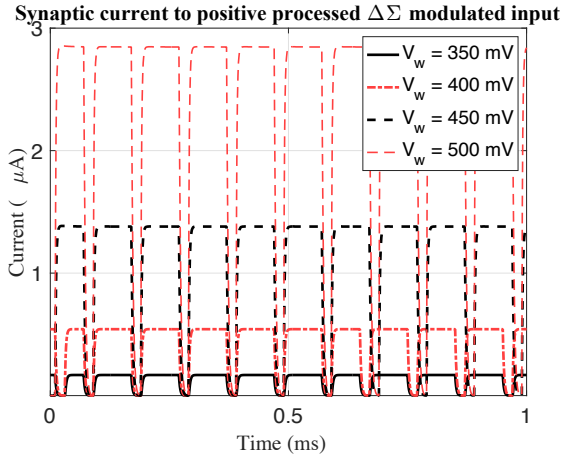

Figure 3.17: The output current of DPI synapse measured in different weight values. The input voltage of the synapse is the $\Delta\Sigma$ modulation of a 0.8 V DC voltage.

3.2.2.2 IF Neuron Circuit

An IF neuron receives the current generated by the synapse and integrates it over a membrane voltage. Then, when the membrane crosses a threshold, the neuron fires a spike and resets its membrane after a delay. The circuit of a LIF neuron with differential pair integrator is discussed in Chapter 2.

The DPI neuron designed in [4] is introduced in Chapter 2. This neuron adds leakage to the membrane. The leakage causes the decay of membrane potential over time, mimicking biological neurons. V_{lk} in Figure 2.7 controls this leakage rate. In this work, we want to encode the input with the best possible resolution, and biological plausibility is not a priority. Equations (3.5) shows that reducing the leakage increases encoding accuracy, especially for inputs near zero. When the input is near zero, the output of the $\Delta\Sigma$ is '0' most of the time and gets '1' occasionally. When the synapse receives '0', it feeds no current to the membrane, and the membrane potential decays over time. Because of this decay, the integrated value is lost, adding non-linearity to the encoding.

This work chooses the circuit of Figure 3.18 as the neuron of the second encoding method. The differential-pair integrator is removed from the LIF neuron's circuit. The synapse has a DPI, and the $\Delta\Sigma$ modulated voltage is already filtered by it. Moreover, the DPI in the neuron adds non-linearity because of the leakage. The only leakage in the designed IF neuron is the non-idealities of the membrane capacitor, which are very small.

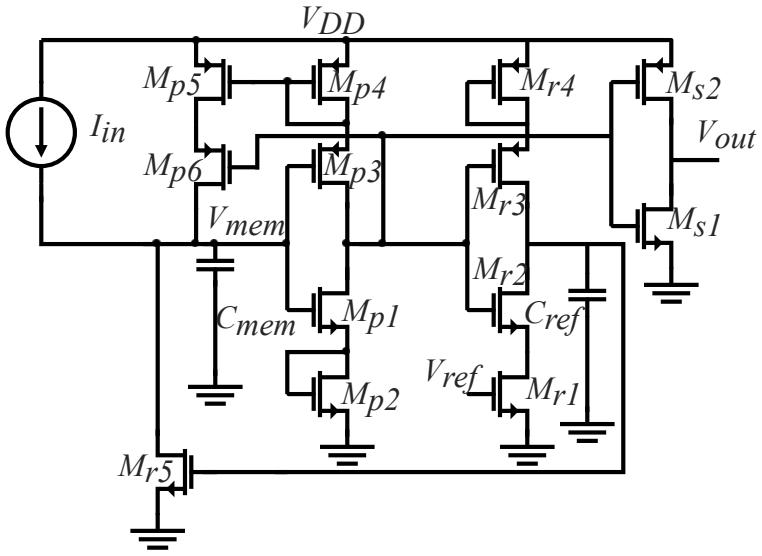

Figure 3.18: The IF neuron circuit design without a leakage.

In this circuit, the neuron receives the $\Delta\Sigma$ modulated I_{in}, and C_{mem} integrates it. Then by raising the membrane voltage, positive feedback is applied, which results in a faster increase of the membrane potential. When it reaches the threshold of 0.5 V, the voltage of C_{ref} rises, turning on M_{r5}, which resets the membrane voltage. The period when the neuron resets is the refractory period, and it cannot fire in that time. This period is controlled by V_{ref}, and C_{ref}.

The MOSFETs' sizing, biasing voltages, and capacitors are chosen in Table 3.6. To check the

Table 3.6: Design parameters of the IF neuron

Parameter	L_{PMOS}	L_{NMOS}	W_{PMOS}	W_{NMOS}	V_{ref}	C_{mem}	C_{ref}
Value	65 nm	65 nm	6 μm	2 μm	360 mV	20 pF	1 pF

response of this input layer neuron, a 0.2 V DC input and 50 kHz clock signal is given to the $\Delta\Sigma$ encoder. The negative-processed output of the $\Delta\Sigma$ is connected to the input voltage of the synapse, and the negative-processed output to 0.2 V DC input is equal to positive-processed one to 0.8 V input. Therefore, the synaptic output current is as Figure 3.17 with an 'ON' current value of 180 nA. With this current, the neuron fires with an average frequency of 14.1 kHz. Figure 3.19 shows the spiking output and the membrane potential due to the mentioned input. As shown, the membrane potential holds value in case of zero input because of no leakage. The width of spikes in this setup is 500 ns, which can vary by changing V_{ref} and C_{ref}.

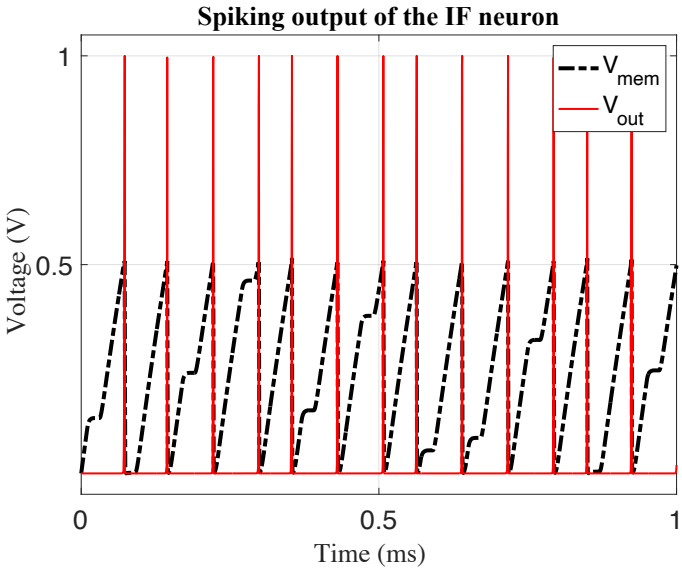

Figure 3.19: The IF neuron circuit membrane potential and spiking output. The input is negative-processed $\Delta\Sigma$ modulated current to 0.2 V DC voltage.

The IF neuron consumes dynamic power. It burns 17.1 pJ per spike. With the values assigned in Table 3.6 and the designed synapse, the maximum firing frequency is 17.5 kHz. Hence, in case of maximum firing, the IF neuron consumes 300 nW on average.

3.3 Design Comparison

The synchronous input neuron consists of a $\Delta\Sigma$ modulator and two DFFs to create spikes from the output of the $\Delta\Sigma$. On the other hand, the neuromorphic input encoder is designed by a DPI synapse and an IF neuron in addition to the same $\Delta\Sigma$ encoder of the synchronous one. In this work, both neurons use a first-order $\Delta\Sigma$ encoder. The neuromorphic A-S converter benefits from higher degrees of freedom than the synchronous one.

There are four biasing voltages and three capacitors that we can change based on the needs of the SNN. In the neuromorphic $\Delta\Sigma$ neuron, we can control the spiking frequency, the spiking pattern, and the width of each spike. On the other hand, in the synchronous $\Delta\Sigma$ input neuron, the only design choice is the phase shifting value of the clock frequency, which controls the width of spikes. The neuromorphic $\Delta\Sigma$ provides these design choices by consuming a much larger area than

the synchronous one. Moreover, the leakage of the IF circuit in the neuromorphic encoder cannot be eliminated completely, which causes a slight non-linearity in lower-end inputs.

The synchronous input neuron based on two DFFs' power consumption is slightly larger than the neuromorphic one. In high-firing frequency applications, the synchronous one consumes smaller energy by using the circuit of Figure 3.11 as the spike generator. In the next chapter, we will simulate these input neurons in different circumstances and compare them in more depth to see which can provide more benefits in various applications.

Chapter 4

Simulation Results

This chapter focuses on different simulation setups that evaluate the performance of the input encoder neurons. All simulations are examined in the analog design environment of Cadence Virtuose. The input neurons' circuits that will be simulated are discussed in Chapter 3 and designed with the TSMC 65 nm technology.

4.1 DC Input

From previous chapters, we know that when neurons receive a DC input, their average firing rate remains constant, defined in (4.1):

$$f_{avg}^{spikes} = \frac{N_s}{t} \tag{4.1}$$

where N_s is the number of spikes, and t is the simulation's runtime. From (3.6) (n=1 for the synchronous A-S converter), we know that the average firing rate of the input neurons represents the input voltage's value. Thus, it is a constant when V_{in} is not changing. A consistent average firing rate does not mean that the Inter-Spike Intervals (ISI) values are unchanging in time. Although the timing between spikes may differ, its average is fixed, representing a DC input.

This section simulates both input neurons' circuits with DC input voltages. We expect the circuit simulation results to match the mathematical ones shown in Figures 3.3 and 3.15.

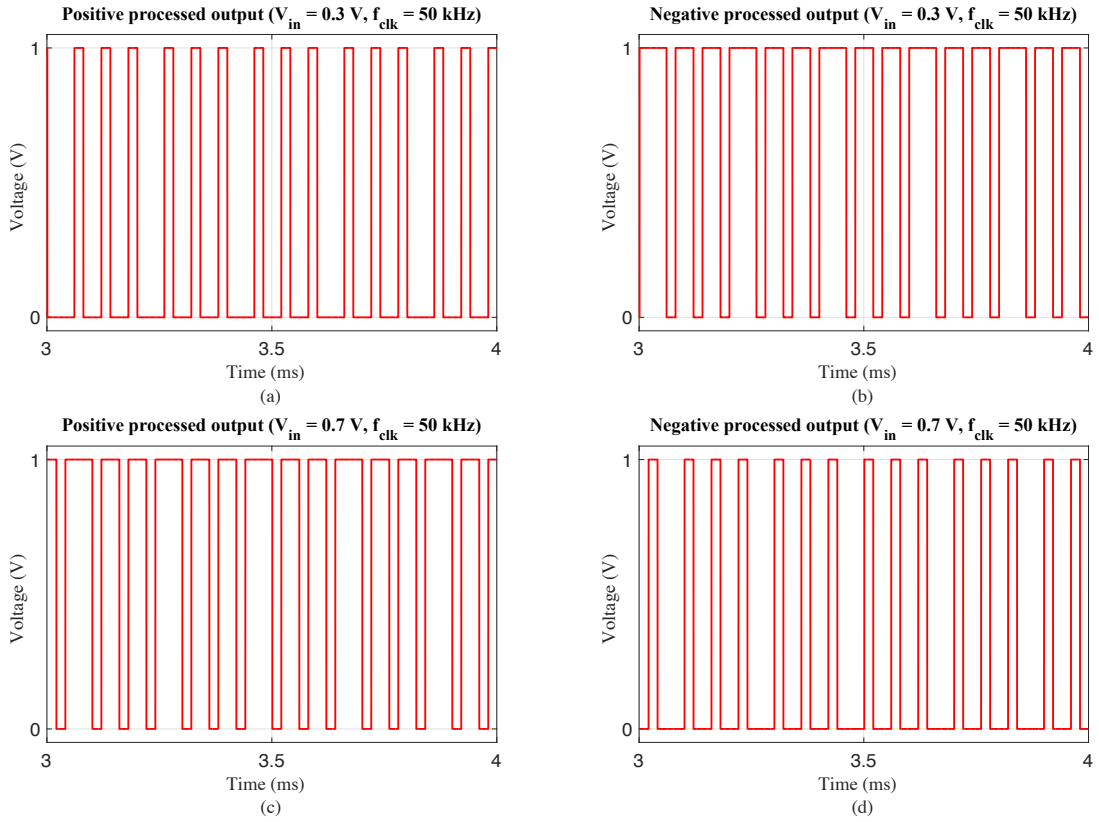

Figure 4.1: $\Delta\Sigma$ encoder's output digital signal to: **a)** positive-processed 300 mV DC input **b)** negative-processed 300 mV DC input **c)** positive-processed 700 mV DC input **d)** negative-processed 700 mV DC input.

4.1.1 Synchronous $\Delta\Sigma$ A-S Converter

This neuron produces spikes directly from the output of the $\Delta\Sigma$ encoder. For example, the $\Delta\Sigma$ circuit is simulated with DC input signals of 300 and 700 mV, and the result is shown in Figure 4.1. In this simulation, when V_{in} = 300 mV, the $\Delta\Sigma$ encoder produces the repeating sequence of '0001001001' at its positive-processed output, as shown in Figure 4.1a. As a result, the output bit is '1' in 3 out of 10 bits representing the 300 mV input. Since the negative-processed output complements its positive one, we see the repeating sequence of '111011011' on this output, produced by V_{in} = 300 mV, in Figure 4.1b. Furthermore, as a result of (3.2), $V_o^+(V_{in} = 0.7) = V_o^-(V_{in} = 0.3)$, and $V_o^-(V_{in} = 0.7) = V_o^+(V_{in} = 0.3)$, which are depicted in Figures 4.1c and 4.1d.

The synchronous $\Delta\Sigma$ A-S converter's spiking output results from the $\Delta\Sigma$ modulated signal shown in Figure 4.1. Figure 4.2 illustrates how this input neuron produces spikes from the digital

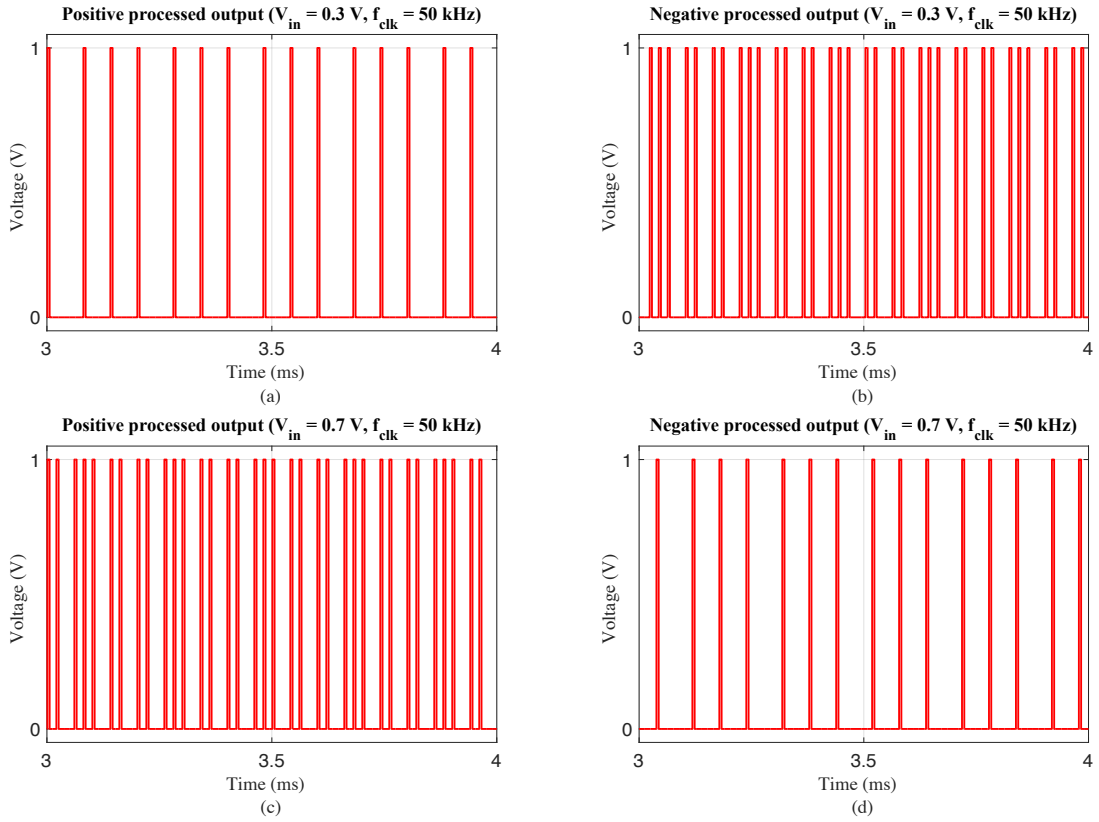

Figure 4.2: Spiking output of the synchronous $\Delta\Sigma$ neuron to: **a)** positive-processed 300 mV DC input **b)** negative-processed 300 mV DC input **c)** positive-processed 700 mV DC input **d)** negative-processed 700 mV DC input.

output of the $\Delta\Sigma$ circuit whenever a '1' bit is monitored at each clock cycle. In this setup, the minimum ISI is 20 μs, equal to the clock period. Hence, the maximum firing frequency is 50 kHz and happens when the input is V_{DD}. In the case of the 300 mV input, 3 out of 10 bits are '1' at the positive-processed output, producing 3 spikes in each 10×20 μs. On the other hand, when the input is 700 mV, the neuron produces 7 spikes at the positive output in the same period. The consecutive spikes in Figure 4.2b and 4.2c are because of consecutive '1' bits at the $\Delta\Sigma$ output signal.

The significant specification to measure in the DC input signal test is the linearity of the encoding. The average spiking frequency is measured with DC input voltages from 0 to 1 V, and the result is plotted in Figure 4.3. The result of this analysis must follow (4.2) for the positive and

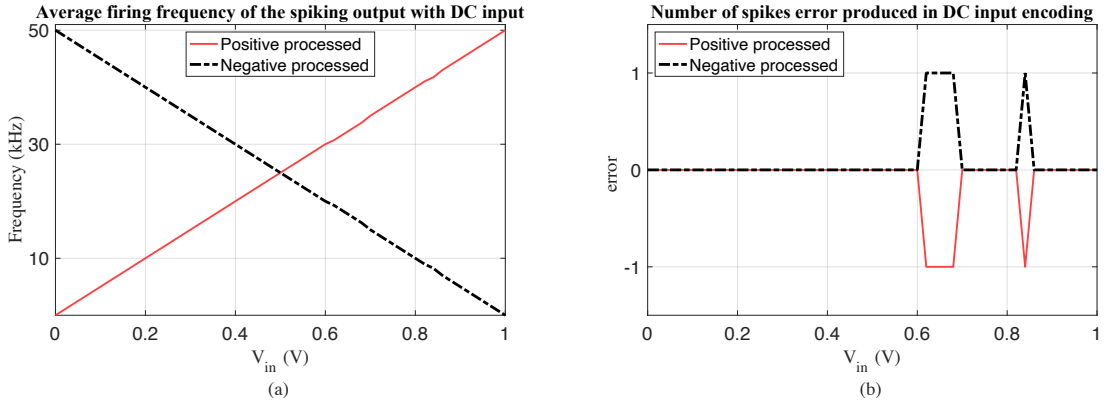

Figure 4.3: **a)** The average spiking frequency of the synchronous $\Delta\Sigma$ neuron's circuit in 10 ms runtime with respect to different DC input voltages **b)** The encoding error in spike numbers in 10 ms runtime.

negative-processed output when the clock frequency is 50 kHz:

$$\begin{aligned} f_{avg}^{spikes}(V_o^+) &= \frac{V_{in}}{V_{DD}} \times 50 \text{ kHz} \\ f_{avg}^{spikes}(V_o^-) &= \frac{1 - V_{in}}{V_{DD}} \times 50 \text{ kHz} \end{aligned} \tag{4.2}$$

The measured frequency of Figure 4.3a, f_{meas}, is compared to its reference frequency, f_{ref}, calculated from (4.2). As a result, the linearity error of the synchronous $\Delta\Sigma$ input neuron is calculated in (4.3):

$$\text{linearity error } = \frac{f_{meas} - f_{ref}}{f_{ref}} \tag{4.3}$$

After calculating the linearity error for 50 points of V_{in} between 0 V and 1 V, the maximum error is 0.41% in the positive-processed output and 0.79% in the negative-processed one in 10 ms runtime. Both processed outputs have a maximum of one spike error in their encoding, as shown in Figure 4.3b. Since the reference frequency when the error is not zero is smaller in the negative-processed output, its spiking frequency's relative error is more significant. The number of spikes error does not change by increasing the simulation runtime. Since this error is produced in a longer time window, the resulting relative frequency error gets smaller. For instance, in 50 ms time window, the maximum produced error is reduced to 0.15%. The linearity error is tiny, and the accuracy of encoding the DC input gets closer to 100% by increasing the observation window further.

4.1.2 Neuromorphic $\Delta\Sigma$ A-S Converter

This input neuron has a synapse that converts the $\Delta\Sigma$ signal into a current received by its post-synaptic IF neuron. The neuromorphic $\Delta\Sigma$ circuit's control parameters are set to have a maximum spiking frequency of $\frac{f_{clk}}{n}$. As a result, n '1's at the output bit sequence of the $\Delta\Sigma$ encoder increases the membrane voltage to its firing threshold. In theory, calculated in the previous chapter, $n = 3$ results in a uniform spike distribution in time with DC input. But in practice, with design parameters of Table 3.5 and 3.6, the maximum spiking frequency is 17.8 kHz, resulting in $n = 2.8$.

One of the benefits of using the neuromorphic encoder instead of the synchronous one is that it can consist of important voltages and currents (V_{mem}, I_{mem}, V_w), which can be accessed by the user for monitoring or adjusting as other layers' neurons. For example, if a system needs training weights in the input layer neurons, it can only use the neuromorphic $\Delta\Sigma$ A-S converter because it is the only one with a synapse in its design. Its additional advantage is that the firing pattern of the input neuron can be more analogous to other neurons' spiking behavior. In the case of the DC input, a typical IF neuron fires uniformly distributed spikes in time, and the ISIs are constant, which can be resembled by the neuromorphic $\Delta\Sigma$ A-S converter.

Figure 4.4 shows this neuron's membrane voltage and output spikes with the same DC inputs used to simulate the synchronous $\Delta\Sigma$ circuit. As displayed, we do not see double or triple consecutive spikes, unlike the synchronous one. Thus, the spikes are distributed more uniformly in time. In this simulation, the input signal of the synapse is modulated as Figure 4.1. Whenever the synapse receives a '1', it injects a current into the membrane, and its voltage increases. Inversely, if the received bit is '0', the synaptic current is 0 A, and the membrane voltage remains still, as shown in Figure 4.4.

The final examination with DC inputs is the linearity of encoding analog voltages into spikes. Like the synchronous input neuron, the number of spikes in a time window defines the average spiking frequency as (4.1). Figure 4.5a depicts the mean spiking frequency of the neuromorphic input neuron with different DC input signals. In this setup, n is 2.8, and the firing frequency is calculated in (4.4).

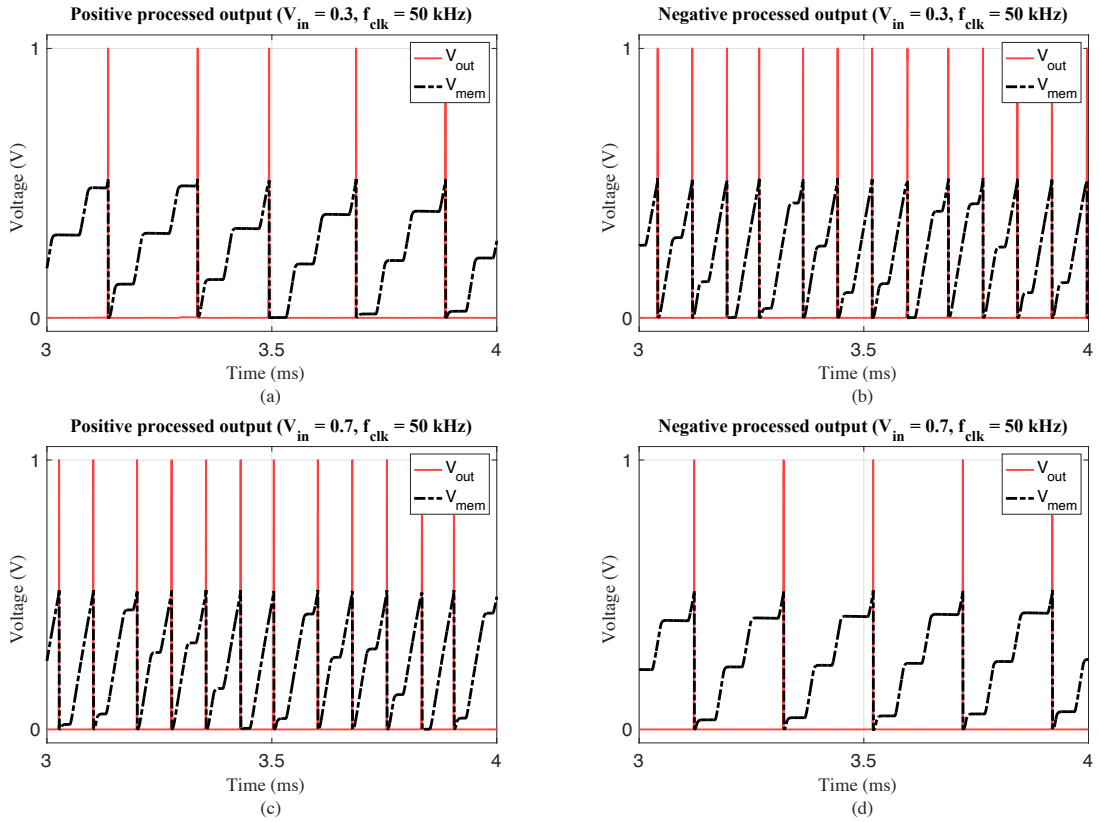

Figure 4.4: Spiking output and membrane voltage of the neuromorphic input neuron to: **a)** positive-processed 300 mV DC input **b)** negative-processed 300 mV DC input **c)** positive-processed 700 mV DC input **d)** negative-processed 700 mV DC input.

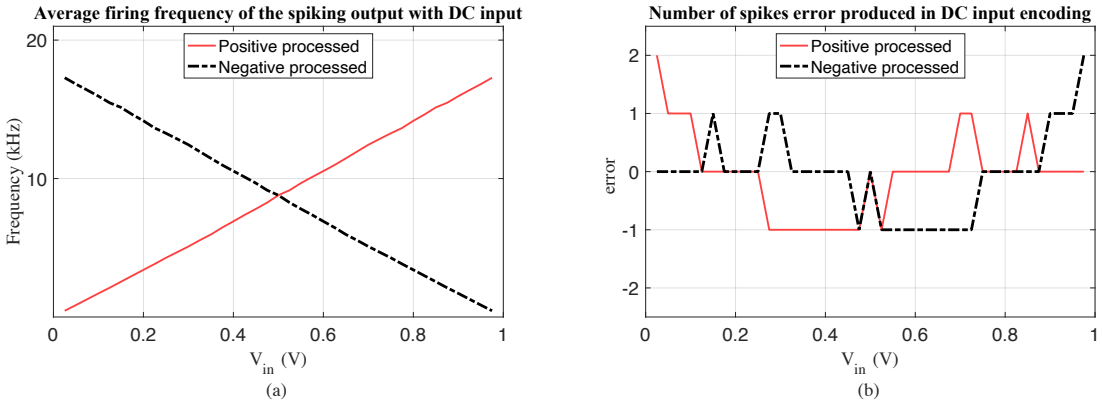

Figure 4.5: **a)** The average spiking frequency of the neuromorphic $\Delta\Sigma$ neuron's circuit in 10 ms runtime with respect to different DC input voltages **b)** The encoding error in spike numbers in 10 ms runtime.

$$f_{avg}^{spikes}(V_o^+) = \frac{V_{in}}{V_{DD}} \times \frac{50\text{ kHz}}{n}$$
$$f_{avg}^{spikes}(V_o^-) = \frac{1 - V_{in}}{V_{DD}} \times \frac{50\text{ kHz}}{n} \tag{4.4}$$

As expected, because of leakage, the neuromorphic input neuron loses accuracy when the $\Delta\Sigma$ circuit's output signal is '0' for a long time. This effect reduces the membrane voltage very slowly and only affects the system when the spiking frequency is low. Although the circuit of the IF neuron in this input encoder improves accuracy by removing leakage, it cannot be entirely eliminated. The error in the number of spikes produced at the output of the neuromorphic input neuron is shown in Figure 4.5b. As shown, the positive-processed encoding is not linear when the input is below 0.1 V. For the same reason, the negative-processed encoding faces non-linearity when the input is above 0.9 V. In these ranges, the number of spikes is small, and a 1-2 spikes error produces a huge non-linearity. Therefore, the input range selected for this A-S converter is between 0.1 and 0.9 V in a $V_{DD} = 1$ V circuit. Thus, when the input voltage is in this range, the maximum number of spikes error is only one spike in both positive and negative outputs. The maximum relative frequency error calculated in (4.3) is 2.73% for both outputs. Therefore, the method used in the neuromorphic input neuron has an encoding accuracy of more than 97% in its defined input voltage range.

4.2 AC Input

The speed of changes in the input signal affects the accuracy of the encoding. In this section, input neurons are simulated with input signals with different frequencies. The SNR is a criterion that defines the accuracy of encoding an AC input signal. To have an 8-bit resolution at the output using a $\Delta\Sigma$ encoding method, 50 dB of SNR is needed [7].

4.2.1 Synchronous $\Delta\Sigma$ A-S Converter

A single-frequency sinusoidal signal is the input signal of this encoding system. In the first step, a 100 Hz sine wave with 0.3 V amplitude is applied to the input. Because of the behavior of the $\Delta\Sigma$ encoders, we expect to see that the 100 Hz tone has the highest power in the power spectral density (PSD) of the output signal. As shown in Figure 4.6a, the density of '1's at the output bit sequence

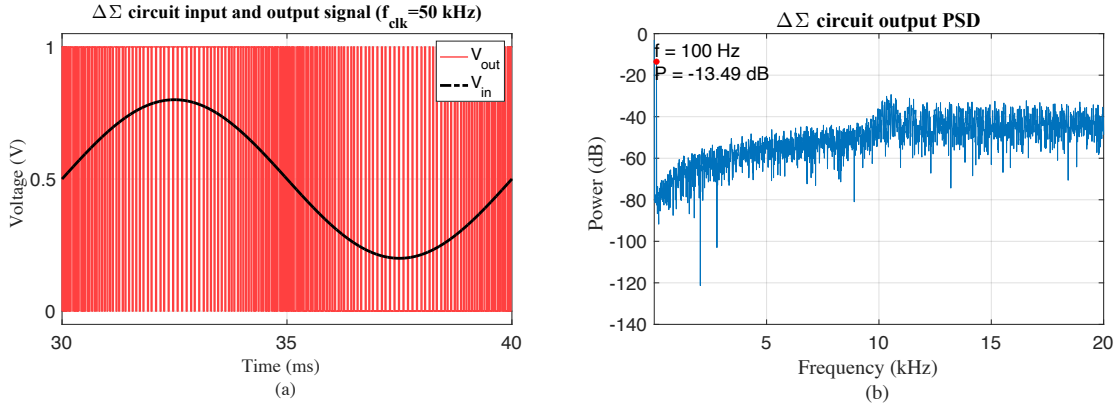

Figure 4.6: **a)** The output signal of the $\Delta\Sigma$ circuit to a sinusoidal input with frequency of 100 Hz. **b)** The PSD of the $\Delta\Sigma$ circuit's output with main harmonic at 100 Hz

of the $\Delta\Sigma$ encoder follows the sinusoidal input value. After 100 ms transient time simulation, the PSD of the $\Delta\Sigma$ encoder's output signal is plotted in Figure 4.6b. In this figure, the main power of the signal is at 100 Hz with a value of -13.49 dB. The amplitude of the sine wave does not affect the power of signal and noise in the output if the amplitude is above 0.1 V. The SNR degrades when the wave amplitude is smaller than 0.1 V. Because of the $\Delta\Sigma$ encoding method of noise shaping, the noise power is filtered in the signal's bandwidth. The Signal-to-Noise Ratio (SNR) is 53.2 dB in the input signal's bandwidth for this signal, calculated as below:

$$\text{SNR} = \frac{P_{signal}}{\sum P_{noise}} = \frac{P_{signal}}{\sum_{f \neq f_{signal}}^{2 \times f_{signal}} P_{PSD}} \tag{4.5}$$

where $\sum P_{noise}$ is calculated from each frequency's power in the PSD up to a bandwidth, excluding the signal's power. The integration bandwidth is the input signal's Nyquist frequency, which is twice the input signal's bandwidth. It is assumed that low-pass filtering of the output attenuates higher frequencies to decode the input signal.

The purpose of the presented input neuron is to encode analog input values. Using a digital encoder can have more advantages if the input is only a DC signal. Therefore, this system must be able to encode a changing analog input in real time with an acceptable resolution. The standard used in this section to analyze the encoding speed of the $\Delta\Sigma$ A-S converters is the same as ADCs [7] since there was no standard method in spiking applications. Larger SNR means a better resolution

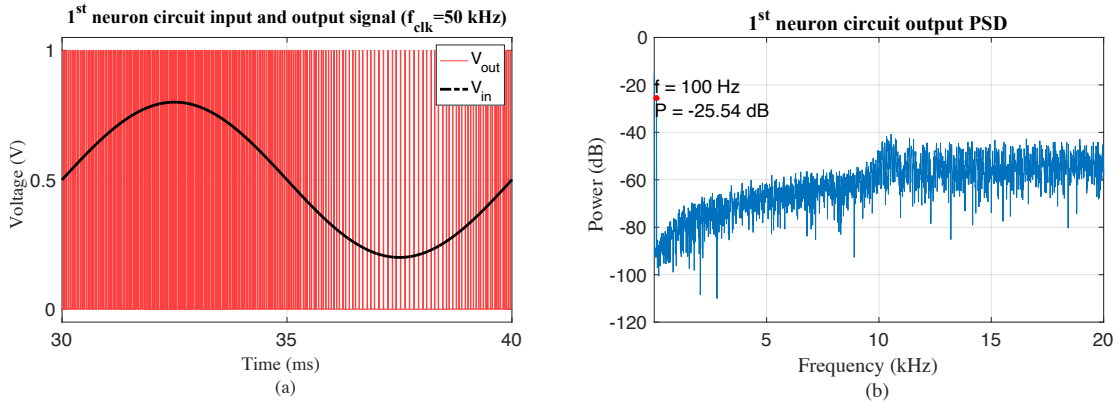

Figure 4.7: **a)** The output signal of the synchronous $\Delta\Sigma$ input neuron's circuit to a sinusoidal input with a frequency of 100 Hz. **b)** The PSD of the spiking output with main harmonic at 100 Hz

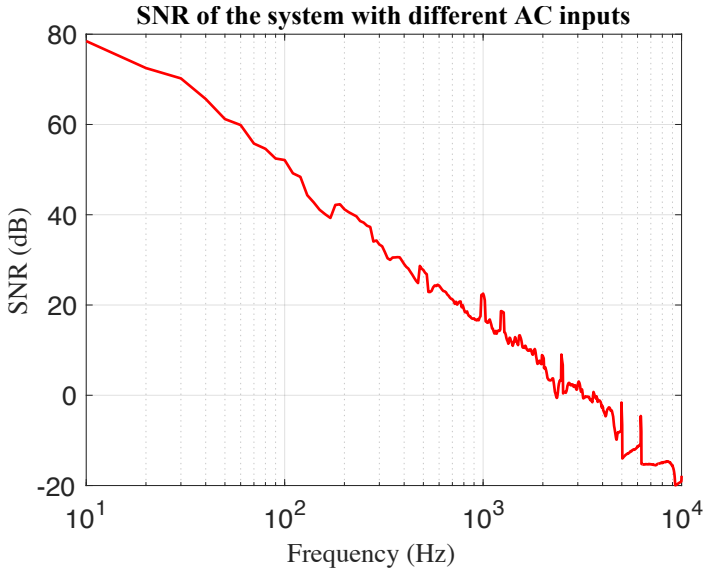

Figure 4.8: SNR of the synchronous $\Delta\Sigma$ input neuron calculated with sinusoidal inputs at different frequencies (f_{clk} = 50 kHz)

of encoding. Therefore, SNR decreases with respect to the input frequency because the encoder cannot follow the changing speed of the input.

The output signal of the synchronous input neuron is the spiking version of the $\Delta\Sigma$ encoder's output. As shown in Figure 4.7a, the spiking frequency follows the sinusoidal input's value. Next, Figure 4.7b shows the spiking output's PSD. Compared to Figure 4.6b, the signal loses 12 dB power because the duration that the spiking output is '1' is $\frac{1}{4}$ of this duration in the $\Delta\Sigma$ modulated output.

Noise power is also reduced for the same reason, and the total SNR in this neuron is 52.1 dB. Since the noise power does not depend only on the '1' duration, the same SNR as the $\Delta\Sigma$ modulating cannot be maintained, and it is decreased by 1.1 dB

Figure 4.8 shows how the SNR of the system degrades by increasing the input signal's frequency. In this simulation, the clock frequency is 50 kHz for all inputs. Furthermore, the input is a sinusoidal signal whose frequency changes between 10 Hz and 10kHz. At each frequency, the PSD of the spiking output is calculated, and the SNR is derived from that. The integration bandwidth for SNR calculation in (4.5) is twice the input's frequency for each sinusoidal input.

The result of the synchronous input neuron in this section was for the positive-processed output. Since the outputs complement each other, the same results are also valid for the negative-processed output. Under the mentioned circumstances, this neuron can encode the changing input signal to a bandwidth of 120 Hz with more than 50 dB SNR. Moreover, as shown in Figure 4.8, the SNR is positive when the input's bandwidth is below 2.2 kHz.

4.2.2 Neuromorphic $\Delta\Sigma$ A-S Converter

This input encoder neuron is examined with the same test benches as the previous one. First, the input is set to a sinusoidal signal with a frequency of 100 Hz. The synapse and IF neuron create the spiking output from this input as shown in Figure 4.9a with a clock signal of 50 kHz. In the spiking output, the ISIs get smaller when the input increases and larger when the input decreases. This circuit's output in time is transformed to the frequency domain, and its PSD is plotted in Figure 4.9b. As shown, the main power of the signal is at 100 Hz with a value of -20.75 dB. Compared to the synchronous input neuron, the in-band noise has greater power in the neuromorphic one. In the next step, the SNR of the output signal is calculated with the signal and noise power in the PSD as (4.5). As a result, with 100 Hz sinusoidal input, the SNR is 46.3 dB.

This input encoder's output SNR is also measured with different AC input frequencies. The result of this test is shown in Figure 4.10. Since the spiking frequency to the same signal is lower in this neuron compared to the first one, a lower input bandwidth is expected. Considering the specification of 50 dB SNR, defined in (4.5), the neuromorphic $\Delta\Sigma$ A-S converter can encode inputs to a bandwidth of 65 Hz. At 1.3 kHz, the SNR is 0 dB, and the signal is not detectable from

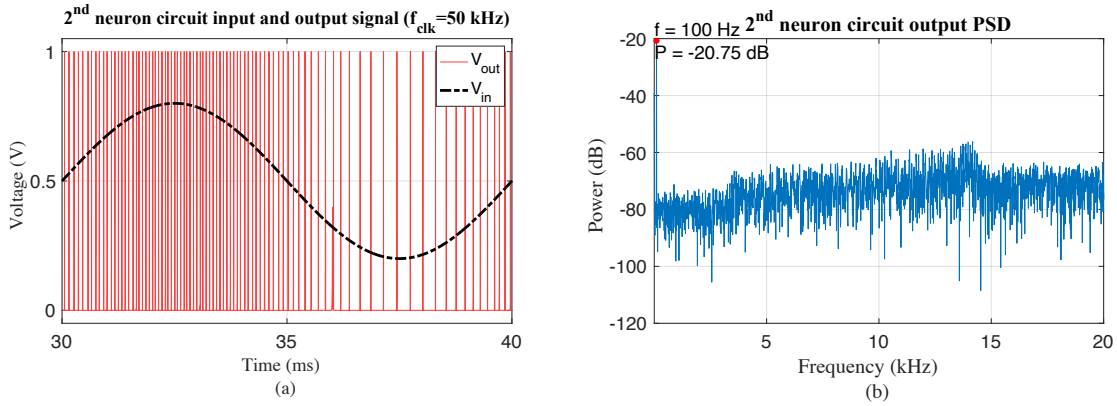

Figure 4.9: **a)** The output signal of the neuromorphic $\Delta\Sigma$ A-S converter's circuit to a sinusoidal input with a frequency of 100 Hz. **b)** The PSD of the spiking output with main harmonic at 100 Hz

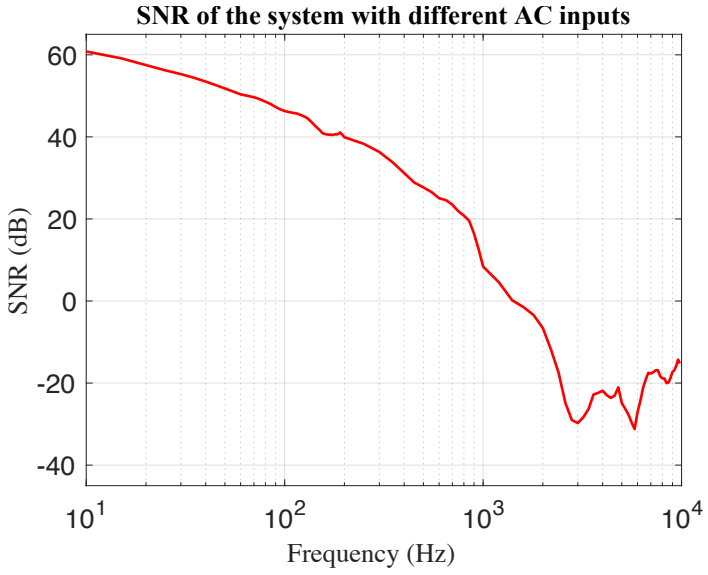

Figure 4.10: SNR of the neuromorphic $\Delta\Sigma$ A-S converter calculated with sinusoidal inputs at different frequencies (f_{clk} = 50 kHz)

the noise above this frequency.

4.3 Power and Energy

Power consumption is a critical part of a neuron design because many neurons work with each other in an SNN chip. How each section consumes power is different because signals at various nodes can be spikes, digital, or analog. If a part of the system draws a constant current from the power source, its power consumption is static and independent of the firing frequency. On the other hand, if the current changes with the arrival of the spikes, the power is dynamic. This section simulates the designed neurons' static or dynamic power and energy.

4.3.1 Synchronous $\Delta\Sigma$ A-S Converter

There are three main components in the synchronous $\Delta\Sigma$ input encoder regarding the power consumption: the Op-Amp, comparator, and spike generator DFF where the first two are parts of the $\Delta\Sigma$ circuit. The energy consumption of the Synchronous $\Delta\Sigma$ and its components are shown in Figure 4.11 in 1 ms runtime. The energy is calculated in (4.6):

$$E(t) = \int V_{DD} I_{DD}(t)\, dt \tag{4.6}$$

where I_{DD} is the current provided by the power source at time t.

In Figure 4.11a, when firing at the maximum frequency, the $\Delta\Sigma$ circuit consumes a static energy of 12 nJ in 1 ms where the Op-Amp uses most of it. In Figure 4.11b, the spike generator DFF consumes energy only in times of spikes; however, its resulting energy consumption is smaller than the $\Delta\Sigma$ circuit. It consumes almost 28 pJ/spike. Figure 4.11c shows the energy consumption of the $\Delta\Sigma$ circuit remains the same when firing at half of the maximum frequency. On the other hand, the spike generator's energy consumption is approximately halved, as shown in Figure 4.11d. The energy and power of this system are measured in 3 firing frequencies in Table 4.1.

It is concluded from Table 4.1 that the $\Delta\Sigma$ circuit's energy and power consumption does not depend on the firing frequency, and it remains constant. Hence, increasing the spiking frequency

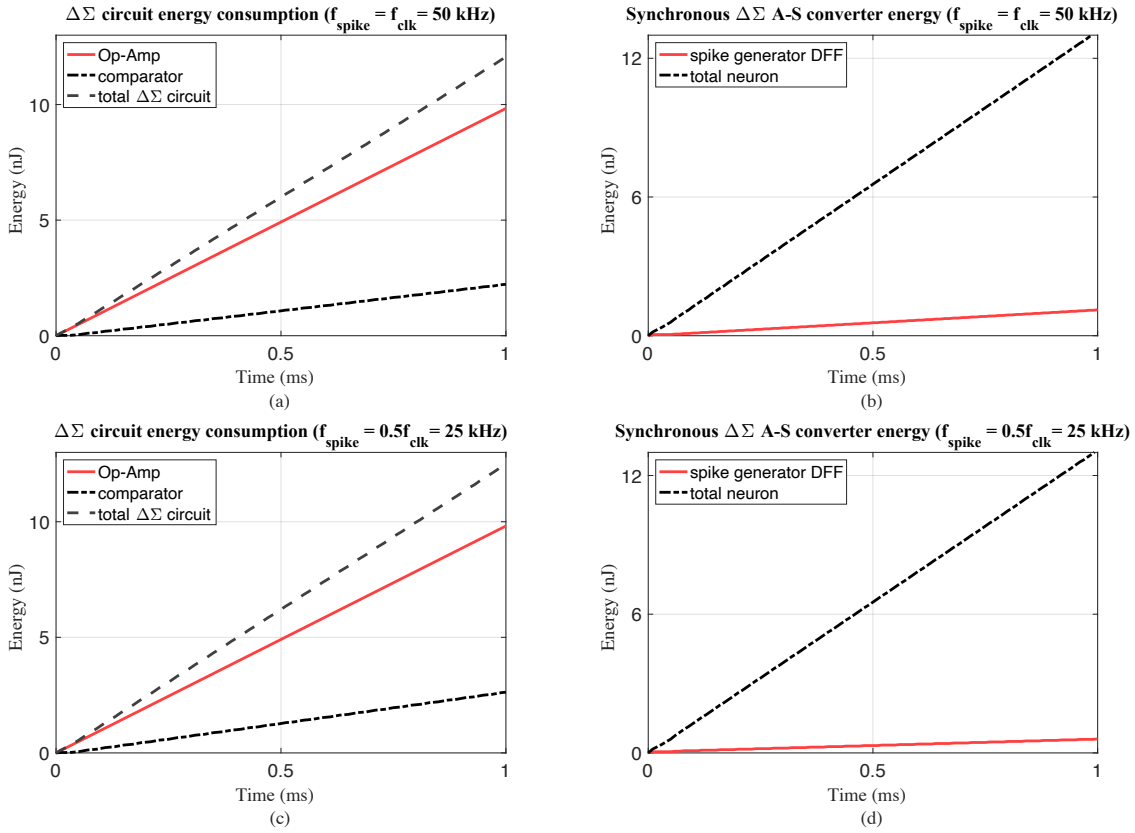

Figure 4.11: Transient energy consumption of: **a)** $\Delta\Sigma$ circuit's components when firing at the maximum frequency **b)** spike generator DFF and complete synchronous $\Delta\Sigma$ neuron when firing at the maximum frequency **c)** $\Delta\Sigma$ circuit's components when firing at half of the maximum frequency **d)** spike generator DFF and complete synchronous $\Delta\Sigma$ neuron when firing at half of the maximum frequency

reduces its energy per spike. On the other hand, the spike generator DFF's energy consumption depends entirely on the spike arrivals. At each spike, the consumed energy increases by almost 28 pJ. The complete circuit of the synchronous $\Delta\Sigma$ A-S converter consumes dynamic and static power from each part of its system, but its static power is dominant. Increasing the input from 0 V to V_{DD} increases the system's power consumption by 11.6%. Thus, its energy consumption increases both in time and at spike arrivals.

Table 4.1: Power and energy of the synchronous $\Delta\Sigma$ A-S converter in 1 ms

component	$\Delta\Sigma$ circuit			spike generator DFF			synchronous $\Delta\Sigma$ A-S converter		
spiking frequency	0 kHz	25 kHz	50 kHz	0 kHz	25 kHz	50 kHz	0 kHz	25 kHz	50 kHz
energy	12 nJ	12 nJ	12 nJ	0 nJ	0.7 nJ	1.4 nJ	12 nJ	12.7 nJ	13.4 nJ
energy/spike	-	0.48 nJ	0.24 nJ	-	28 pJ	28 pJ	-	0.51 nJ	0.27 nJ
average power	12 μW	12 μW	12 μW	0 μW	0.7 μW	1.4 μW	12 μW	12.7 μW	13.4 μW

4.3.2 Neuromorphic $\Delta\Sigma$ A-S converter

The spike generator circuit in the neuromorphic $\Delta\Sigma$ A-S converter is an IF neuron and a synapse. These circuits operate in the sub-threshold domain of MOSFETs, and one of the advantages of the sub-threshold region is that it consumes a small current. The energy consumption of the neuromorphic $\Delta\Sigma$ and its components are shown in Figure 4.12.

In Figure 4.12a, when firing at maximum frequency, the input of the synapse is always '1' received from the $\Delta\Sigma$ circuit. Thus, it consumes an average power of 400 nW when the circuit receives '1' and produces a current. At the same time, the IF neuron's power consumption happens when it creates a spike. In the time window of Figure 4.12a, it fires 3 times, consuming 18.4 pJ per spike. As expected, it is shown in Figure 4.12b and 4.12d that the dominant power-consuming part of this system is the $\Delta\Sigma$ encoder draining 12 μW on average. The energy that is used by the neuromorphic spike generator part is neglectable. At last, in Figure 4.12c, we see that the synapse is draining a current almost half of the time when the input of the $\Delta\Sigma$ circuit is 0.5 V. In this situation, the IF neuron only produced one spike consuming 18.7 pJ. The energy and power of this system are measured in 3 firing frequencies in Table 4.2.

It is concluded from Table 4.2 that the synapse's energy consumption depends on the output bit of the $\Delta\Sigma$ circuit, and the IF neuron's energy depends on whether it creates a spike or not. The power of the neuromorphic $\Delta\Sigma$ A-S converter does not change in time, neglecting the effect of the synapse and IF neuron. We see a slight increase of 700 nW in the power of the total system when the input rises from its minimum to maximum voltage. As a result, the power of this system is around 12 μW, and the input signal can change it only by 6%.

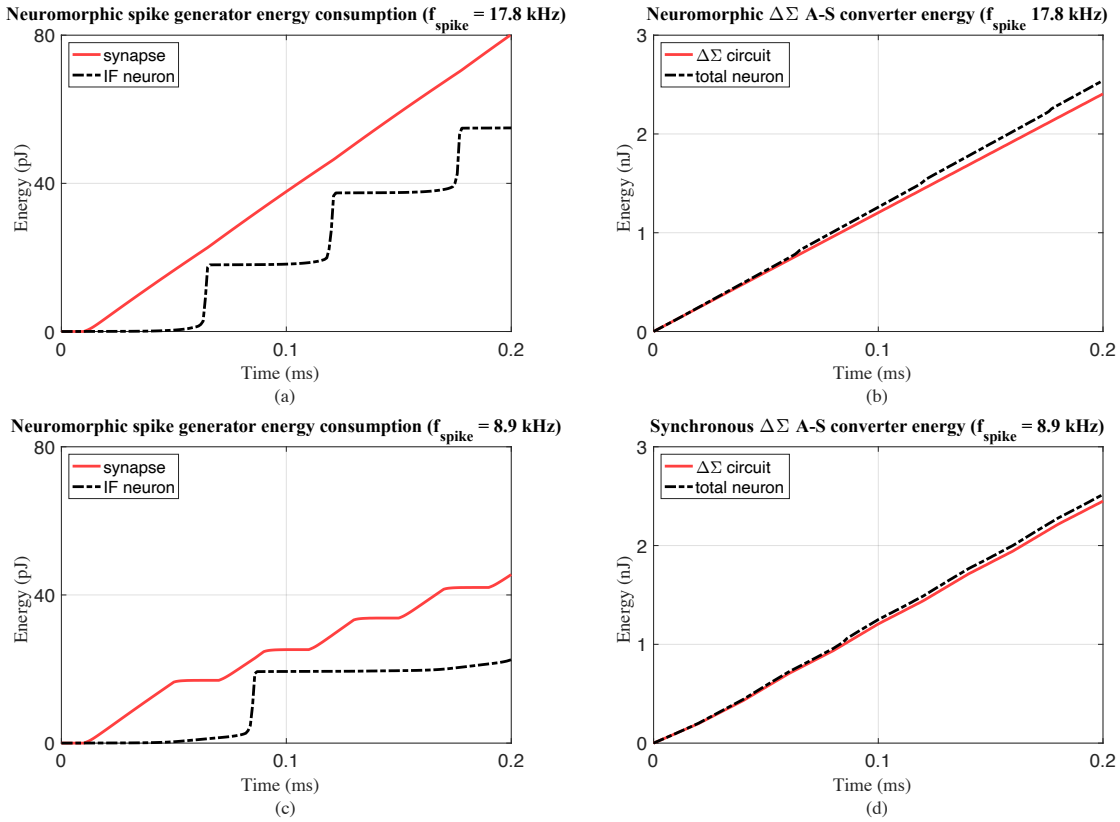

Figure 4.12: Transient energy consumption of **a)** the DPI synapse and the IF neuron firing at the maximum frequency **b)** $\Delta\Sigma$ circuit and complete A-S converter at the maximum frequency **c)** the DPI synapse and the IF neuron firing at half of the maximum frequency **d)** $\Delta\Sigma$ circuit and complete A-S converter at half of the maximum frequency

Table 4.2: Power and energy of the neuromorphic $\Delta\Sigma$ A-S converter in 1 ms

component	**DPI synapse**			**IF neuron**			**neuromorphic $\Delta\Sigma$ A-S converter**		
spiking frequency	0 kHz	8.9 kHz	17.8 kHz	0 kHz	8.9 kHz	17.8 kHz	0 kHz	8.9 kHz	17.8 kHz
energy	0 pJ	220 pJ	400 pJ	0 pJ	150 pJ	312 pJ	12 nJ	12.3 nJ	12.7 nJ
energy/spike	-	24.4 pJ	23.5 pJ	-	18.7 pJ	18.4 pJ	-	1.5 nJ	747 pJ
average power	0 nW	220 nW	400 nW	0 nW	150 nW	312 nW	12 μW	12.3 μW	12.7 μW

4.4 Area

The area is another important criterion that has to be determined in each neuron. Since each encoder is copied in the SNN system as many times as the number of input neurons, each one has to be designed with the minimum possible area. Since the simulations are done on schematics in this work, the layout is not executed. The area used by MIM capacitors is defined, and we have to estimate the area of MOSFETs. The area for each circuit is estimated based on an inverter designed

with $\frac{W}{L}(\text{PMOS}) = \frac{5.6\ \mu\text{m}}{65\ \text{nm}}$, and $\frac{W}{L}(\text{NMOS}) = \frac{1.3\ \mu\text{m}}{65\ \text{nm}}$. This inverter occupies 9 μm^2. Moreover, the MOSFETs and capacitors are implemented in different metals of the chip. Thus, they can share an area. Therefore, the area of the entire chip is estimated based on its MOSFETs or capacitor's area, depending on which one of them is larger. Every component is used from TSMC 65 nm library.

4.4.1 Synchronous $\Delta\Sigma$ A-S Converter

In this system, the area is used by two capacitors and different MOSFETs. The capacitors are 500 and 250 fF and digital and analog MOSFETs are the remaining parts. The area used by each part is shown in Table 4.3.

Table 4.3: Area of different parts of synchronous $\Delta\Sigma$ A-S converter

	$\Delta\Sigma$ circuit					**synchronous spike generator**	
component	C_f (MIMCap)	C_s (MIMCap)	digital circuits	Op-Amp	comparator	monitoring DFF	resetting circuit
area	$243.36\mu\text{m}^2$	$121\mu\text{m}^2$	$165\mu\text{m}^2$	$98\mu\text{m}^2$	$30\mu\text{m}^2$	$50\mu\text{m}^2$	$46\mu\text{m}^2$

As shown in the table, The capacitive area is estimated as 420 μm^2 on the top metal. On the other hand, the MOSFETs occupy an estimated area of approximately 400 μm^2. Therefore, we can assume that the entire area of this chip is smaller than 500 μm^2.

4.4.2 Neuromorphic $\Delta\Sigma$ A-S Converter

The neuromorphic $\Delta\Sigma$ spike generator consists of two 1 pF and a 20 pF capacitors. Thus, more than 99% of the area is used by the capacitors, and the area of other circuit parts is insignificant. Furthermore, the $\Delta\Sigma$ circuit is the same as the previous one, and the area used by C_s and C_f are calculated in the previous part. Table 4.4 shows the area of different parts of this neuron.

Table 4.4: Area of different parts of neuromorphic $\Delta\Sigma$ A-S converter

	$\Delta\Sigma$ circuit	**DPI synapse**	**IF neuron**	
component	complete circuit	C_{syn}	C_{mem}	C_{ref}
area	$500\mu\text{m}^2$	$492.84\mu\text{m}^2$	$9856.8\mu\text{m}^2$	$492.84\mu\text{m}^2$

Most area is used by the membrane capacitor, which is 20 pF. In this system, the synaptic circuit produces the membrane current between 100 and 500 nA. With this current, the membrane capacitor cannot be reduced for this setup to have an acceptable firing frequency with a clock frequency

of 50 kHz. As a result, this neuron uses 0.011 mm^2, which is almost 22 times larger than the synchronous one.

4.5 Changing the Clock Frequency

Up to this section, the clock had a fixed frequency of 50 kHz, and all parts of the circuit were optimized for this clock signal. The clock signal directly correlates with the maximum firing frequency of both neurons. The synchronous $\Delta\Sigma$ A-S converter's circuit can perform with different clock frequencies. The bandwidth of the OTA and the delay of digital circuits limit the behavior of the $\Delta\Sigma$ circuit. The dominant delay of the digital part is set by the clock-Q delay of the DFF, which is 1 ns. On the other hand, the gain-bandwidth product of the OTA 20.2 MHz. The input of the OTA is the sampled signal at the clock frequency, which changes its value with the clock speed. Therefore, the OTA must be able to follow these changes, and it cannot provide a gain for clock frequencies near 20.2 MHz. Therefore, the transfer function of the switch-cap system is not an integrator in this region. Furthermore, the OTA is expected to lose performance in smaller frequencies than the clock speed, which will be explored in this section with different input voltages. Since the $\Delta\Sigma$ circuit is shared with both synchronous and neuromorphic $\Delta\Sigma$ A-S converter, they cannot perform for larger than 20.2 MHz clock frequency. Furthermore, Section 3.1.2.3 shows the high-frequency spike generator of the synchronous $\Delta\Sigma$ that replaces two DFFs when the clock frequency is above 1 MHz.

In the neuromorphic $\Delta\Sigma$ A-S converter, capacitors must be updated for different clock signals. Unlike the synchronous one, the circuit designed for 50 kHz clock cannot perform well when the clock frequency is altered. In this circuit, the capacitors play an essential role in the filter's time constant, the neuron's spiking rate, and its refractory period. Hence, their value depends on the clock frequency of the $\Delta\Sigma$ circuit to perform as expected. As a result, for each clock signal, the value of the new capacitor is updated as (4.7):

$$C'_x = C_x \times \frac{50\text{ kHz}}{f'_{clk}} \tag{4.7}$$

where C'_x is the new value of any capacitor in the synapse and IF neuron, and C_x is its value with

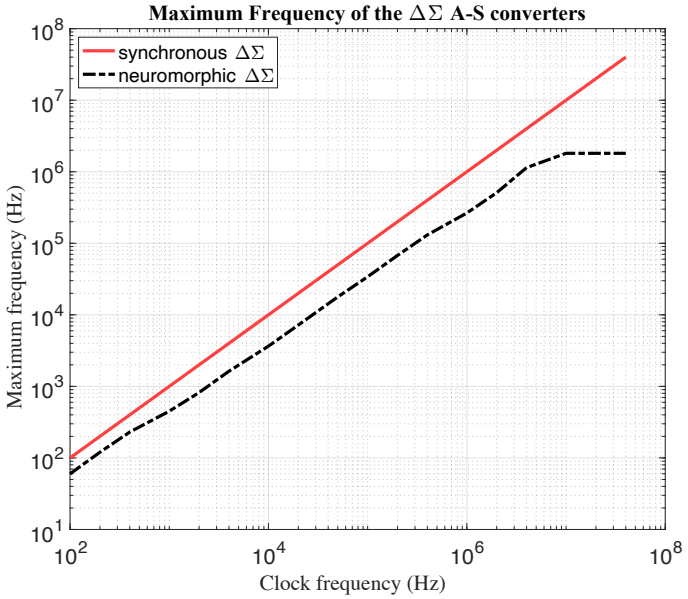

Figure 4.13: Maximum firing frequency of both $\Delta\Sigma$ A-S converters with respect to different clock frequencies.

50 kHz clock signal. Furthermore, the new clock frequency is f'_{clk}, which cannot be larger than 5 MHz in this neuron because the smallest MIMcap is 10 fF. This section reruns all of the above simulations with different clock frequencies to see the effect of the clock signal on the linearity, bandwidth, power, and area of both input encoder neurons. The selected clock signals to test these two neurons are 10^n, 2×10^n, and 4×10^n, where $n = 2, 3, 4, 5, 6, 7$.

4.5.1 DC Input

The A-S converters are updated for each clock frequency, and at the first step, the input is set to V_{DD} to measure the maximum spiking frequency. The clock signal is swept from 100 Hz to 40 MHz with 3 points in each decade. The maximum spiking frequency of each neuron is measured for each clock signal, and the result is shown in Figure 4.13.

The synchronous $\Delta\Sigma$ circuit's maximum spiking frequency always equals the clock signal's frequency. On the other hand, the neuromorphic $\Delta\Sigma$'s firing frequency gets saturated when the clock signal reaches 10 MHz. This is because of the minimum available capacitors. After 10 MHz all capacitors of the synapse and IF neuron are 10 fF, and increasing the clock signal's frequency

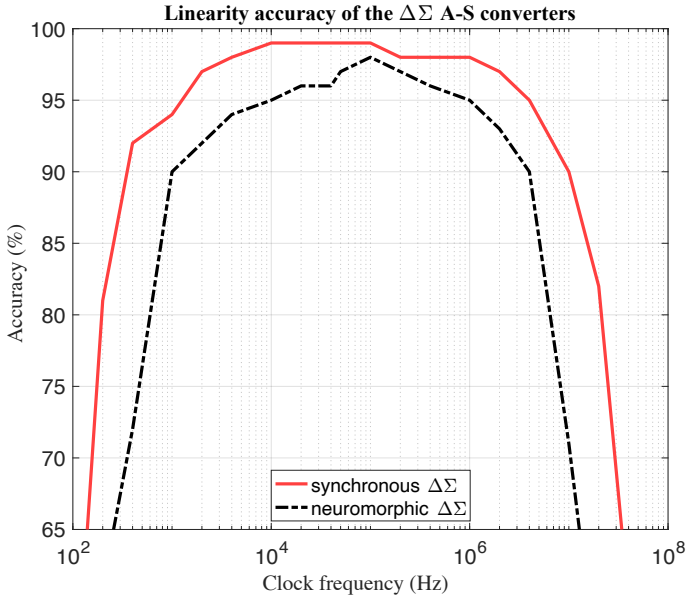

Figure 4.14: Positive-processed DC input encoding accuracy of each $\Delta\Sigma$ A-S converter with respect to different clock frequencies.

does not affect the neuron's spiking rate, which is maxed at 1.8 MHz.

The next step is to calculate the encoding accuracy for DC inputs. To compare the linearity in different frequencies, the simulation time is limited to 500 clock periods. The linearity error calculated in (4.3) determines the accuracy of encoding DC signals in the input neurons in the defined runtime. In this simulation, the error is measured for different DC inputs, and the maximum measured error is e_{max}. The linearity accuracy for both A-S converters is calculated in (4.8) for each clock frequency with DC inputs between 0.1 V and 0.9 V. The results are shown in Figure 4.14.

$$\text{linearity accuracy} = (1 - e_{max}) \times 100\% \tag{4.8}$$

In Figure 4.14, the $\Delta\Sigma$ circuit's clock signal must be in a certain range for both neurons to have an acceptable encoding accuracy. The synchronous $\Delta\Sigma$ encoder works better in lower and higher boundaries of the clock frequency compared to the neuromorphic one. As a result of this simulation, the synchronous $\Delta\Sigma$ A-S converter's linearity accuracy is above 90% when the clock signal is in the range of 400 Hz and 10 MHz. In the low range, the switch-cap part of the digital integrator cannot accumulate input data with acceptable accuracy because the capacitors are small and produce a

very large impedance at the low frequency. Every integrator's first pole should be at 0 Hz which is impossible in real circuits. Hence, the integrator is a low-pass filter with a very low corner frequency that acts as an integrator when the input's frequency is higher than the filter's corner frequency. The input of the SC integrator in the $\Delta\Sigma$ modulator changes with the clock frequency. Therefore, when the clock frequency is low, the Bode diagram magnitude of the integrator is a flat line in this region and does not act as an integrator. As a result, the system loses accuracy when the clock signal gets lower than 1 kHz. Also, in the high range of the sampling speed, the OTA's bandwidth is insufficient and cannot follow the changes in its input. The clock frequency lower and upper limit is reduced to 1 kHz and 2 MHz for the neuromorphic $\Delta\Sigma$ circuit to encode the DC input with an accuracy of larger than 90%. In this neuron, when the DC input is low, the leakage of the membrane reduces linearity. Thus, the maximum encoding error gets bigger in lower clock frequencies because of the leakage. On the other hand, when the clock signal is above 2 MHz, the capacitors of the circuit are too small, and the parasitic capacitors become comparable to them. As a result, these parasitic capacitors produce non-linearity in different firing frequencies of the neuron.

4.5.2 AC Input

In the AC input section with the clock frequency of 50 kHz, SNR of 50 dB was the criterion that measured the input bandwidth. The same standard is also applied here. In this simulation, the clock signal is set to the same frequencies between 100 Hz and 40 MHz as the DC input simulation in the previous part. The $\Delta\Sigma$ circuit receives sinusoidal input at different frequencies. Afterward, the output PSD is calculated, and the signal and noise power are measured. For each sinusoidal frequency, the SNR is evaluated as (4.5), and the input bandwidth must support an SNR higher than 50 dB. This procedure is repeated for different clock signals, and the input bandwidth is shown in Figure 4.15 for each clock frequency.

In this simulation, when the clock signal is below 1 kHz, the $\Delta\Sigma$ circuit cannot provide an SNR of 50 dB independent of the input signal's frequency. The same is true for the clock signals above 4 MHz because of the OTA's bandwidth. Moreover, the neuromorphic encoder's clock signal is limited to 2 MHz for an acceptable encoding SNR. As a result of this test, the synchronous $\Delta\Sigma$ A-S converter can have a maximum input bandwidth 4 kHz with an appropriate clock signal. Moreover,

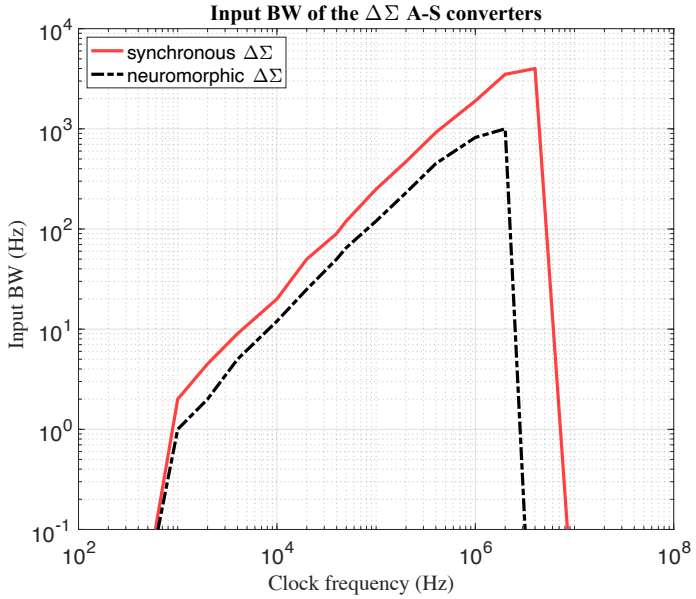

Figure 4.15: The input BW in which the neurons can encode the changing input with an SNR higher than 50 dB in different clock frequencies.

this maximum input bandwidth is 1 KHz for the neuromorphic $\Delta\Sigma$.

4.5.3 Power

The power and energy section mentioned that these input neurons burn dynamic and static power. Furthermore, two circuits were designed for the synchronous $\Delta\Sigma$ A-S converter depending on the clock frequency, in which the slow one consumes much more energy per spike. Hence, Table 4.1 shows the slow spike generator's power consumption at the clock frequency of 50 kHz. In this work, the low-power, high-speed spike generator replaces the DFF version when the clock signal reaches 1 MHz. Figure 4.16 shows the average power of the converters in different clock signals when their input is set to V_{DD}.

Both converters' maximum power dissipation is the $\Delta\Sigma$ circuit's power, which is 12 μW. The synchronous $\Delta\Sigma$'s power consumption increases with a higher slope than the neuromorphic $\Delta\Sigma$ because of its higher energy per spike and larger firing frequency. In this figure, both the high-speed and low-speed spike generator's power consumption is shown for the synchronous $\Delta\Sigma$ A-S converter. The maximum clock signal for which the low-speed one can be utilized is 400 kHz in this

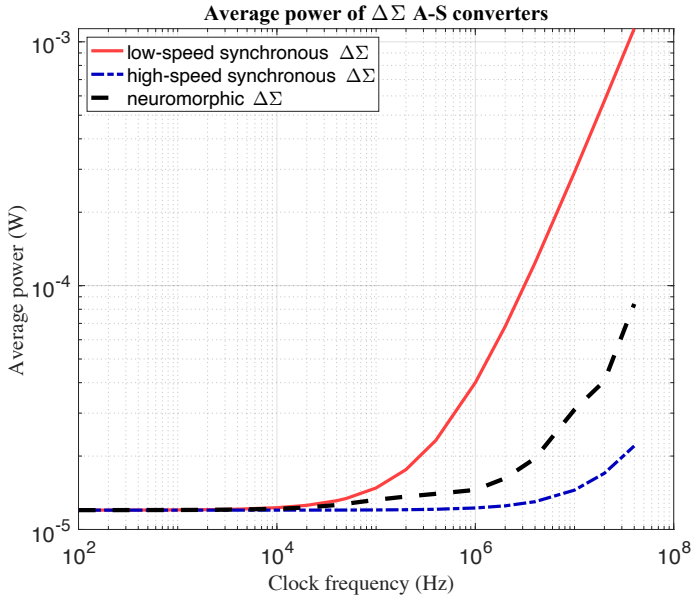

Figure 4.16: The average power consumed by each A-S converter while firing at the maximum frequency

simulation; for this frequency, the power dissipation is 23 μW approximately. On the other hand, the high-speed one's dominant power consumer is the $\Delta\Sigma$ circuit, which remains constant with spiking frequency. Moreover, the neuromorphic $\Delta\Sigma$ circuit's average power consumption slightly increases by the frequency of spikes. The significant power increase for the high-speed synchronous and the neuromorphic $\Delta\Sigma$ is seen for the clock frequencies higher than 4 MHz, which is above the operating region of the $\Delta\Sigma$ A-S converters.

4.5.4 Area

The area of each circuit depends on the size of its capacitors, shown in Figure 4.17. The synchronous neuron's capacitors do not change with respect to the clock frequency. Hence, its size remains constant at 500 μm^2. Conversely, the neuromorphic circuit's capacitors vary with respect to the clock frequency calculated in (4.7). Due to previous simulations, the minimum clock frequency for the neuromorphic $\Delta\Sigma$ to work appropriately is 1 kHz. At this frequency, the chip size is 0.54 mm^2, which is dominantly occupied by the 1 nF membrane capacitor. The area of synchronous and neuromorphic $\Delta\Sigma$ A-S converters are much closer in the high firing frequency range, and the

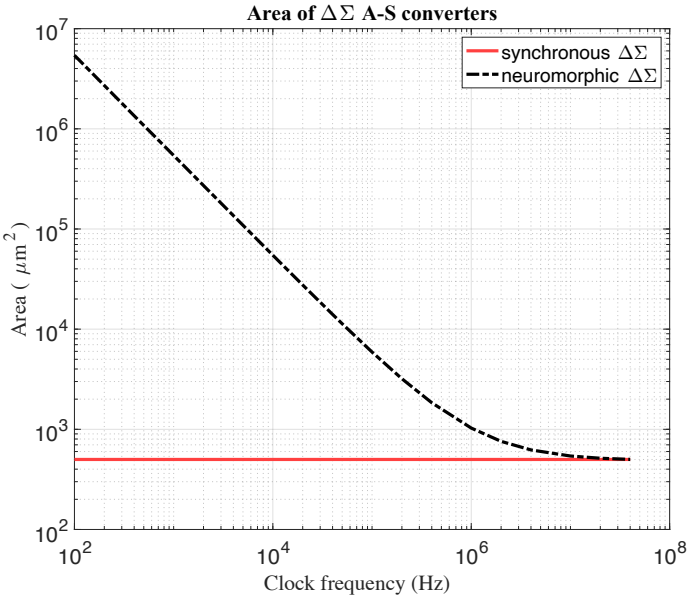

Figure 4.17: The area occupied by each A-S converter in different clock frequencies

area benefit of the synchronous one is almost diminished in this region with a high-speed clock signal.

4.6 SNN System

This work designs the input layer neuron that converts the analog value of the input signal into a spiking signal. In the previous sections of this chapter, two $\Delta\Sigma$ A-S converters were simulated in different conditions, and their reliability was evaluated. The summary of the final results is provided in Table 4.5. Each neuron's result is shown in its minimum, optimized, and maximum operating clock frequency to support a wide range of spiking rates. This work assumption by showing the resolution of the input encoder is that the designed $\Delta\Sigma$ A-S converters can replace any prior designs of input neurons in a variety of SNN systems. The comparison between the designed input neuron and other input interfaces used in SNN applications is provided in Table 4.6. All encoders in the table and the reported parameters are based on a single input neuron.

The testbench of digit classification for the MNIST dataset requires 784 input neurons to encode 28×28 grayscale digit images [37]. A digit classifier also contains 100 hidden and 10 output layer neurons in addition to the input ones. Furthermore, a feed-forward fully connected network needs

Table 4.5: Simulation result's summary of synchronous and neuromorphic $\Delta\Sigma$ A-S converters

	synchronous $\Delta\Sigma$ A-S converter			**neuromorphic $\Delta\Sigma$ A-S converter**		
clock frequency	1 kHz (min)	50 kHz (optimized)	4 MHz (max)	1 kHz (min)	50 kHz (optimized)	2 MHz (max)
spike generator	low-speed	low-speed	high-speed	IF neuron	IF neuron	IF neuron
maximum spiking frequency	1 kHz	50 kHz	4 MHz	442 Hz	17.8 kHz	510 kHz
Input range	$0.1\text{ V} - 0.9\text{ V}$	$0\text{ V} - 1\text{ V}$	$0.1\text{ V} - 0.9\text{ V}$	$0.1\text{ V} - 0.9\text{ V}$	$0.1\text{ V} - 0.9\text{ V}$	$0.1\text{ V} - 0.9\text{ V}$
maximum DC input encoding error	5.8%	0.8%	8.1%	9.8%	2.7%	7.5%
DC encoding accuracy	94.2%	99.2%	91.9%	90.2%	97.3%	92.5%
input bandwidth (SNR$>$ 50 dB)	2 Hz	120 Hz	4 kHz	1 Hz	65 Hz	1 kHz
average power	12 μW	13.4 μW	13 μW	12 μW	12.7 μW	16.2 μW
area	500 μm^2	500 μm^2	500 μm^2	0.542 mm^2	0.011 mm^2	710 μm^2

Table 4.6: Performance comparison

input encoder	**encoder types**	**maximum rate**	**linear inputs**	**input range**	**average power**	**area**
MoS_2 transistor [21]	rate, count, time	100 Hz	0, 0.5, 1..., 5 V	0 – 5 V	500 pW	NR*
asynchronous Δ modulation [10]	change of input	500 Hz	–	0 – 1.8 V	7.3 μW	0.15 mm^2
non-linear synapse [5]	positive-negative rate	400 Hz	-0.5 – -0.2V	-1 – 1 V	NR*	NR*
linear V-I converter [9]	-	neuron-dependent	0 – 0.7 V	0 – 1 V	10 μW*	$>$ 0.06 mm^2*
this work: synchronous $\Delta\Sigma$ (optimized)	positive-negative rate	50 kHz	0 – 1 V	0 – 1 V	13.4 μW	500 μm^2
this work: neuromorphic $\Delta\Sigma$ (optimized)	positive-negative rate	17.8 kHz	0.1 – 0.9 V	0 – 1 V	12.7 μW	0.01 mm^2

$*$These parameters are not reported directly in the source paper, and they are based on estimation. NR means not reported.

79400 synapses in this application. The designed input encoders of this work are replicated 784 times at the input interface and are connected with 78400 synapses to the next hidden layer. Other layers of neurons in the system can be implemented using the LIF neuron of Figure 2.7. The programmable and changeable parameters in the system are synaptic weights, synaptic current's MOSFET size, neuron's integration gain, and the clock signal of the input neuron's $\Delta\Sigma$. The

system can be trained by changing the weight voltage of each synapse.

Based on scaling and estimation, the first complete network can be built with 784 synchronous $\Delta\Sigma$ inputs and 110 other layers' neurons connected with 79400 synapses. This network occupies almost 2 mm^2 on-chip while consuming 12 mW on average. The synchronous $\Delta\Sigma$ input layer neurons occupy 0.4 mm^2, consuming 10.5 mW average power in this network. On the other hand, the second complete network can be built with 784 neuromorphic $\Delta\Sigma$ inputs and the same number of neurons and synapses. This network occupies almost 10 mm^2 on-chip while consuming 11 mW on average. The neuromorphic $\Delta\Sigma$ input layer neurons occupies 8.6 mm^2 consuming 10 mW average power in this system.

Chapter 5

Conclusion and Discussion

5.1 Conclusion

Spiking Neural Networks is a developing area in Artificial Intelligence applications that focuses on replicating the behavior of the brain's neurons. This algorithm's primary benefit is saving area and power in computation and memory compared to typical Artificial Neural Networks. In the SNN, the communication of neurons is based on sending and receiving spikes, while the signals in the outside world produced by sensors and received by the SNN are not in a spiking form. Thus, the input layer neuron must receive the sensory information and convert it into spiking signals. Therefore, the neurons in the input layer differ from the other layers.

The focus of this work was to design two different input layer neurons which convert the analog input data into spikes and can be implemented on the same chip as other layers. Both neurons used a $\Delta\Sigma$ encoder circuit at their input interface to encode the analog input into a digital bit sequence. Their difference was the spike-generating method based on the $\Delta\Sigma$ modulated digital signal. Synchronous $\Delta\Sigma$ input neuron monitored the $\Delta\Sigma$ digital signal with the same clock frequency and produced a spike in case of receiving '1'. On the other hand, the neuromorphic $\Delta\Sigma$ neuron created a current from the received bits by a synaptic circuit and injected it into the membrane of an IF neuron to create spikes.

These designed $\Delta\Sigma$ Analog to Spike (A-S) converters introduce a method to encode sensory information into spikes that can be used as the input layer neurons in any SNN system. In Chapter 2,

the background needed to understand neuron and synapse systems and circuits was reviewed in addition to the $\Delta\Sigma$ systems and different input encoders. Then, in Chapter 3, the circuits of the synchronous and neuromorphic $\Delta\Sigma$ encoders were introduced and optimized for a clock frequency of 50 kHz after proving that both methods work at the system level. Afterward, in Chapter 4, both transistor-level circuits were simulated with DC and AC inputs as well as power and area with a clock frequency of 50 kHz. Finally, in the last section of Chapter 4, the clock frequency was changed from 100 Hz to 40 MHz, and each criterion was measured again for different clock signals. The result summary is provided in Table 4.5.

Compared to the neuromorphic $\Delta\Sigma$ input neuron, the synchronous one provides a better encoding accuracy and higher input bandwidth while occupying a smaller area. On the other hand, the neuromorphic one utilizes an actual neuron and synaptic circuit and has a better synergy with other layer neurons while consuming a lower average power. The synchronous $\Delta\Sigma$ A-S converter's anomaly in high clock frequency is because a low-power, high-speed spike generator was specially designed for this neuron when operating in higher than 1 MHz clock frequency. As a result of the replacement, this neuron's power consumption drops at high-speed clock signals. Furthermore, as seen in the table, the difference in power and area of two input encoders drops in high-speed clock applications.

5.2 Discussion

This work presents two methods to encode analog sensory information into spikes. In the current state of SNN systems, there are no widespread designs for the input layer's neurons that can convert the input data into spikes with an acceptable resolution. The existing methods either use a specific type of encoder that can work only in that structure or try to produce spikes outside the chip and then feed the spikes into the SNN. The contribution of this work is that it can be used in any SNN system operating in a wide range of spiking frequencies, and the whole network can be implemented on the same chip. Any SNN application that needs high-resolution data encoding, such as colored image recognition or control systems, can benefit from this work's design.

Furthermore, the following questions can be investigated as research in the future based on the

results of this work:

1. How designing a new synchronous spike generator can improve power dissipation and work at low and high clock speeds, maintaining the same encoding accuracy?

2. Is using other ways to design the OTA and comparator of the $\Delta\Sigma$ circuit save power while keeping the encoder's resolution? For example, the comparator-based SC gain stage structure [41] can be investigated in this topic.

3. Can oversampling technique apply to other parts of a neural network and improve the system? The oversampling technique can be analyzed in designing neuron models, and their benefits can be explored.

4. Can higher order feedback loops and increasing the integrator number in the $\Delta\Sigma$ architecture benefit the encoding resolution of the analog to spike converter? It is known that using a 2^{nd} order loop instead of 1^{st} order improves the SNR of a $\Delta\Sigma$ ADC by 12 dB [7]. The effects of this improvement can be explored in the input neuron design in SNN applications.

5. Does implementing the encoders with a differential input improve SNR? Since the positive and negative-processed output of the $\Delta\Sigma$ creates a differential output, a full-differential implementation of these input encoders is possible with a more complex circuit and consuming more power.

6. If the IF neuron in the neuromorphic encoder becomes more biologically plausible by adding leakage, adaptive spiking frequency, and varying threshold, does the input encoder lose resolution? If so, in what applications is it beneficial to have a biologically plausible neuron?

7. How does the neuromorphic input encoder perform after replacing the IF model with other models of neurons, like Izhikevich [30], or exponential IF [31]? In SNN systems where the neuron model is not a linear IF, the neuromorphic input encoder's neuron can match other layers' neuron models.

8. Build a complete spiking neural network with the designed encoder and use the encoder to convert the training and test data into spikes. Moreover, test the same SNN system with

an ideal input encoder. How much does using the introduced A-S converters degrade the accuracy of the SNN compared to the ideal version? An ideal input encoder converts the data into spikes outside the SNN chip, and the hidden layer neurons receive this spiking signal. A programmed FPGA can map the sensory data into digital spikes.

9. Multiple SNN systems use a non-linear encoding method. How much does using the introduced input neurons, instead of the non-linear one, improve the accuracy of the SNN's output in its testbenches?

Bibliography

[1] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of Physiology*, vol. 117, no. 4, pp. 500–544, aug 1952. [Online]. Available: https://doi.org/10.1113%2Fjphysiol.1952.sp004764

[2] Z. F. Mainen, J. Joerges, J. R. Huguenard, and T. J. Sejnowski, "A model of spike initiation in neocortical pyramidal neurons," *Neuron*, vol. 15, no. 6, pp. 1427–1439, Dec. 1995. [Online]. Available: https://doi.org/10.1016/0896-6273(95)90020-9

[3] A. van Schaik, "Building blocks for electronic spiking neural networks," *Neural Networks*, vol. 14, no. 6-7, pp. 617–628, Jul. 2001. [Online]. Available: https://doi.org/10.1016/s0893-6080(01)00067-3

[4] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. van Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, and K. Boahen, "Neuromorphic silicon neuron circuits," *Frontiers in Neuroscience*, vol. 5, 2011. [Online]. Available: https://doi.org/10.3389%2Ffnins.2011.00073

[5] A. Hazan and E. E. Tsur, "Neuromorphic analog implementation of neural engineering framework-inspired spiking neuron for high-dimensional representation," *Frontiers in Neuroscience*, vol. 15, Feb. 2021. [Online]. Available: https://doi.org/10.3389/fnins.2021.627221

[6] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri, "Neuromorphic electronic circuits for building autonomous cognitive systems," *Proceedings of the IEEE*, vol. 102, no. 9, pp. 1367–1388, Sep. 2014. [Online]. Available: https://doi.org/10.1109/jproc.2014.2313954

[7] R. J. Baker, *CMOS: Mixed-Signal Circuit Design*, 2nd ed. Wiley-IEEE, 2008. [Online]. Available: http://CMOSedu.com

[8] S. Rabii and B. Wooley, "A 1.8-v digital-audio sigma-delta modulator in 0.8-m CMOS," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 6, pp. 783–796, Jun. 1997. [Online]. Available: https://doi.org/10.1109/4.585245

[9] B. Fotouhi, "All-MOS voltage-to-current converter," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 1, pp. 147–151, 2001. [Online]. Available: https://doi.org/10.1109/4.896241

[10] M. Sharifshazileh, K. Burelo, J. Sarnthein, and G. Indiveri, "An electronic neuromorphic system for real-time detection of high frequency oscillations (HFO) in intracranial EEG," *Nature Communications*, vol. 12, no. 1, May 2021. [Online]. Available: https://doi.org/10.1038/s41467-021-23342-2

[11] C. Mead, "Neuromorphic electronic systems," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1629–1636, 1990. [Online]. Available: https://doi.org/10.1109%2F5.58356

[12] C. Eliasmith, T. C. Stewart, X. Choo, T. Bekolay, T. DeWolf, Y. Tang, and D. Rasmussen, "A large-scale model of the functioning brain," *Science*, vol. 338, no. 6111, pp. 1202–1205, nov 2012. [Online]. Available: https://doi.org/10.1126%2Fscience.1225266

[13] S. Lek and Y. Park, "Artificial neural networks," in *Encyclopedia of Ecology*. Elsevier, 2008, pp. 237–245. [Online]. Available: https://doi.org/10.1016%2Fb978-008045405-4.00173-7

[14] B. Han, A. Sengupta, and K. Roy, "On the energy benefits of spiking deep neural networks: A case study," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jul. 2016. [Online]. Available: https://doi.org/10.1109/ijcnn.2016.7727303

[15] D. Liu, H. Yu, and Y. Chai, "Low-power computing with neuromorphic engineering," *Advanced Intelligent Systems*, vol. 3, no. 2, p. 2000150, Dec. 2020. [Online]. Available: https://doi.org/10.1002/aisy.202000150

[16] G. Indiveri and S.-C. Liu, "Memory and information processing in neuromorphic systems," *Proceedings of the IEEE*, vol. 103, no. 8, pp. 1379–1397, Aug. 2015. [Online]. Available: https://doi.org/10.1109/jproc.2015.2444094

[17] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014. [Online]. Available: https://doi.org/10.1126/science.1254642

[18] D. Zhang and Z. Tan, "A review of optical neural networks," *Applied Sciences*, vol. 12, no. 11, p. 5338, May 2022. [Online]. Available: https://doi.org/10.3390/app12115338

[19] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. P. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, and G. W. Burr, "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature*, vol. 558, no. 7708, pp. 60–67, Jun. 2018. [Online]. Available: https://doi.org/10.1038/s41586-018-0180-5

[20] K. Yamazaki, V.-K. Vo-Ho, D. Bulsara, and N. Le, "Spiking neural networks and their applications: A review," *Brain Sciences*, vol. 12, no. 7, p. 863, Jun. 2022. [Online]. Available: https://doi.org/10.3390/brainsci12070863

[21] S. S. Radhakrishnan, A. Sebastian, A. Oberoi, S. Das, and S. Das, "A biomimetic neural encoder for spiking neural network," *Nature Communications*, vol. 12, no. 1, Apr. 2021. [Online]. Available: https://doi.org/10.1038/s41467-021-22332-8

[22] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers in Computational Neuroscience*, vol. 9, Aug. 2015. [Online]. Available: https://doi.org/10.3389/fncom.2015.00099

[23] Y. Wang, D. Wu, Y. Wang, X. Hu, Z. Ma, J. Feng, and Y. Xie, "A low-cost hardware-friendly spiking neural network based on binary MRAM synapses, accelerated using in-memory computing," *Electronics*, vol. 10, no. 19, p. 2441, Oct. 2021. [Online]. Available: https://doi.org/10.3390/electronics10192441

[24] W. Guo, M. E. Fouda, A. M. Eltawil, and K. N. Salama, "Neural coding in spiking neural networks: A comparative study for robust neuromorphic systems," *Frontiers in Neuroscience*, vol. 15, Mar. 2021. [Online]. Available: https://doi.org/10.3389/fnins.2021.638474

[25] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going deeper in spiking neural networks: VGG and residual architectures," *Frontiers in Neuroscience*, vol. 13, Mar. 2019. [Online]. Available: https://doi.org/10.3389/fnins.2019.00095

[26] Y. Zeng, T. C. Stewart, Z. I. Ferdous, Y. Berdichevsky, and X. Guo, "Temporal learning with biologically fitted SNN models," in *International Conference on Neuromorphic Systems 2021*. ACM, Jul. 2021. [Online]. Available: https://doi.org/10.1145/3477145.3477153

[27] N. Sengupta and N. Kasabov, "Spike-time encoding as a data compression technique for pattern recognition of temporal data," *Information Sciences*, vol. 406-407, pp. 133–145, sep 2017. [Online]. Available: https://doi.org/10.1016%2Fj.ins.2017.04.017

[28] D. V. Christensen, R. Dittmann, B. Linares-Barranco, A. Sebastian, M. L. Gallo, A. Redaelli, S. Slesazeck, T. Mikolajick, S. Spiga, S. Menzel, I. Valov, G. Milano, C. Ricciardi, S.-J. Liang, F. Miao, M. Lanza, T. J. Quill, S. T. Keene, A. Salleo, J. Grollier, D. Marković, A. Mizrahi, P. Yao, J. J. Yang, G. Indiveri, J. P. Strachan, S. Datta, E. Vianello, A. Valentian, J. Feldmann, X. Li, W. H. P. Pernice, H. Bhaskaran, S. Furber, E. Neftci, F. Scherr, W. Maass, S. Ramaswamy, J. Tapson, P. Panda, Y. Kim, G. Tanaka, S. Thorpe, C. Bartolozzi, T. A. Cleland, C. Posch, S. Liu, G. Panuccio, M. Mahmud, A. N. Mazumder, M. Hosseini, T. Mohsenin, E. Donati, S. Tolu, R. Galeazzi, M. E. Christensen, S. Holm, D. Ielmini,

and N. Pryds, “2022 roadmap on neuromorphic computing and engineering,” *Neuromorphic Computing and Engineering*, vol. 2, no. 2, p. 022501, May 2022. [Online]. Available: https://doi.org/10.1088/2634-4386/ac4a83

[29] G. Indiveri, F. Stefanini, and E. Chicca, “Spike-based learning with a generalized integrate and fire silicon neuron,” in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. IEEE, May 2010. [Online]. Available: https://doi.org/10.1109/iscas.2010.5536980

[30] E. Izhikevich, “Simple model of spiking neurons,” *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1569–1572, Nov. 2003. [Online]. Available: https://doi.org/10.1109/tnn.2003.820440

[31] S. A. Aamir, P. Muller, G. Kiene, L. Kriener, Y. Stradmann, A. Grubl, J. Schemmel, and K. Meier, “A mixed-signal structured AdEx neuron for accelerated neuromorphic cores,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 5, pp. 1027–1037, Oct. 2018. [Online]. Available: https://doi.org/10.1109/tbcas.2018.2848203

[32] S. A. Nasrollahi, A. Syutkin, and G. Cowan, “Input-layer neuron implementation using delta-sigma modulators,” in *2022 20th IEEE Interregional NEWCAS Conference (NEWCAS)*. IEEE, Jun. 2022. [Online]. Available: https://doi.org/10.1109/newcas52662.2022.9842225

[33] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal dynamics*. Cambridge, England: Cambridge University Press, Jul. 2014.

[34] D. E, N. Pradhan, S. R, and A. Sreedevi, “Analysis of the dynamic behaviour of a single hodgkin-huxley neuron model,” in *2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*. IEEE, Dec. 2015. [Online]. Available: https://doi.org/10.1109/erect.2015.7499056

[35] L. Abbott, “Lapicque’s introduction of the integrate-and-fire model neuron (1907),” *Brain Research Bulletin*, vol. 50, no. 5-6, pp. 303–304, Nov. 1999. [Online]. Available: https://doi.org/10.1016/s0361-9230(99)00161-6

[36] S. ichi Amari, "Backpropagation and stochastic gradient descent method," *Neurocomputing*, vol. 5, no. 4-5, pp. 185–196, Jun. 1993. [Online]. Available: https://doi.org/10.1016/0925-2312(93)90006-o

[37] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. [Online]. Available: https://doi.org/10.1109/5.726791

[38] D. Auge, J. Hille, E. Mueller, and A. Knoll, "A survey of encoding techniques for signal processing in spiking neural networks," *Neural Processing Letters*, vol. 53, no. 6, pp. 4693–4710, Jul. 2021. [Online]. Available: https://doi.org/10.1007/s11063-021-10562-2

[39] D. Mahajan, V. Kakkar, and A. K. Singh, "Analysis of delta sigma modulator," in *2011 International Conference on Computational Intelligence and Communication Networks*. IEEE, oct 2011. [Online]. Available: https://doi.org/10.1109%2Fcicn.2011.36

[40] M. M. Krishna and M. Vadivel, "Design and analysis of first order sigma-delta modulator based on switched capacitor integrator (130nm)," in *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*. IEEE, Sep. 2021. [Online]. Available: https://doi.org/10.1109/icses52305.2021.9633907

[41] M.-C. Huang and S. luan Liu, "A fully differential comparator-based switched-capacitor delta-sigma modulator," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 56, no. 5, pp. 369–373, May 2009. [Online]. Available: https://doi.org/10.1109/tcsii.2009.2019166