AUTHOR VERSION

Revisiting the Ultra-Low Power Electronic Neuron Towards a Faithful Biomimetic Behavior

Théo Prats Rioufol, Zalfa Jouni, Thomas Soupizet, Pietro M. Ferreira
Université Paris-Saclay, CentraleSupélec, CNRS, Lab. de Génie Électrique et Électronique de Paris,91192,Gif-sur-Yvette,France.
Sorbonne Université, CNRS, Lab. de Génie Électrique et Électronique de Paris, 75252, Paris, France.
Email: theopratsrioufol@gmail.com, maris@ieee.org

***Abstract*—The spiking neural networks (SNN) have been considered as the third generation of neural networks, since it better mimics the biological behavior of the brain cortex which processes spike trains. This paper proposes to revisit electronic neuron implementations in a novel way able to deal with SNN limitations in deep learning. Post-layout simulation results prove that non-linear action function is able to shift from right to left considering a reference sizing to create excitation and inhibition synaptic current. Proposed model is extended to handle up to three neurons in a parallel association, i.e. SNN with a fan-in/fan out of 3 synaptic branches. Revisited electronic neuron achieves an area of $10.8 \times 9.7\ \mu m^2$ for a energy efficiency below 10 fJ/spike, while including synapse circuitry. Relative standard deviation of synapse current is below 1.8% and synapses weight mismatch leads to less than 4 pA current error.**

***Index Terms*—ultra-low-power, neuromorphic circuits, spiking neural networks**

I. Introduction

Neuromorphic computing is an established complementary to von Neumann systems exploring the Artificial Intelligence (AI) paradigm in electronics [1]. Digital implementations have often been proposed in FPGAs and GPUs [2] due to their advantages of reconfigurability, reusability and reduced implementation costs. Feed-forward neural networks (FNN) are the most common architecture for Deep Learning. FNNs present interneuron connections with a linear scaling controlled by the weight coefficients (synapses). The spiking neural networks (SNN) have been considered as the third generation of neural networks. Different from conventional FNNs which process digital data, the SNN better mimics the biological behavior of the brain cortex which processes spike trains [3].

A biological cortex neuron membrane potential (V_m) is excited by a current pulse (I_{ex}) of few hundreds of pico-Amps. Thus, it operates in an average firing rate (f_{spike}) of few Hz with an energy efficiency ($E_{eff} = P_{rms}/f_{spike}$) of 2.45 pJ/spike. Such neurons have an average membrane capacitance (C_m) of 245 pF and operate with an action potential (V_d) of 100 mV [4]. Analog electronic neurons (*eN*) cannot afford to work on such low f_{spike} and having such high C_m. *eN* have been implemented in the literature [5], [6] with f_{spike} of a few hundreds of kHz, using weak inverted transistors to save power and C_m dozens of fF to save area. However, recent neuromorphic hardware is often single neuron circuits [7], [5], [8], or small neural networks [9], [6]. Ou and Ferreira have proved that the spiking frequency variability may be compensated for $\pm 10\%$ of supply voltage variation while providing a tradeoff counterbalance for temperature and process variations [10].

In previous work [11], the authors have already questioned if *eN* are usable on analog SNN using deep neural network training tools and tried and narrow the gap between hardware and software AI in [11]. Assumptions were made considering analog electronic synapse (*eSyn*) implementation originally published in [12] and its ultra-low-power version from [9]. Published *eSyn* presents a neuron output f_{spike} (the pre-synaptic signal) conversion to an output (the post-synaptic signal) i_{syn} controlled by the trained synaptic weight. However, published *eSyn* implementation have presented a behavior which does not mimic the biological functionality of synapses. Besides, published models fails in implementing the neuron plasticity and training only found in memristor's synapses [13]. SNN scalability is a challenge without a bio-inspired synapses in silicon-based solution.

To establish a silicon-based SNN competitive to memristor's based, this paper proposes to revisit *eN* and *eSyn* implementations in a novel way able to deal with SNN scalability from a bio-inspired connection between neurons. Figure 1(a) illustrates the biological neuron model composed of soma, axon, and few dendrites. Figure 1(b) mimics biological behavior considering: an *eN* soma using a biomimetic Moris-Lecar model according to [5] implementation; an *eN* Axon having propagation delays from M_0 and C and a transconductance gain from M_2; *eN* dendrites exciting the V_m to spike accordingly. The *eSyn* circuit topology is considered from [11], but here it is redefined as the connection between *eN* axon and dendrites on the following layer. Revisited *eN* now considers the presence of one to three dendrites, which are connected to the different *eN* axons of the previous layer. The *eSyn* weights ($\omega_{h,j}$) is then the current mirror gain from M_3 presented in *eN* axon and $M_{h+1,j}$ dendrite transistors in the k^{th} *eN* in the $h+1^{th}$ layer.

II. Revisited Electronic Neuron

Let's consider two eNs connected together: j,k of layer h, $h+1$ in Fig. 1(b) The electronic Soma (eSoma) of $eN_j^{(h)}$ is driven by an input current $i_{in,h,j}$ and the electronic dendrite $eD_{j,k}^{(h+1)}$ making the interneuron connection have an output

979-8-3503-1834-0/23/$31.00 ©2023 IEEE

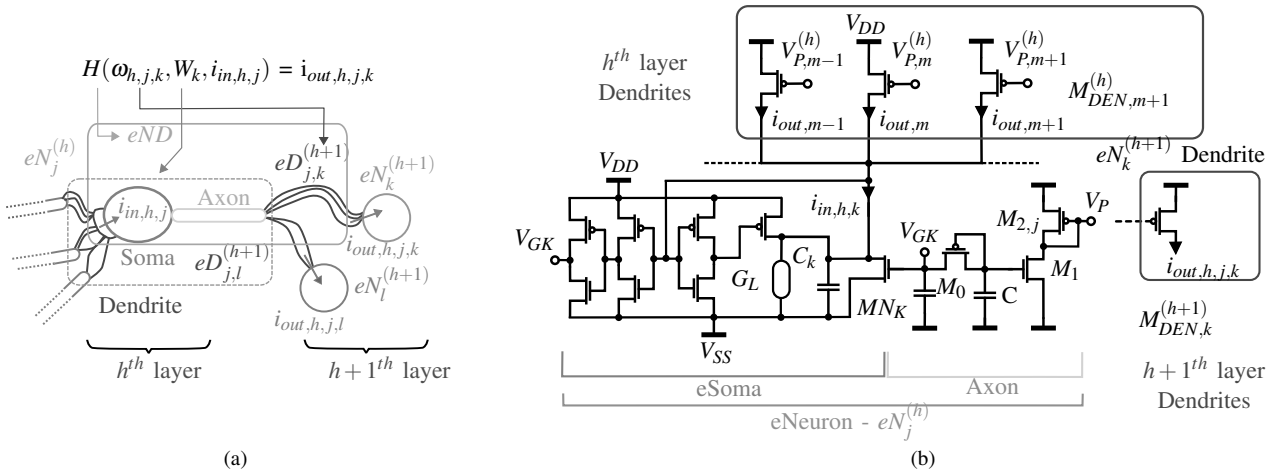


Fig. 1. (a) Biological inspired neural network topology. Each soma is connected to one axon and make connections through dendrites. (b) Schematic for an analog neuron and dendrites ($V_{DD} = 100\ mV$, $V_{SS} = -100\ mV$). For each new connection, a single PMOS transistor (dendrite) is added to form a new synapse. Transistor sizing is obtained from [11]. Modified sizing are: M_0 with 810/60 nm; $M_{2,j}$ with 2100/60, $C = 16.88$ fF, M_1 with 135/60 nm.

current $i_{out,h,j,k}$. The neuron behavior (H) of this set eSoma, eAxon, and eDendrite referred as *eND* is defined as:

$$i_{out,h,j,k} = H(\omega_{h,j,k}, W_K, i_{in,h,j}) \tag{1}$$

where W_K is the width of the MN_K transistor responsible for H bias shift; $\omega_{h,j,k}$ is the synaptic weight between the two *eNs*, controlled by the sizing ratio from $M_{2,j}^{(h)}$ and $M_{DEN,k}^{(h+1)}$. Besides, software-based NN have the corresponding behavior

$$i_{out,h,j,k} = AF_{software}(i_{in,h,j}) \times \omega_{h,j,k} \tag{2}$$

where $AF_{software}$ is the activation function of the neuron (i.e. the neuron behavior) and the synaptic weight is implemented through a product.

The revisited *eN* must achieve a non-linear $i_{in,h,j}$ to $i_{out,h,j,k}$ relationship. Moreover the *eN* should be able to be inhibited, i.e behave like if its input current was lower: i_{in} becoming $i_{in} - i_{tr}$, where i_{tr} is a positive threshold current. A exitation correspond to $i_{in} + i_{tr}$. This operation results in a shift of the plot $H(i_{in})$ to the rigth for inhibition or the left for exitation (see 4(a)).

For clarity, as ω is dependent on the width W of the dendrite, (1) is simplified to $i_{out} = H(W, W_K, i_{in})$ for an *eND*. Then, the contribution of the *eD* and the *eN* is separated as

$$H(W, W_K, i_{in}) = f(W, AF(W_K, i_{in})), \tag{3}$$

where f is the function of the dendrite; AF is the activation function of the *eN*. As the voltage between aAxon and eDendrite is complex, the AF is introduced for a specific synapse weight as $AF(W_K, i_{in}) = H(W_{ref}, W_K, i_{in})$ to get a current-to-current function.

The model is extended to handle up to three neurons in a parallel association. To test the influence of an *eN* connected in parallel to an *eND* eN_1, the following cases are studied:

1) The eN_1 is alone to simulate its characteristic response. Another eN_{load} is connected to eN_1 as an output impedance through a dendrite having a width $W_{DEN,1}$.
2) A second eN_2 is connected to the input of eN_{load} in parallel of eN_1, with a constant medium current excitation i_{N2}. The dendrite used to connect eN_2 has a width $W_{DEN,2} = 0.135\mu$m.
3) Same circuit as case 2, but with $W_{DEN,2} = 1.35\mu$m.
4) eN_3 is added to the input of eN_{load} with a constant current excitation i_{N3}. $W_{DEN,2} = W_{DEN,3} = 0.675\ \mu$m. i_{N3} is chosen to avoid beat phenomenon using $|f_{spike,eN_2}(i_{N2}) - f_{spike,eN_3}(i_{N3})| >> \frac{1}{\tau}$ where τ is the simulation time. Moreover, $i_{N2} \approx i_{N3}$ to get roughly the same excitation level.

A. Negative bias proposition and impact on AF

In the literature, analog *eN* inhibition is a challenge in circuit design trade-offs. However, in machine learning the use of negative bias can strongly increase the performance of the network [14]. For software-based NN, bias is implemented using (4) [14]. Choosing $i_{biais} \leq 0$ causes a right shift of the activation function, which models neuron inhibition.

$$i_{out} = H_{software}(\omega, i_{in} + i_{bias}). \tag{4}$$

Changing the gate width (W_K) of the discharging transistor MN_K (Fig. 1(b)), enables the *eN* to emulate the function

$$H(W_0, W_K, i_{in}) = \begin{cases} i_{min}\ for\ i_{in} \leq i_{tr} \\ G \cdot H(W_0, W_{K,ref}, i_{in} - i_{tr})\ for\ i_{in} > i_{tr} \end{cases} \tag{5}$$

where G and i_{tr} are respectively a constant gain and a threshold current depending on the value of W_K; $W_{K,ref}$ being a fixed value of W_K; i_{min} the minimal *eN*'s RMS output current

Two effects can be noticed. First, H is shifted to the right and starts increasing after a threshold current. Indeed,

increasing W_K causes a greater discharging current of the capacitor C_k (see Fig. 1(b)). This capacitor will require a greater current i_{in} to increase the spike frequency above the rest frequency (spiking due to *eN*'s noise in the power line).

The second effect is attenuation ($G \leq 1$) of i_{out} post-threshold. Since a bigger discharging current change the spike's wave form by reducing the voltage drop time, the average voltage of the spike decreases while the high pass characteristic of the synapse selects $<V_{GK}>$ causing i_{out} to be less than the reference for a fixed frequency. G is assumed independent of i_{in} to keep the model simple in order to be efficiently trainable by digital frameworks such as TensorFlow.

B. $f(W,AF)$ dependency for a single dendrite

The case 1 is used to model the behaviors of a $h+1^{th}$ layer dendrite. Adding a dendrite after an *eN* at the node V_P create a current mirror. Using classical current mirror equation in weak inversion region from [15], and by assuming that the load effect of dendrites in V_P (see Fig. 1(b)) is negligible, the output current of the mirror i_{out} is

$$i_{out} = \frac{I_{ref} W}{W_{ref}} \frac{1 - e^{\frac{V_{DS}}{\phi_T}}}{1 - e^{\frac{V_{DS,ref}}{\phi_T}}}. \tag{6}$$

where I_{ref} is the output current of the mirror at $W = W_{ref}$; V_{th} is the threshold voltage; $n \cdot \phi_T$ the slope factor and thermal voltage taken as constants; V_{DS} the drain to source voltage since bulk is connected to source in 1(b). $I_{ref}(i_{in})$ is defined as $I_{ref}(i_{in}) = AF(W_K, i_{in})$: the output current for the reference width and $V_{DS,ref}$, V_{DS} in this situation. Thus, one can deduce from (6)

$$i_{out} = g_{ideal}(W) \cdot AF(W_K, i_{in}), \tag{7}$$

assuming $g_{ideal}(W)$ independent of the input current i_{in}. To include analog defects, leakage currents are considered. Leakage depending on the excitation rate of the *eN* will be handled in a non-ideal $g(W)$ term. Besides, constant leakage in the dendrite is described by a current $i_L(W)$. Finally, the dendrite effect is summarized in

$$H(W, W_K, i_{in}) = g(W) \cdot AF(W_K, i_{in}) + i_L(W), \tag{8}$$

$$f(W, x) = g(W) \cdot x + i_L(W), \tag{9}$$

where by definition, $g(W_{ref}) = 1$, $i_L(W_{ref}) = 0$. f is then obtained via (3). $i_L(W)$ is mainly due to band-to-band-tunneling current between the substrate bias at V_{DD} to the drain. This leakage is proportional to the electrical field of the dendrite's transistor junction [16]. Assuming side effects negligible, this field is proportional to $+W$ then positive value of $i_L(W)$ is expected for $W > W_{ref}$.

C. eN's environment influence

To simulate the *eND* behavior as it would be in a neural network, the electrical environment of the $(h-1)^{th}$ and $(h+1)^{th}$ layer should be considered. Moreover, a key parameter of a neural network is the weights of the synapses. Thus a model of the analog response of *eND* should only take as parameters the weight of its neighbors and omit their excitation to use the same paradigms as software-based networks.

The model of the *eND* should consider: 1) the impact of the inputs dendrites, 2) the influence of the output connections number (see V_P in Fig. 1(b)) and 3) the dependency for the same layer eNs (connected together on the next layer dendrites).

1) Input dendrite impact on H: The model is established for only one input dendrite. In a complete network scenario, a dendrite should be driven by a complete *eN*. To simplify the study, the dendrite is replaced by a PMOS impedance controlled by a constant gate voltage. Notice that a group of input dendrites can be approximate by PMOS transistors in parallel with the supply voltage line.

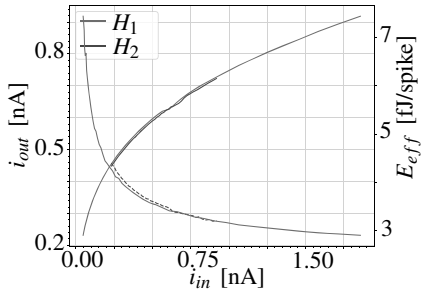


Fig. 2. Energy efficiency and *eND* behavior for two driving methods. Blue represent those functions for a *eND* driving circuit and green in a case of a constantly biased PMOS driving circuit.

To justify this approximation, the characteristic values H and E_{eff} of an *eND* are plotted in Fig. 2 for a PMOS approximation driven method as H_1 in green, and for a complete *eN* driving circuit H_2, shows in blue. The *eND* is connected to a second *eN* (with an output dendrite to ground) to emulate a realistic output load impedance. Figure 2 illustrates the *eND*'s root main square (RMS) output current i_{out} (continuous line) and the energy efficiency E_{eff} (dashed line) as a function of the RMS value i_{in}. In the two driving method, the *eND* has the same results with a relative root mean square error $\sigma_e = 0.34\%$ on E_{eff} and $\sigma_e = 1.9\%$ on i_{out}.

As the results are similar, all the analysis of Sec. III shall consider a PMOS input impedance as dendrites. In the Fig. 2 the input current domain [0, i_{max}] isn't totally solicited for H_2. This is the case because the dendrite width W_{DEN} is set to 0.675 μm, but any input current in [0, i_{max}] is accessible by using the correct W_{DEN} between W_{min} and W_{max}. Consequently the author will assume that the *eND*'s response for the domain [0, i_{max}] can be evaluated using a PMOS driving method.

2) $(h+1)^{th}$ layer's connection influence on AF: For each connection of an *eN* to another one in the $(h+1)^{th}$ layer, a

dendrite is connected to the axon. Since the input impedance of a dendrite is equal to the gate impedance of PMOS, the load of the dendrite on the axon is low. As a result, only the case with one output dendrite connected is studied.

3) Parallel association: Finally, the effect of a parallel association is quantified on AF and f. The result is obtained for a single bias W_K, then W_K will be omitted from notations. The function $\hat{AF}_i(W, i_{in})$ of (10) is introduced to track the effect of the i^{th} environment on the activation function. This represents a estimated activation function based on the response $H(W, i_{in})$ for the i-th case and a dendrite width equal to W by decoupling the influence of the dendrite from the eN.

$$\hat{AF}_i(W, i_{in}) = \frac{H(W, i_{in}) - b_i(W)}{a_i(W)} \text{ and} \tag{10}$$

$$H_{real}(W, i_{in}) \approx a_i(W) \cdot AF(i_{in}) + b_i(W), \tag{11}$$

where $a_i(W)$ and $b_i(W)$ for (11) is fitted using least square method on the whole range [0, i_{max}] (i_{max} = 0.9 nA). As the coupling between axon and dendrite is low, results are expected to validate $\hat{AF}_i(W) \approx AF$. In this situation $a_i(W) = g(W)$ and $b_i(W) = i_L(W)$ for the case i. The impact of the environment on the dendrite is shown by the influence of i on $a_i(W), b_i(W)$.

III. Results and Discussions

The circuits are implemented using the BiCMOS SiGe 55 nm technology from ST Microelectronics. Figure 3 illustrates the layout of an eNeuron, occupying a $10.8 \times 9.7\ \mu m^2$ area. For the next sections, all simulation results are based on Cadence Virtuoso analysis post-layout simulation (PLS). For each variation of a transistor's width, the layout of the whole circuit was redesigned to obtain a correct post-layout result.

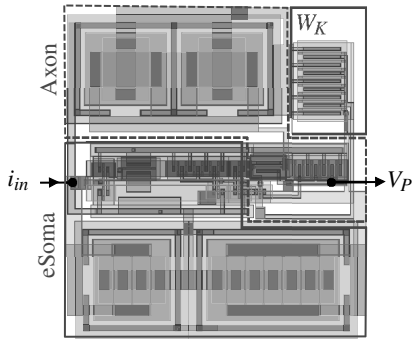


Fig. 3. Layout of an eN without dendrites in the maximum area configuration for W_K = 10 μm (on the top right corner). Area is 10.8 × 9.7 μm^2.

A. Negative bias PLS validation

To simulate the behavior of an eND constantly inhibited, the eN is connected to a PMOS dendrite, with constant gate voltage $V_0 \in [0, 120]$ mV. Another eN is connected as a load. i_{in} and i_{out} are obtained by the RMS input and output current of the eN computed on 300 μs after a transitory regime.

Figure 4(a) shows the influence of the gate width W_K of MN_K on the eND's response $H(W, W_K, i_{in})$. Two effects are observed: H is shifted to the right by a threshold current i_{tr}; the response is attenuated for wide W_K. For a current lower than i_{tr}, $i_{out}(i_{in} \leq i_{tr}) \approx i_{min}$, validating the $(i < i_{tr})$ part of (5).

Concerning the energy efficiency, Fig. 4(b) summarizes how E_{eff} is affected by the bias, i.e. the modification of W_K. The dashed line corresponds to the efficiency having W_K set to his original value (2 μm, first used in [11]) for the same layout technology. Adding a constant negative bias increase the energy efficiency near the non-linearity point (often solicited in a deep learning process) with a maximum increase of 68.0% higher than $E_{eff}(i_{max} \approx 2nA)$. However, in the high excitation zone, E_{eff} is 24.5% lower than in the standard configuration. In a power perspective, P_{eN} decrease for i_{in} less than i_{tr} making a silicon area trade-off rather than a consumption issue, between deep-learning abilities and analog implementation.

The function $M(W_K, i_{in})$ defined by

$$M(W_K, i_{in}) = \frac{H(W_0, W_K, i_{in} - i_{tr}(W_K))}{H(W_0, W_{K,model\ ref}, i_{in} - i_{tr,model\ ref}))}, \tag{12}$$

is introduced to quantify the representative of the model (5). $i_{tr}(W_K)$ is the threshold current of $H(W_0, W_K, i_{in})$; $i_{tr,model\ ref} = i_{tr}(W_{K,model\ ref})$; $W_{K,model\ ref}$ a width arbitrarily chosen. M represents the ratio of the shifted activation function from the reference one ($W_{K,model\ ref}$), assumed constant to $G(W_K)$ by the model in (5).

Figure 4(c) plots M using the simulation results of Fig. 4(a). The maximum relative gap from a constant $G(W_K)$ is 21 % and 14.5 % for $W_{K,model\ ref}$ of 2.5 and 8 μm respectively. The deviation of M to G is wider near the threshold point. However, the model's accuracy has to be related to the transitor's width variability of the manufacturing process. The model has a better correspondence for $W_K \approx W_{model\ ref}$. In a training process, the network can be trained by a midrange $W_{model\ ref}$ at first, and then, by a refined model with $W_{model\ ref}$ adapted near the values of width corresponding to each i_{tr} obtained.

Figure 5 represents i_{tr} (solid line) and G (dashed line) as a function of W_K. The two functions are monotonous with a fairly constant slope, which makes those values less sensitive to the variability of W_K during manufacturing. High inhibition would use more silicon area.

B. $f(W,x)$ PLS results for a single dendrite

The activation function used in the four cases is obtained for $W_{DEN,1} = 0.135\ \mu m$ and W_K set to 2.5 μm as it provided the minimal inhibition while presenting $AF(i_{in} \approx 0) \approx i_{min}$.

Figure 6 shows $\hat{AF}_1(W, i_{in})$ in the environment of case 1 for 8 evenly spaced widths from $W_{min} = 0.135\ \mu m$ to $W_{max} = 1.35\ \mu m$. The blue line represents the typical activation function $\hat{AF}_{typ}(W, i_{in})$ computed as the average on the width of all estimated AF. The surface colored in light blue represents the maximal deviation of $\hat{AF}_1(W, i_{in})$ from the typical one.

On this first environment, all $\hat{AF}_i(W, i_{in})$ match the reference with a σ_e = 1.8 %, and a maximum absolute error of 0.8 pA. This validates the model $H_W = f(W, AF)$ where the influence of the dendrite is decoupled from the eN behavior. Indeed, the

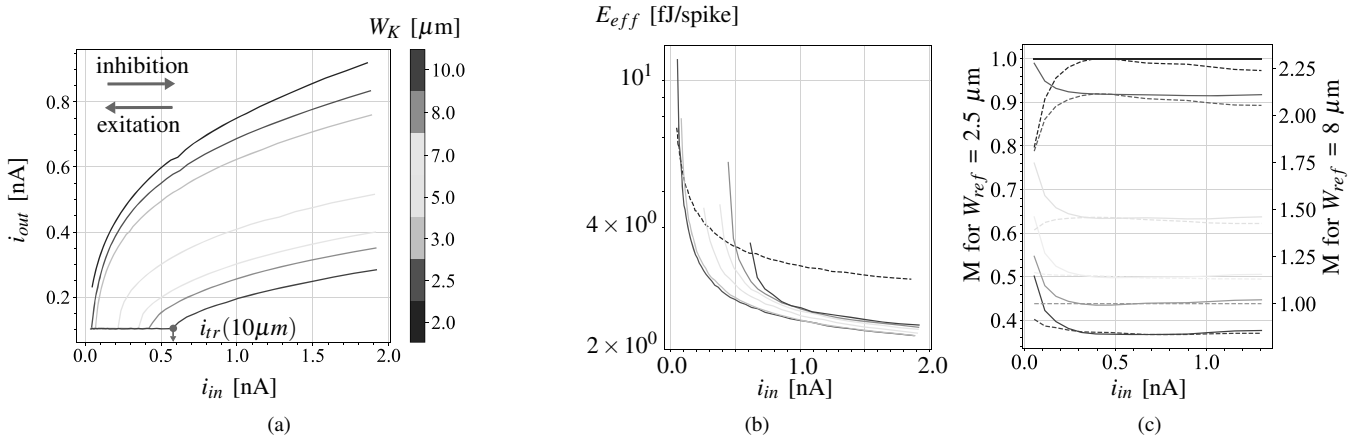


Fig. 4. (a) Activation function for 8 values of W_K. Reference activation function has $W_K = 2.5\mu m$ (b) energy efficiency against the input current i_{in} of the *eN*. (c) present the variable $M(i_{in})$, representing the post-threshold gain. In continuous line, M is computed for W_{ref} = 8 μm and 2.5 μm for the dashed one.

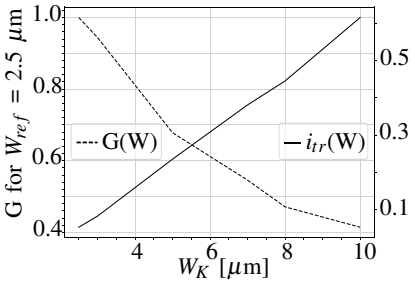


Fig. 5. Variation of the threshold current i_{tr} in black line and post-threshold activation function gain G with the width of MN_K

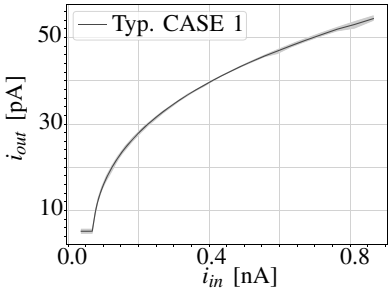


Fig. 6. Activation function of the eN_1. Changing the width $W_{DEN,1}$ of the next-layer dendrite cause the AF to slightly deform, as it's plotted by the blue surface (functions $AF((W), i_{in})$). Relative standard deviation is σ = 1.8 % and the maximum deviation is 0.8 pA

AF is independent of dendrites' width (i.e. synaptic gain) as the deviation of $\hat{AF}_i(W, i_{in})$ is low. Besides, the linear fitting of f validate that V_{DS} in (6) marginally depend on i_{in}, i.e $g(W, i_{in}) \approx g(W)$.

Figure 7(a) represents the estimated standardized gain $C_i(W) = \frac{a_i(W)}{W}$ for the four cases. The dashed black line represents $C_1(W)$ for case 1. According to (6) in traditional current mirror assumption, $C_1(W)$ should be a constant equal to $\frac{1}{W_{ref}}$. However, Fig. 7 shows that $C_1(W)$ changes significantly with a factor of 2.12 between small (0.135 μm) and large geometries (1.35 μm). This illustrates the impact of small geometries on the constants V_{th} and n, studied in [17]. Dissipative phenomena are put in evidence for small width. Because $C_1(W) > \frac{1}{W_{ref}}$ for $W > W_{ref}$, a positive drain leakage appears for big widths. This leakage is proportional to the excitation level of the *eN* as it impacts the gain $C_1(W)$.

Figure 7(b) represents the bias current $b_1(W)$ in dashed black line. Two things can be observed: $b_1(W)$ is positive and tend to a linear asymptote. Indeed, $b_1(W)$ is the constant (of i_{in}) drain's leaking current of the dendrite i_L in relation to the leaking of the reference dendrite. Since the reference dendrite has the minimal width, all dendrites considered increase leakage so $b_1(W) > 0$. The linear asymptote highlights the proportional dependency of i_l on W for band-to-band-tunneling leakage current. Notice that $b_1(W)$ is in the same order of magnitude as i_{in} so cannot be neglected for high gain. Associating eNs in parallel would produce an extra excitation to the following *eN*.

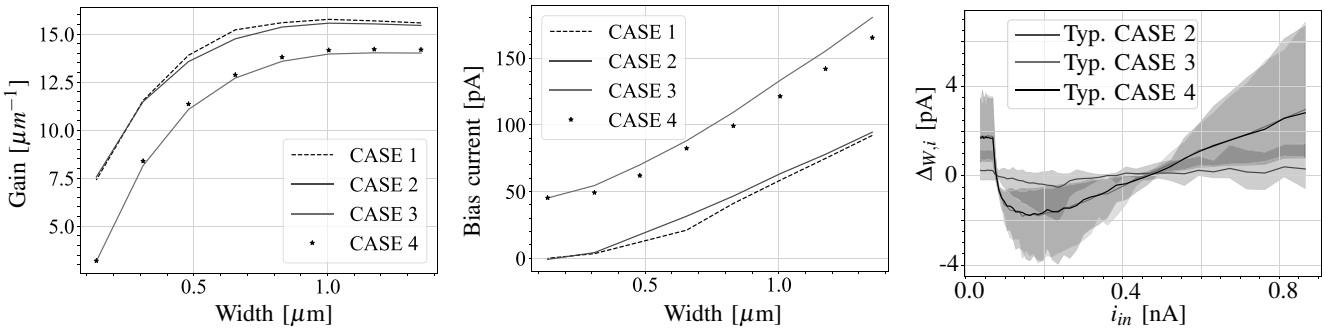
C. Parallel association impact

To evaluate the impact of the parallel association on the behavior H of the *eND*, cases 2,3 and 4 are considered. $\Delta_{W,i}$ is introduced as:

$$\Delta_{W,i} = \hat{AF}_i(W, i_{in}) - AF_{d,1,typ}. \tag{13}$$

to quantify the dependency of the activation function on the i-th case environment.

Figure 7(c) represents the typical value of $\Delta_{W,i}$ (solid line) and its maximum deviations (in transparent surfaces), showing a slight modification of $\hat{AF}$. The deviation of $\hat{AF}_i(W, i_{in})$ seems to increase with a complex environment but remains small (pA << nA for i_{in}). However, the typical $\Delta_{W,i}$ is constantly bias and can be partially corrected by adapting the model of AF for each layer depending on the *eN*'s number.



(a) (b) (c)

Fig. 7. Evolution of the normalized gain, and the bias current of the linear model of $f_{W,AF}$. Leakage make those parameters nonconstant. Variation between cases 1 to 4 shows the dependency of $f(W,AF)$ from his environment. (a) plot the parameter $C_i(W)$, (b): $b_i(W)$ for the whole input current range. (c): Deviation of $\hat{AF}_i(W,i_{in})$ in case 2,3 and 4 from the typical $\hat{AF}_{1,typ}(i_{in})$ when eN_1 is alone. Transparent surface represents the maximum deviation from the typical value (i.e. the average) for each case.

Figure 7(a) presents how the gain of the current mirror is affected by the gain of the other dendrites. The fact that $C_i(W)$ decreases with the gain of the other neurons on the same layer can be understood by the apparition of a leakage current from the eD to the other ones (because $C_2, C_3, C_4 \leq C_1$), depending on their width. For instance, $C_3(W)$ of case 3 is the same as $C_4(W)$ of case 4, while $W_{N2,case\ 3} = W_{N2,case\ 4} + W_{N3,case\ 4}$. The shift of C(W) seems to depend on the sum of the width of the other synapses of the same layer. This effect is dominant with wide $W_{DEN,k>1}$, as the gain $C_2(W)$ is close to $C_1(W)$, while $W_{N2,case\ 2}$ is small, equal to W_{min}.

Figure 7(b) shows the dependency of $b_i(W)$ to the environment. An offset of positive current going from the synapse to his environment can be observed. Like $C_3(W) \approx C_4(W)$, $b_3(W) \approx b_4(W)$, the shift of $b_i(W)$ seems to increases with the sum of the other dendrites' width of the same layer.

IV. Conclusion

Proposed eNeuron review considered the implementation of a biomimetic model of biological soma, axon, and dendrites to enable a novel way dealing with SNN limitations. Post-layout simulation results proved a non-linear action function shift from excitation and inhibition synaptic current. Proposed model is extended to handle up to three neuron connections in a parallel association. Therefore, SNN deep learning is feasible for a fan-in/fan out of 3 synaptic branches. Revisited electronic neuron achieves an area of $10.8 \times 9.7\ \mu m^2$ for an energy efficiency below 10 fJ/spike. Moreover, such a result considers the total power consumption (soma, axon, dendrites), while literature presented eNeurons and synapses separately. Relative standard deviation of synapse current below 1.8% and synapse weight mismatch ≤ 4 pA current error are in agreement with deep learning training requirements.

References

[1] C. D. Schuman, et al., "A Survey of Neuromorphic Computing and Neural Networks in Hardware," pp. 1–88, 2017. [Online]. Available: http://arxiv.org/abs/1705.06963

[2] E. E. Tsur and M. Rivlin-Etzion, "Neuromorphic implementation of motion detection using oscillation interference," *Neurocomputing*, vol. 374, pp. 54–63, 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2019.09.072

[3] S. Davidson and S. B. Furber, "Comparison of Artificial and Spiking Neural Networks on Digital Hardware," *Frontiers in Neuroscience*, vol. 15, no. April, pp. 1–7, apr 2021.

[4] J. Hasler and B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," *Frontiers in Neuroscience*, vol. 7, pp. 1–29, sep 2013.

[5] I. Sourikopoulos, et al., "A 4-fJ/spike artificial neuron in 65 nm CMOS technology," *Frontiers in Neuroscience*, vol. 11, no. 123, pp. 1–14, mar 2017.

[6] P. M. Ferreira, et al., "Neuromorphic analog spiking-modulator for audio signal processing," *Analog Integrated Circuits and Signal Processing*, vol. 106, no. 1, pp. 261–276, jan 2021. [Online]. Available: https://link.springer.com/10.1007/s10470-020-01729-3

[7] G. Indiveri, et al., "Neuromorphic silicon neuron circuits," *Frontiers in Neuroscience*, vol. 5, no. MAY, pp. 1–23, may 2011.

[8] F. Danneville, et al., "A Sub-35 pW Axon-Hillock artificial neuron circuit," *Solid-State Electron.*, vol. 153, no. January, pp. 88–92, 2019.

[9] F. Danneville, et al., "Sub-0.3V CMOS neuromorphic technology and its potential application," in *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, no. 1. Lille, France: IEEE, jun 2021, pp. 1–6.

[10] J. Ou and P. M. Ferreira, "A Tunable Morris-Lecar Spiking Neuron in CMOS," in *IEEE Midwest Symp. Circuits Syst.*, 2023, p. pp.

[11] T. Soupizet, et al., "Analog Spiking Neural Network Synthesis for the MNIST," *J. Integr. Circuits and Syst.*, vol. 18, no. 1, pp. 1–12, 2023.

[12] C. Bartolozzi and G. Indiveri, "Synaptic Dynamics in Analog VLSI," *Neural Computation*, vol. 19, no. 10, pp. 2581–2603, oct 2007.

[13] W. Xu, J. Wang, and X. Yan, "Advances in Memristor-Based Neural Networks," *Frontiers in Nanotechnology*, vol. 3, no. March, pp. 1–14, 2021.

[14] F. Chollet, *Deep Learning with Python*. New York, NY, USA: ACM, nov 2018.

[15] M. C. Schneider and C. Galup-Montoro, *CMOS Analog Design Using All-Region MOSFET Modeling*. Singapore: Cambridge University Press, 2010. [Online]. Available: http://ebooks.cambridge.org/ref/id/CBO9780511803840

[16] P. F. Butzen and R. P. Ribas, "Leakage Current in Sub-Micrometer CMOS Gates," in *Universidade Federal do Rio Grande do Sul*, 2007, pp. 1–30. [Online]. Available: http://www.inf.ufrgs.br/logics/docman/book_emicro_butzen.pdf

[17] J. Ou and P. M. Ferreira, "Implications of Small Geometry Effects on gm/ID Based Design Methodology for Analog Circuits," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 66, no. 1, pp. 81–85, jan 2019.