Smart IoT - Implementation of Neuromorphic Circuits

Artur Piana, Siqi Wang, Joao Frischenbruder Sulzbach, Paolo Zaniboni, Raphael El-Haddad, A. Benlarbi-Delai, Geoffroy Klisnick, Pietro Maris Ferreira

Smart IoT - Implementation of Neuromorphic Circuits

Artur Piana, Siqi Wang, João F. Sulzbach, Paolo Zaniboni, Raphael El-Haddad, Aziz Benlarbi-Delai, Geoffroy Klisnick, Pietro M. Ferreira

Sorbonne Université, CNRS, Lab. de Génie Électrique et Électronique de Paris, 75252, Paris, France
Université Paris-Saclay,CentraleSupélec,CNRS,Lab. de Génie Électrique et Électronique de Paris,91192,Gif-sur-Yvette,France

***Abstract*—This paper presents a simple analog neuromorphic system in microelectronics suitable for an audio source localization problem. The receptor can find the relative position of an audio source by the information of angles and distances according to acquire acoustic signals. In this paper, we focus on the angle detection, but some information of the distance is also presented. This paper also presents the development of some alternatives to the most important circuit blocks in neuromorphic systems: the neurons and the synapses. The results are validated in two different levels: the system level and the transistor level.**

***Index Terms*—Internet of Things (IoT), neuromorphic circuits, spiking neural networks, ultra-low power**

I. Introduction

Nowadays, the Internet of Things (IoT) becomes gradually more extensive. With the development of artificial intelligence in both software and hardware [1], these connected objects are becoming more complex and more intelligent, which paves a way for numerous applications in the future. The information processing based on digital signal processing techniques is usually implemented in an architecture (e.g. Von Neumann, Harvard) that has shown good performance on computing power, memory capacity, and accessibility. The technology of integrated circuits based on CMOS technology [2] has been rapidly developed as described by Moore's law, particularly in terms of miniaturization and heat dissipation. However, with the end of Moore's law [3], the intelligent and powerful IoT objects come at the price of high-power consumption, as simple software and hardware optimization do not follow the advance in the same pace. Above that, the power consumption of an integrated circuit is always a challenge. One can clearly see the conflict between performance and power efficiency, which suggests the need of power optimization of the current electronic circuits. One promising solution for that is the "More than Moore" paradigm, relying on the development of alternative electronic systems beyond the traditional and well-known Moore's Law. For this reason, this work highlights neuromorphic systems as a design solution under this paradigm [4]. Neuromorphic systems are inspired by brain's biological nervous systems, which can be energy efficient and promising for information processing [5]. Most popular neural networks nowadays (software and digital) still have room for power optimization and further improvements. The spiking neural network (SNN) becomes a very promising solution since analog systems can be power efficient [6], [7]. In the literature [8], [9], the analog implementation of neuromorphic systems exhibit ultra-low power (ULP) consumption [10], [11], as well as excellent miniaturization perspectives. In this paper, we firstly design, both in schematics and layout, low-level circuit blocks (neurons and synapses) for SNN with ultra-low power. Secondly, we present a conception of high-level circuit models, *i.e.* neurons and synapses, which describes the behavior of analog neural networks.

II. Neuron and Synapse in SNN

A. Artificial Neurons

Neurons of the brain cortex are electrically excitable cells that respond to excitation current upon its membrane with voltage spikes that are conducted from its soma to its axons and then to the next neurons. These spiking voltage can be of many different forms [12]. Artificial spiking neurons are inspired by the behavior of the biological neurons. They can be characterized as an electronic circuit with:

- **Excitation current**: the spiking neuron is excited by a current, which is usually in order of the picoamperes (pA). To make the neuron fire spikes with a certain frequency, this current is usually constant.
- **Membrane voltage**: the neuron transmits its voltage in form of spike trains, which is usually in order of the millivolt (mV). The voltage is periodic with a frequency proportional to the input.

The neuron can be seen as a frequency modulated signal driven by an input current. Moreover, it can be considered as a converter, which converts physical signals to information coded in spikes.

B. Artificial Synapses

Artificial synapses are the interconnections between neurons, which transmit electrical signals from one neuron to another. They are very numerous in number because a neuron can be connected to several others at the same time. In this case, the artificial synapse can be characterized with following signals:

- **Input**: the input of a synapse is usually a spiking voltage of a pre-neuron.

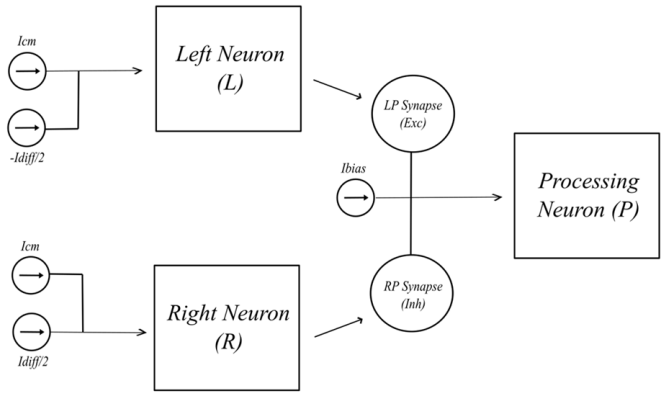

Fig. 1. Schematics of the proposed system.

- **Output**: the output of a synapse is a step current in steady state and to be used as to excite or inhibit the post-neuron.

To put it simply, synapses can be of two distinct types:

- **Excitation Synapse**: the synapse behaves as a current source, injecting a positive current to the post-neuron, leading to an increase in spiking frequency.
- **Inhibition Synapse**: the synapse behaves as a current sink, reducing the excitation current of the post-neuron, leading to a decrease in spiking frequency.

III. Proposed framework

We propose a simple, two-layer neural network with three different neurons and two synapses as illustrated in Fig. 1. It is composed of following blocks:

1) **Input signals**: Two separate and independent current I_{cm} and I_{diff} are taken as inputs, which for example could be the output currents of a sound sensor.
2) **Input neurons**: Two separate and independent neurons are named as Left Neuron (L) and Right Neuron (R). Each of them receives independent current inputs as described above. The firing frequency of each neuron is an increasing function of the input signals.
3) **Synapses**: Two independent synapses of excitatory and inhibitory receive the voltage spikes and providing the excitation (resp. inhibition) current to post-neurons.
4) **Bias current source**: The source of a fixed current is used to provide a bias current. This results in a fixed operating point for the post-neuron.
5) **Processing neuron**: A single neuron receives the sum of the currents of the bias point and of the synapses. As the bias point being fixed, its spiking frequency depends solely on the inputs.

As the Left Neuron has an increase on input current, its spiking frequency also increases. The LP synapse will then have a larger output current, which will raise up the Processing Neuron's spiking frequency. The same applies for the Right Neuron, except that the RP synapse is an inhibitory synapse which provides a low output current and decreases the spiking frequency of the post-neuron.

IV. Circuits design and layout simulation

The project is designed in two different levels:

1) **Circuit Level Design**: design of neurons and synapses at the transistor level using Cadence.
2) **System Level Design**: design of neurons and synapses at the system level using high-level languages, e.g. Matlab and VerilogA.

A. Circuit-Level Design

The technology used in this paper is the BiCMOS 55 nm by STMicroelectronics. In this paper, the designed circuits are tested firstly in the schematic domain (electrical simulation) and afterward in the layout domain (post-layout simulation with all parasites).

1) Synapse: The designed synapses have a voltage input and a current output which can be both positive and negative. In this paper, we name it as "complementary synapses". The principle of the synapses is composed of three stages:

1) A Low Pass Filter made up with a diode and a capacitor. This filter is needed to extract the mean value from the output voltage of the previous neuron.
2) A transconductance, which is provided by an NMOS transistor. This is needed for getting a current output from the voltage input.
3) A mirroring stage to provide both positive (for excitation) and negative (for inhibition) current outputs.

The first-order filter was chosen according to the trade-off performance to and complexity. The parasitic effects (especially capacitance) could then also be less harmful to the circuit's performance compared with higher order filters. A classical filter is usually composed of a diode-connected transistor (acting as a diode) and a capacitance, creating the simple first-order filter. However, in this paper, we replace the transistor with a pure diode since the transistors in the subthreshold domain were hardly controllable in terms of conduction. In addition, the pure diode provides a much higher resistance, which increases the resistive effect in the RC circuit and therefore introduces a higher delay, comparable to the neuromorphic computing window of 1 ms. Two different types of schematics were analyzed in this paper as depicted in Fig. 2 and 3. They will be called respectively 5T synapse and 3T synapse in the following part of the paper. The main difference between them is in the mirroring stage: the one in 2 uses 5 transistors in the mirroring stage, and the one in 3 uses only 3. The latter will present, of course, lower power consumption and lower surface area in the layout stage, although with a possible reduction in performance.

The performances of these synapses are validated on the test bench as depicted in Fig. 4. The test bench is composed of an input neuron, a complementary synapse and two output neurons. The input neuron is excited by different currents and we analyzed the currents generated by the synapse. The behaviors of both synapses are given in Fig. 5. The difference between these two synapses is that the value of the inhibitory current of the 3T topology is lower. We can observe a

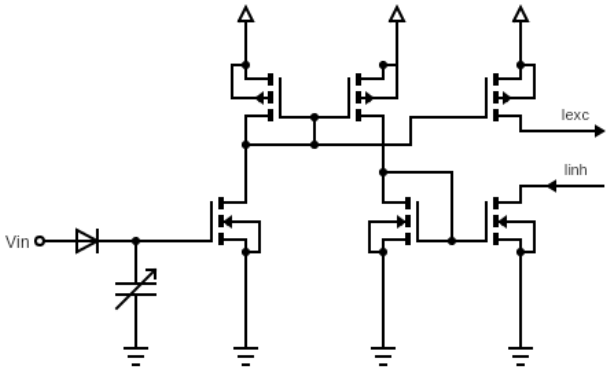

Fig. 2. Schematics of the 5T synapse.

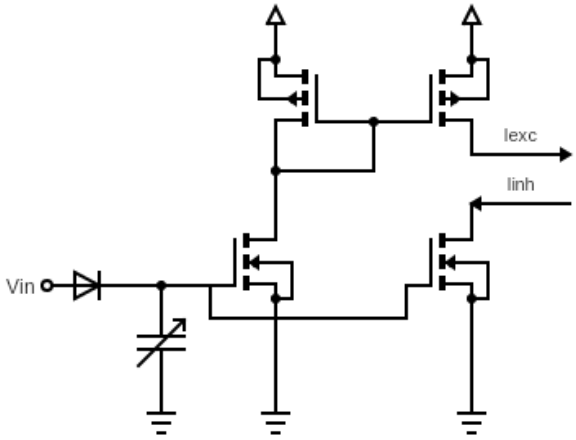

Fig. 3. Schematics of the 3T synapse.

mismatch between the excitatory and the inhibitory currents. It is due to channel-length modulation effect, since the drain to source voltages of all the transistors are different and even oscillating for the output stage. This is a strong non-ideality that will strongly affect the behavior of our neural network. The result of layout simulation for the 5T topology is shown in Fig. 6. The non-ideality brings a significant drop in both output currents. This drop is due to charge-sharing phenomena between the diode equivalent capacitance and the variable capacitance, as well as all the parasitic capacitance that is added between this node and the ground. This charge-sharing brings a drop in the voltage at the gate of the first transistor, which results in a small output current.

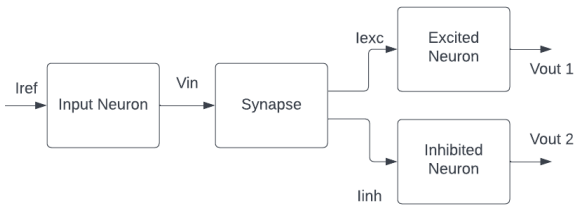

Fig. 4. Synapse Testbench.

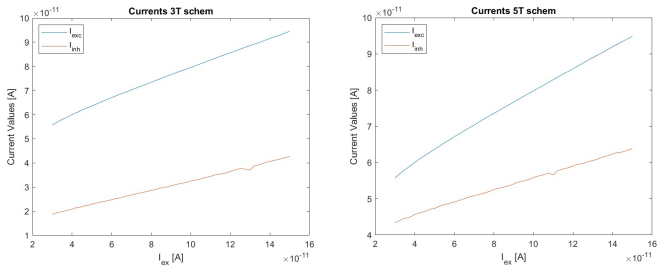

Fig. 5. Output Currets for the 3T and 5T Synapse.

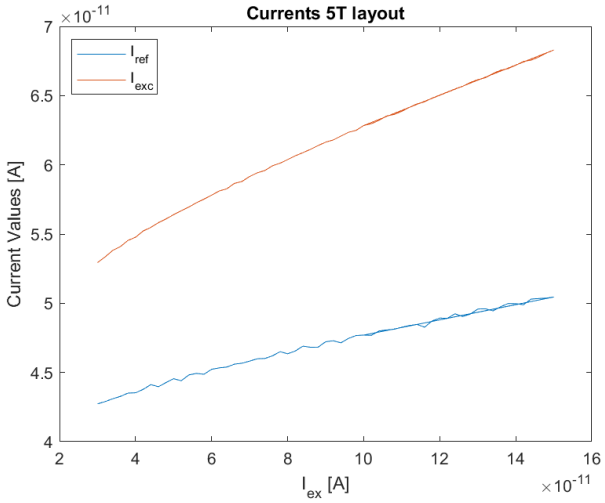

Fig. 6. Output Currents for the 5T Synapse (Layout).

2) Neural Network: We construct neural networks with neurons designed in [9] and the complementary synapses designed in this paper as illustrated in Fig. 1. This neural network has two inputs current: common-mode current which affects all neurons equally, and differential-mode current which is the deviation from the common-mode input current and with absolute equal value but opposing signs for each input neuron.

The simulation result is given in Fig. 7. The output neuron spiking frequency is illustrated in Fig. 8. The neurons have an opposing but symmetric response to the differential input currents, as one increases but one decreases in a similar speed.

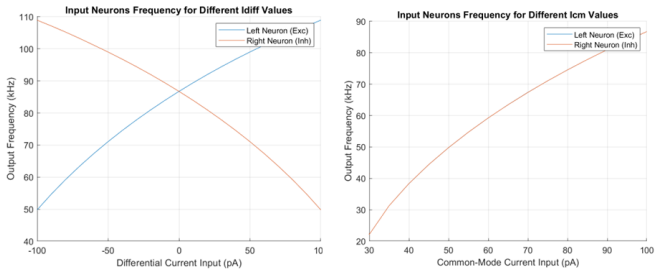

Fig. 7. Neural Network Tests for Low-Level: Input Neurons Spiking Frequency in function of Differential (left) and Common-Mode (right) Input Currents.

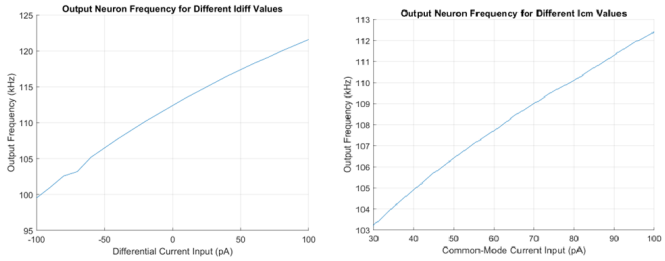

Fig. 8. Neural Network Tests for Low-Level: Output Neurons Spiking Frequency in function of Differential (left) and Common-Mode (right) Input Currents.

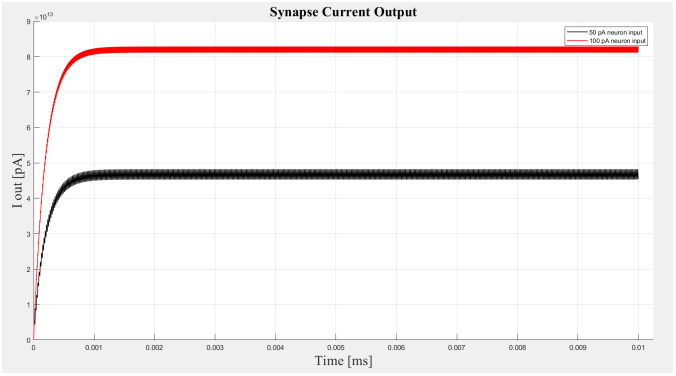

Fig. 9. Matlab Synapse current modulation.

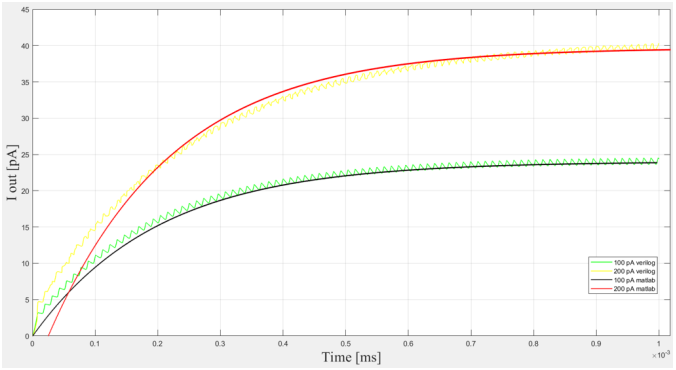

Fig. 10. Comparison between synapses in MATLAB and VerilogA.

Moreover, the spiking frequency of both neurons intersect at zero differential input currents as expected. For the common-mode current variation with zero differential current, the result is the same for both neurons. The output neuron frequency varies almost linearly with the differential input, which is a good first point for the neural network. Furthermore, the inverse sensitivity is of about 9.1 pA/spike with a resolution of approximately 4.4 bits. The latter one can be even increased if the bias point is changed and moved up to higher currents, since the full scale is mostly limited by the minimum input current for a spike in the input neurons. Therefore, the differential sensing works quite well and could be used to detect the angle of a source.

B. System-Level (Matlab and VerilogA)

1) Neuron: For system level modeling of neural networks, we propose a neuron model based on the Fast Spiking (FS) neuron model in [13] by Henider. We modify the calculation of the spike waveform by defining a T_{on} period where the function is greater than zero, and a T_{off} period where the function is equal to zero:

$$T_{on} + T_{off} = T = \frac{1}{f_s} \tag{1}$$

where fs is the spike frequency. Interpolated data from Cadence is used for model construction. The range of membrane potential is between 0 and 100 mV which is observed experimentally in Cadence. In VerilogA, we make a periodic function with a "sub-threshold" behavior. The neuron does not spike when I_{ex} is below 30 pA. The spike of membrane voltage is between 0 mV and 80 mV.

2) Synapse: A synapse model was developed according to the hardware circuit. A low pass filter is used in synapse modeling to fetch the DC component and to smooth the curve by adding a delay on the system. We model each component of the circuit individually and then connect all of them. The Matlab model assumes all hardware components are in ideal condition. In VerilogA, a low-pass filter in the first order has been used. Concerning the parameters of this filter, the real parameter of the physical model has been used for the time constant RC with $R = 6.48$ GΩ which correspond to the diode resistor and $C = 1.85$ fF. The gain is set to match with the simulation curves. Then we choose the inverse gain to have an excitatory synapse instead of a inhibition one. We connect the output of a neuron to the input of the synapse. The output current of the synapse is modulated by the frequency of the spiking neuron as illustrated in Fig. 9. The graph containing the comparison between MATLAB and VerilogA models for inhibition synapse is displayed in Fig. 10. A current value has been chosen (100 pA or 200 pA) as the input of the neuron connected to the synapse. One can notice that the inhibition properties are matched. Then it can be seen that MATLAB and VerilogA models matched well.

3) Neural Network: The model of the whole circuit of neural networks in Fig. 1 assembles the models of neuron and synapse. The mathematical model is aimed at working in the same way as the physical model but with an advantage of computing way faster and be easily editable. We will also make tests varying the differential current and common-mode current of the input in order to observe the sensitivity of the output frequency in relation to their variation. We firstly build the model in Fig. 11 in VerilogA and design another similar branch as in Fig. 1. The result is illustrated in Fig. 12. The upper subfigure is the output of the first neuron, the figure below is the output of the final neuron. The frequency is lower and the signal even starts to be constant because of a low current at the input. In Fig. 13, one can notice that both the synapses and the neurons are working as expected (above and sub threshold mode). In Matlab we connect the necessary components and measure the outputs of the left and right

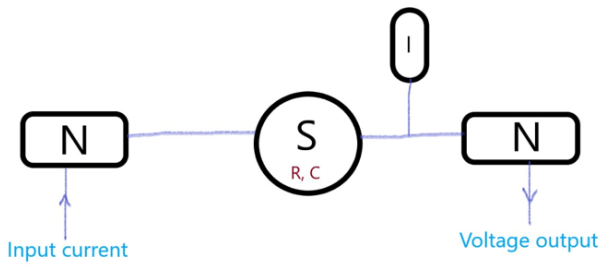

Fig. 11. Test circuit VerilogA.

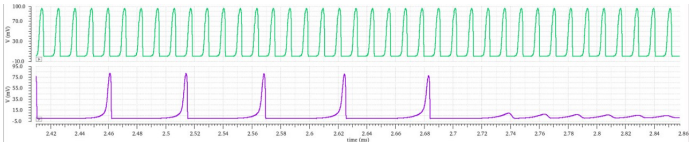

Fig. 12. VerilogA results on one branch.

neurons and synapses. An example simulation is displayed in Fig. 15. In this example, the common-mode current is 75 pA and the differential current is 50 pA. The output spikes of the processing neuron are displayed in Fig. 14. The resulting frequency is 63 kHz.

4) Results analysis: We first analyze the results of VerilogA modeling. In case of the common mode where we increase I_{cm} in figure 1 and I_{diff} is always equal to zero, we observed that the output frequency of the post-neuron is constant as expected. It is close to the ideal one, differently from what was seen for the low-level design. In the case of the differential mode where we keep I_{cm} constant and increasing I_{diff}, we observed that $I_{cm} = 50pA$ and I_{diff} goes from $-100pA$ to $100pA$. The frequency is increasing, with an inverse sensitivity of 13 pA/spike. This sensitivity is relatively large and the output varies linearly with the current, confirming previous expectations. We then analyze the results of Matlab modeling. We perform the same tests varying the common-mode and differential-mode currents and measure each neuron sensitivity to the variations. For the pre-neurons, varying the common-mode and differential-mode currents is simply a current variation. Therefore, the common-mode sensitivity in Fig. 17 is equal for both left and right neurons. The differential-mode sensitivity in Fig. 16 is inverted as the left neuron is fed by a negative differential current and the right neuron is fed by a positive difference current. The absolute value of the

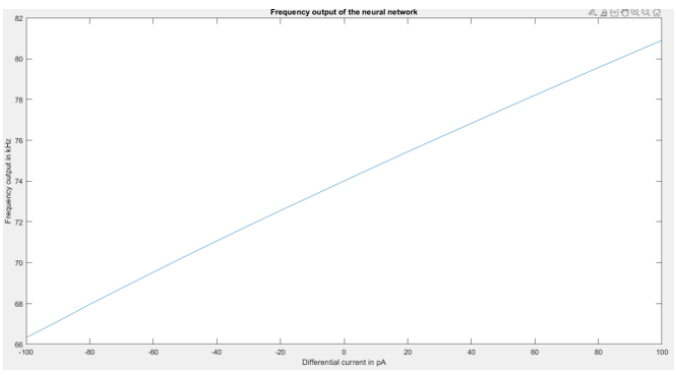

Fig. 13. Frequency output of the VerilogA circuit.

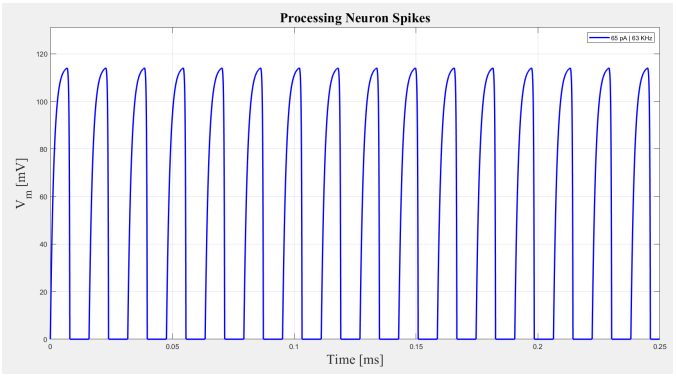

Fig. 14. Matlab Neural Network Processing Neuron Spikes.

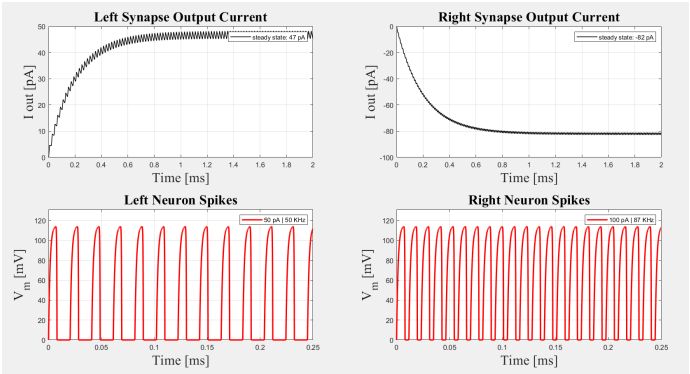

Fig. 15. Matlab Neural Network Left and Right Neurons and Synapses.

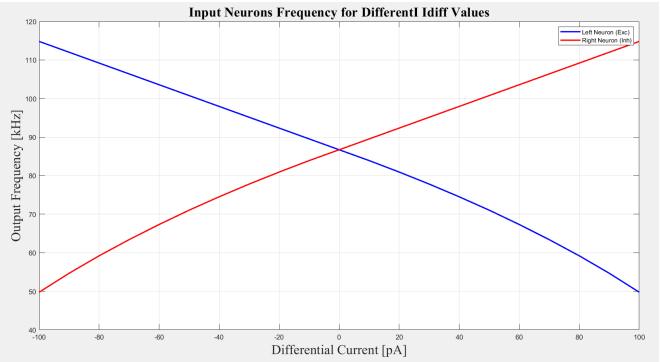

Fig. 16. Neural Network Input Neurons Frequency for Different I_{diff} Values.

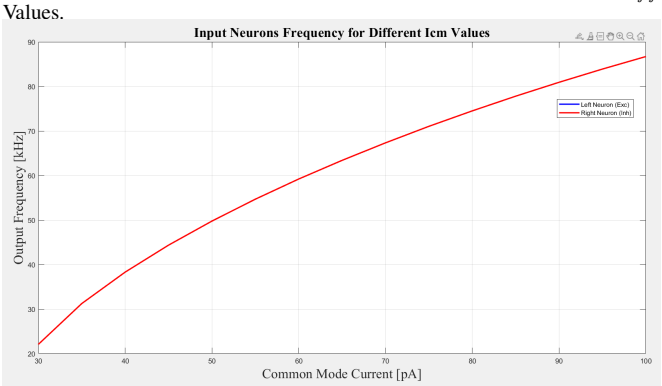

Fig. 17. Neural Network Input Neurons Frequency for Different I_{cm} Values.

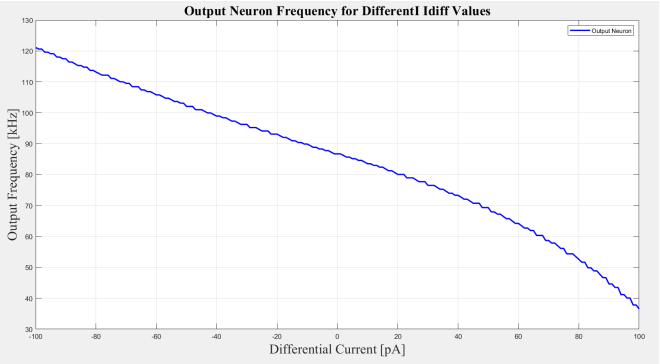

Fig. 18. Neural Network Processing Neuron Frequency for Different Idiff Values

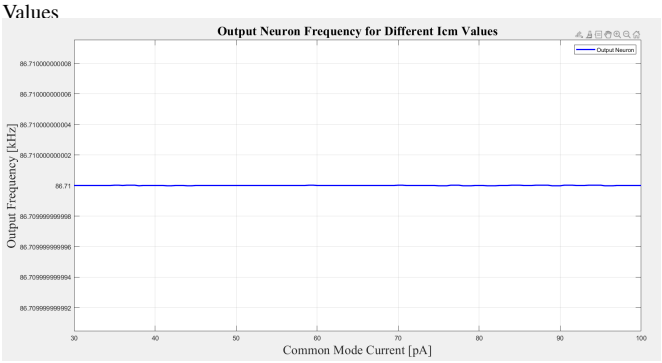

Fig. 19. Neural Network Processing Neuron Frequency for Different Icm Values

sensitivity is always 0.8 kHz/pA (where the inverse sensitivity is 1.25 pA/ kHz) in the linear region.

The post-neuron current common-mode sensitivity in Fig. 19 is zero. We cannot change its frequency by adding equal input to both left and right neurons at the same time. For the differential-mode current, we measure a sensitivity of 0.33 kHz/pA (inverse sensitivity of 3 pA/kHz) in the linear region. This value is high compared to the Cadence circuit and Verilog models, which confirms that the post-neuron is indeed affected by a current difference in the pre-neurons, therefore being able to detect a difference in position of the signal resource (angle).

V. Conclusion

In summary, a proof of concept of the neuron network in the system-level is achieved. Regarding the system-level models, they are fully parametrized so that the results can be adapted according to changes in the circuit blocks. The differential and the common mode have very good sensing. Moreover, it provides insights into what should be corrected in the circuit-level design. The inverse sensitivities are different between models since every slight change in the neuron and synapse model will influence the whole circuit. It is suggested to tune the parameters in each model in order to have the same sensitivity which could allow the team to make a test on a desired platform. The simulation results confirm the concept of the spiking neural network based on low-power neuromorphic circuits.

References

[1] S. Moradi, N. Qiao, F. Stefanini, and G. Indiveri, "A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (dynaps)," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 1, pp. 106–122, 2018.

[2] H. Wang, "Review of cmos millimeter-wave radio frequency integrated circuits," in *2015 IEEE MTT-S International Microwave and RF Conference (IMaRC)*, 2015, pp. 239–242.

[3] T. N. Theis and H.-S. P. Wong, "The end of moore's law: A new beginning for information technology," *Computing in Science Engineering*, vol. 19, no. 2, pp. 41–50, 2017.

[4] B. Wen and K. Boahen, "A silicon cochlea with active coupling," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 3, no. 6, pp. 444–455, 2009.

[5] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Netw*, vol. 10, no. 9, pp. 1659–1671, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608097000117

[6] T. Matsubara and H. Torikai, "An asynchronous recurrent network of cellular automaton-based neurons and its reproduction of spiking neural network activities," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 836–852, 2016.

[7] Y. Cao, Y. Chen, and D. Khosla, "Spiking deep convolutional neural networks for energy-efficient object recognition," in *International Journal of Computer Vision*, 2015, pp. 54–66.

[8] P. M. Ferreira, J. Nebhen, G. Klisnick, and A. Benlarbi-Delai, "Neuromorphic analog spiking-modulator for audio signal processing," in *Analog Integrated Circuits and Signal Processing*, vol. 106, 2021, pp. 261–276.

[9] C. Loyez, K. Carpentier, I. Sourikopoulos, and F. Danneville, "Subthreshold neuromorphic devices for spiking neural networks applied to embedded a.i," in *2021 19th IEEE International New Circuits and Systems Conference (NEWCAS)*, 2021, pp. 1–4.

[10] M. Yang, S.-C. Liu, M. Seok, and C. Enz, "Ultra-low-power intelligent acoustic sensing using cochlea-inspired feature extraction and dnn classification," in *2019 IEEE 13th International Conference on ASIC (ASICON)*, 2019, pp. 1–4.

[11] A. Jimenez-Fernandez, E. Cerezuela-Escudero, L. Miro-Amarante, M. J. Dominguez-Morales, F. de Asis Gomez-Rodriguez, A. Linares-Barranco, and G. Jimenez-Moreno, "A binaural neuromorphic auditory sensor for fpga: A spike signal processing approach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 4, pp. 804–818, 2017.

[12] E. Izhikevich, "Simple model of spiking neurons," *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1569–1572, 2003.

[13] R. Henider, "Systèmes neuromorphiques pour les applications RF large bande," 2021.