A Sub-35 pW Axon-Hillock artificial neuron circuit

Francois Danneville, Christophe Loyez, K. Carpentier, I. Sourikopoulos, E. Mercier, A. Cappy

▶ **To cite this version:**

Francois Danneville, Christophe Loyez, K. Carpentier, I. Sourikopoulos, E. Mercier, et al.. A Sub-35 pW Axon-Hillock artificial neuron circuit. Solid-State Electronics, 2019, 153, pp.88-92. ⟨10.1016/j.sse.2019.01.002⟩. ⟨hal-03330291⟩

HAL Id: hal-03330291

https://hal.science/hal-03330291v1

Submitted on 21 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-NC 4.0 - Attribution - Non-commercial use - International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S0038110118305562
Manuscript_0fec821725a3996cb7e3668c12529e74

SOLID-STATE ELECTRONICS

Contents lists available at ScienceDirect

Solid State Electronics

Journal homepage: www.elsevier.com

A Sub-35 pW Axon-Hillock artificial neuron circuit

F. Danneville[a,*], C. Loyez[a], K. Carpentier[b], I. Sourikopoulos[b], E. Mercier[c], A. Cappy[a]

[a] *Centre National de la Recherche Scientifique, Université Lille, USR 3380 - IRCICA, Lille, France and Centre National de la Recherche Scientifique, Université Lille, ISEN, Université Valenciennes, UMR 8520 - IEMN, Lille, France*

[b] *SATT NORD, 25 Avenue Charles Saint-Venant, Lille, France*

[c] *CEA, LETI, MINATEC Campus, Grenoble, France*

ARTICLE INFO

Article history:
Received 00 December 00
Received in revised form 00 January 00
Accepted 00 February 00

Keywords:
CMOS
Axon-Hillock
Artificial Neuron
Ultra low power
high energy efficiency

ABSTRACT

Artificial Intelligence (AI) applications are developing at a high rate, facing soon a tremendous energy challenge. In this context, the original Axon-Hillock (AH) Artificial Neuron (AN) has been optimized to achieve ultra-low power (ULP) consumption. The membrane capacitance was taken out, and in order to drastically reduce its power consumption, the (feedback) capacitance is lowered to 5 fF, the transistors gate width is reduced to 120 nm and the supply voltage is decreased to as low as 200 mV. Designed and fabricated using 65 nm CMOS Technology, the refined AH neuron features a standby power of 11 pW, and when excited, a power consumption that does not exceed 30 pW for a firing frequency of 15.6 kHz. Its energy efficiency per spike is lower than 2 fJ / spike when the DC power is included (around 1 fJ / spike excluding the DC power), for an area of 31 µm². These performance confer to this ULP AH neuron a high potential for future development of highly energy efficient Spiking Neural Networks, required to design future neuroprocessors embedded in various applications (smart visual sensors for autonomous vehicles, robotics).

© 2018 xxxxxxxx. Hosting by Elsevier B.V. All rights reserved.

1. Introduction

Moore's law is reaching its end, which paradoxically may be viewed as an opportunity [1]. This motivates to investigate new paradigms particularly to address the energy dissipation challenge for huge AI applications. In this context, spiking neural networks (SNNs) constitute an interesting alternative to process information, in view of providing cognitive characteristics. Hence, the development of high energy efficient artificial neurons and synapses using standard CMOS technology is important. In a previous work [2], we have proposed an AN having an outstanding energy efficiency of few fJ / spike, whose features are very close to the biology, because it constitutes an approximation of biological Morris-Lecar (ML) model. In its simplest version, this ML AN uses only six transistors and two capacitances (Fig. 1). Its performance was basically achieved by applying the following design rules: decreasing as much as possible the supply voltage *and* the membrane capacitance values. Following this work, we have carefully looked throughout the abundant literature for other CMOS based AN architectures [3,5-7], to identify which one would possibly be competitive in terms of reducing both the DC power and the energy efficiency against the ML AN. It turned out that the Axon-Hillock (AH) artificial neuron originally proposed by C. Mead [4] was worth to be optimized to address the ultra-low power challenge, and this is the main objective of this work.

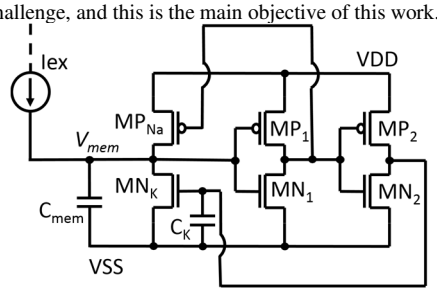

Fig. 1. Simplified Morris-Lecar Neuron Circuit.

* *Corresponding author.*
E-mail address: francois.danneville@univ-lille.fr

© 2019 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

The rest of this article is organized as follows: The original AH AN architecture is recalled and the simplified version that was implemented onto silicon is presented. Then the electrical behavior, power consumption, temperature and supply voltage dependence, process variability robustness, are discussed for the refined AH AN. All simulations were carried out by Spectre simulator, on typical corner PDK models, and performed at 300°K (besides temperature sweeps). Details about the circuit fabrication and measurement setup follow and then experimental AH AN results and performances are presented at 300°K. This article concludes with a benchmarking of state-of-the-art artificial neurons in conclusion.

2. From the Original Axon-Hillock to its simplified ultra low power version

2.1 Original Axon-Hillock

The original Axon-Hillock circuit is presented in Fig. 2 [3, 4]. The circuit is made of a Voltage Amplifier (VA) –usually made using two inverters cascaded in series- and uses two capacitances: the membrane capacitance C_{mem} and a feedback capacitance C_f. I_{ex} is the excitatory current, which mimics the overall synaptic current flowing out of the dendritic tree.

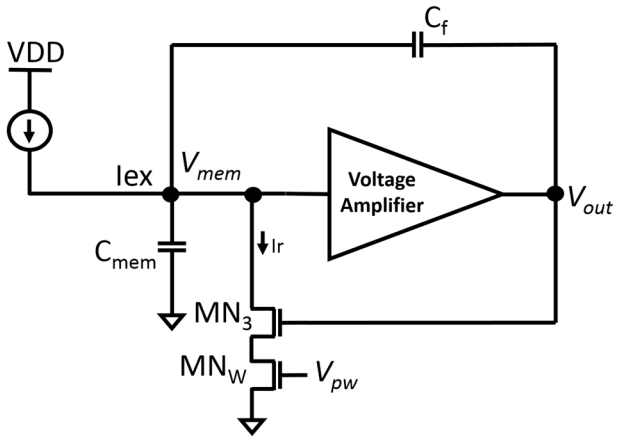

Fig. 2. Original Axon-Hillock artificial neuron (as drawn in [3]).

Because the main objective was to achieve extremely low DC power and an outstanding energy efficiency for any spike generation, the following arrangements were applied with respect to the original circuit shown in Fig. 2. The first one was to remove the explicit membrane capacitance C_{mem}, keeping only the parasitic component, which corresponds to the first inverter input capacitance. The second one was to remove MN_W, the nMOS to which the "weight" voltage $V_{pw,}$ is applied in Fig. 2; the current I_r can be set by adjusting transistor dimensions as explained in the next section. The (nominal) supply voltage VDD was chosen to 200 mV, ensuring that MOSFETs will operate in deep sub-threshold regime. The source to drain voltage of any MOSFET conductive channel cannot exceed VDD. Such a supply voltage favors low power consumption. After these arrangements, the refined AH architecture is depicted in Fig. 3.

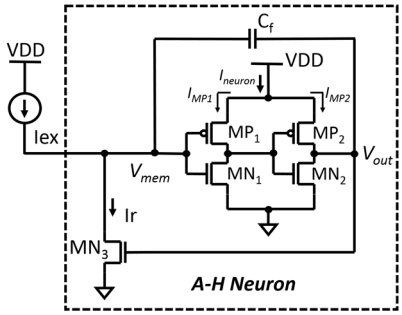

Fig. 3. Refined AH artificial neuron.

2.2 Basic behavior of the refined Axon-Hillock circuit

The basic behavior of the circuit drawn in Fig. 3 is as follows: without excitatory current (I_{ex} = 0 A), the output and membrane voltages (V_{mem} *and* V_{out}) are 0 V, C_f is not charged and MN_3 (nMOS) is OFF. When an excitatory (DC) current I_{ex} is applied, a charge is stored by C_f and V_{mem} increases. When the magnitude of I_{ex} is sufficient, V_{mem} reaches the switching voltage of the first inverter, both inverters change states, and V_{out} rises towards VDD (Fig. 4). Meanwhile, a positive feedback occurs through C_f, pulling up the membrane voltage V_{mem} to a positive value higher than V_{out} and MN_3 turns ON. V_{out} magnitude is limited by the voltage, which develops between drain and source of MP_2, following the increase of the reset current I_r. I_r is now set by the conductance ratio (and effectively the sizing) of the pull-up (MP_2) and pull-down (MN_3) transistor. With V_{out} in a high state, having an I_r much larger than I_{ex}, will cause V_{mem} to decrease and to reach (again) the switching voltage of inverter 1 (Fig. 4); as a result, the inverters switch again, forcing V_{out} to 0 V. V_{mem} changes sign, thus source and drain electrodes for MN_3 are interchanged. Through the adding currents (I_{ex} + $|I_r|$), the membrane potential is rising towards positive values and when the membrane potential crosses 0 V, source and drain for MN_3 recover their original location (that is, the source is connected to the ground in Fig. 3) and leading MN_3 to be turned OFF again. C_f is charged again through only I_{ex}, and the cycle starts again.

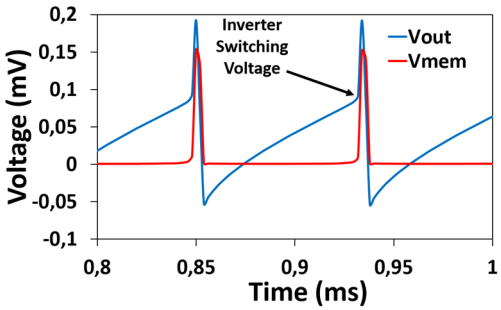

Fig. 4. Membrane and output voltages (Cadence simulation), for Iex = 10 pA (W/L=120nm/65nm for transistors, C_f =5 fF).

2.3 Power consumption of the refined AH circuit

First of all, when the neuron is not excited (Iex = 0 A), the total DC power consumption of inverter 1 (Pinv1) and of inverter 2 (Pinv2) is around 1 pW in simulation (W/L=120nm/65nm for transistors, C_f =5 fF). This low power consumption is explained by the fact that that both the membrane and output voltages are 0 V when the neuron is not excited. Nevertheless, when an excitatory current is applied while not sufficient to trigger a spike, the membrane voltage increases; it turns out that both MP1 and MN1 are in an "on-state", which contributes to increase the DC current. Hence, just before the AH is going to fire, the DC current of inverter1 has increased to around 20 pA, which leads to a dissipated DC power of 4 pW. At the mean time, Pinv2 is negligible.

When the AH is firing, the power delivered to each inverter is plotted in Fig. 5 (note that a power consumption analysis of the Axon-Hillock neuron can be found in [8]). The power is obtained through the product of VDD times the DC current (that is, the mean value) for each inverter.

In Fig. 5, the dissipated power of inverter 1 (Pinv1) is rather constant while those of inverter 2 (Pinv2) continuously increases. The capacitive load of inverter 1 is limited to a low value, set by the parasitic capacitance of inverter 2, making its "switching power" negligible. Nevertheless, because the membrane voltage slowly varies as a function of time (see Fig. 4), the current I_{MP1} is also slowly varying with a rising time close to the period of the spike; thus, the inverter 1 dissipates a "short-circuit power" [9], almost independent of the firing frequency (Fig. 5).

The situation is very different for inverter 2. Indeed, as shown in Fig. 4, out of T_H duration, V_{out} is close to 0 V, forcing the current flowing in MP2/MN2 to a low value; indeed, the supply voltage mainly delivers the current I_{MP2} during the spike duration T_H (Fig. 4), which is found almost independent through simulation. Because during T_H, I_{MP2} equals the capacitive current flowing in C_f, Pinv2 is frequency dependent, signature of a "switching power".

From this qualitative analysis, it turns out that Pinv1 is the major neuron dissipated power for the low excitatory currents (or low firing frequencies) while Pinv2 is the major one for the highest excitatory currents (or high firing frequencies).

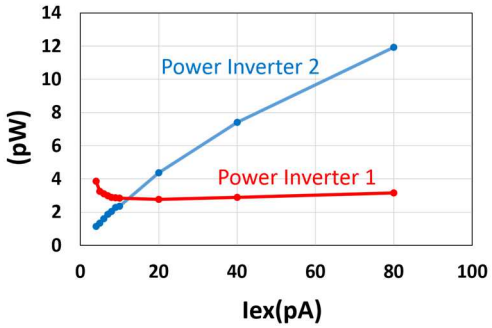

Fig. 5. Power dissipated (Cadence simulation) in each inverter of the AH refined neuron when excited (*W/L*=120nm/65nm for transistors, C_f =5 fF).

In this context, the elements sizing for the designed AH NA were chosen by following these considerations: (i) in order to decrease as much as possible the DC/short-circuit powers, the gate width (W) for all transistors were chosen as low as possible (for the CMOS technology used), (ii) in order to reduce as much as possible the switching power, the value for feedback capacitance C_f was chosen as low as possible, paying attention that it remains much higher than parasitic capacitance of inverter 1.

2.4 Impact of Temperature

The variation of spike frequency is plotted as a function of temperature in Fig. 6, for Iex = 10 pA. When temperature increases from 25°C to 40°C, a 20% spike frequency decrease is observed. Such a drop is directly related to the deep sub-threshold operation of the CMOS transistors [11].

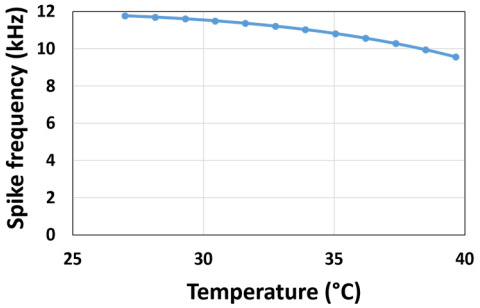

Fig. 6. Firing frequency (Cadence simulation) as function of temperature, for Iex = 10 pA (*W/L*=120nm/65nm for transistors, C_f =5 fF).

2.4 Impact of supply voltage (VDD) variation

The variation of spike frequency is plotted as a function of the supply voltage in Fig. 7. For constant Iex = 10 pA and a supply voltage variation of 0.2 ± 0.05 V the frequency shift is of ± 25% from the nominal firing frequency (that is 12 kHz for VDD = 0.2V).

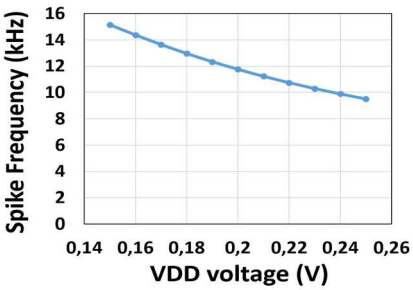

Fig. 7. Firing frequency (Cadence simulation) as function of VDD, for Iex = 10 pA (*W/L*=120nm/65nm for transistors, C_f =5 fF).

Fig. 7 provides also good information about how the AH AN could behave to potential VDD noise perturbation.

2.5 Process variability robustness

As previously mentioned, the transistors are operating in deep-threshold regime, and their current-voltage characteristics are subject to variations. Fig. 8 shows a Monte Carlo simulation with one thousand simulations with the criterion to have the AH AN spiking output peak-to-peak voltage greater than half of VDD. As it is seen in Fig. 8, the yield was 820/1000.

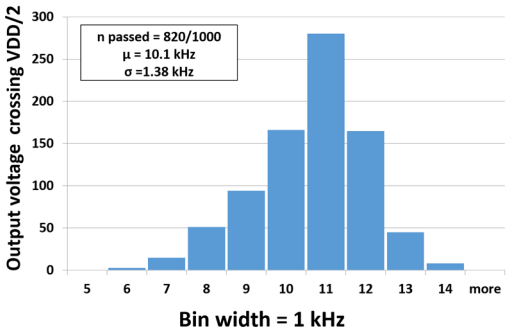

Fig. 8. Monte Carlo (Cadence) simulation: the criterion is to have the AH AN output voltage magnitude exceeding 50% VDD (half VDD).

In conclusion to this section, we should underline that the extremely low excitation and consumption characteristics of the proposed circuit are indeed accompanied by an important margin of operation to what regards PVT variation. This should be taken into account when designing multi-neuron networks with information representation schemes utilizing spiking frequency, inter-spike interval, time to first spike etc.

3. Circuit fabrication and measurement setup

The AH neuron implemented onto silicon was designed using *W/L*=120nm/65nm transistors (all transistors have the same dimensions,

with all nMOS body contacts connected to ground and all pMOS body contacts connected to VDD) and a MOM feedback capacitance C_f=5fF. It was fabricated using TSMC 65 nm process in the LP option.

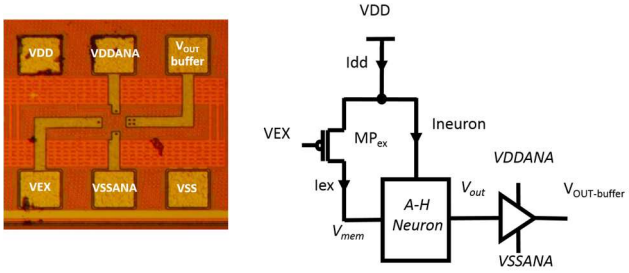

Fig. 9. Chip photograph (left) and block diagram (right).

It used external biasing, as shown in Fig. 6 ("VDD", "VSS" and VEX pads). The excitation current was tuned through on-chip externally biased transconductance MP_{ex} (Fig. 9) modeling the synaptic pre-neuronal excitation. The electrical measurements were performed on a probe station and voltage supplies resolution in the order of 100 fA.

In order to save area on the chip (this one comprised other circuits and the available area was constrained), a common VDD –set to 200 mV- was connected to both the AH neuron and the excitation transconductance. The AH neuron output was monitored through an on-chip unity gain buffer ("V_{OUT}-buffer" pad in Fig. 9) that was designed ensuring that the frequency response of the neuron circuit would not be affected. More precisely, the buffer featured a high input impedance (corresponding to a capacitance equal to 3 fF, the impact of parasitic DC current is negligible). The output buffer also featured independent DC supply (connected to the pads "VDDANA" and "VSSANA" in Fig. 9) to enable accurate power consumption measurements. The AH neuron core area is 31 μm^2.

4. Experimental results

Measurements, using the protocol described in the previous section, were performed on the chip.

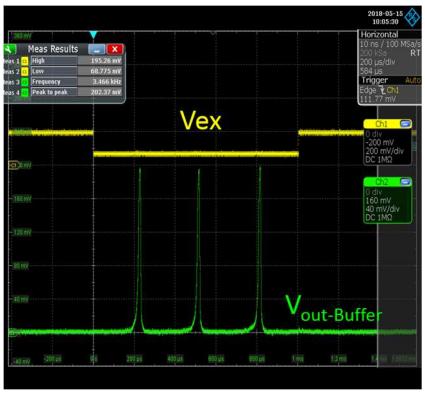

Fig. 10. Waveform of the AH AN output voltage.

A step response is reported in Fig. 10. A voltage VEX close to VDD, so that the neuron doesn't spike, is initially applied to the gate of MPex, thus corresponding to an off-state. After 400µs, a VEX step is applied and the AH neuron generates spikes for as long as VEX is sustained. When VEX returns to its initial value (≈ VDD), the neuron is not excited anymore. Because it was not possible to check the exact buffer gain value (theoretically equal to unity), the DC bias current source connected to the buffer was tuned so that the maximum voltage for $V_{OUT\text{-}buffer}$ reaches 200 mV (VDD). As it is shown, the shape of the output waveform (Fig. 10) agrees well with the simulated waveform in Fig. 4. After verifying the correct behavior of the simplified AH neuron, its performance was measured in terms of frequency, power consumption and spike energy efficiency. Because the excitation current is tuned by adjusting the gate voltage applied on pad "VEX" of the transconductance (Fig. 9), the measured (DC) current Idd delivered by the supply voltage -the only one accessible- varies. As shown in Fig. 9, Idd corresponds to the summation of both the excitatory current I_{ex} (flowing out the transconductance) and the current consumed by the neuron I_{neuron} (transconductance and neuron supply voltages are physically connected to the same supply pad). Thus, the actual power consumption and energy efficiency for the designed AH neuron are likely slightly better.

In Fig. 11, the AH neuron starts firing with a frequency equal to 290 Hz (Idd = 57 pA), then the firing frequency continuously increases up to 15.6 kHz (Idd = 150 pA), covering more than a decade of frequency. The power consumption, which is reported in Fig. 11, corresponds to the product of the supply voltage times the supply current *VDD x Idd* (VDD=200mV). On one hand, it corresponds to the "DC power" when the AH neuron is not excited. On the other hand, it includes both the DC and dynamic powers when the AH neuron is excited. As shown in Fig. 11, the DC power is around 11 pW, while when the AH neuron reaches its highest frequency, the power consumption goes up to 30 pW; the excess of power -19 pW- corresponds to the dynamic power.

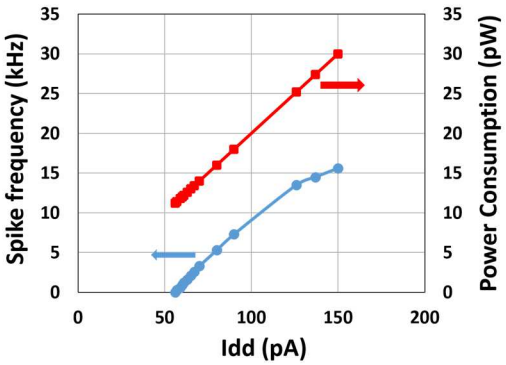

Fig. 11. Firing frequency and power consumption as a function of the DC current (I_{dd}) delivered by the supply voltage VDD.

Knowing the power consumption and firing frequency, the energy efficiency per spike of the AH neuron, including or excluding the DC power, was calculated (Fig. 12). In this plot, when the DC power is included, the energy efficiency is equal to 10.2 fJ / spike for a spike frequency of 1.2 kHz and drops afterwards to attain its lowest value of 1.9 fJ / spike for the highest spike frequency. As discussed in Section 2 related to the power dissipation discussion, when the AH neuron is spiking at low frequencies, the dissipated power of inverter 1 (almost frequency independent) is the most important, which explains the decrease of energy efficiency in Fig. 12. On the other hand, for the highest firing frequencies, the inverter 2 dissipated switching power explains why the energy efficiency becomes independent.

It is to be noted that when the DC power (namely, the power when Iex = 0 A) is deducted, the energy efficiency stands around 1 fJ/spike, whatever the firing frequency.

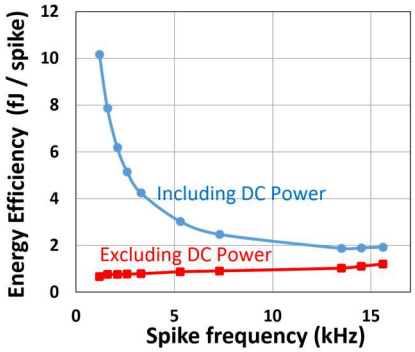

Fig. 12. Energy efficiency as a function of the spiking frequency

5. Conclusion

The performance of recent state of the art neuron implementations is reported in Table 1.

Table 1
Comparison of recent silicon artificial neuron implementations

Ref.	Node (nm)	Area (μm^2)	Spiking frequency (Hz)	Power (W)	Energy Efficiency (pJ/ spike)
[5]	350	1887	100	1.74n	17.4
[6]	65	120	$1.9\ 10^6$	78µ	41
[7]	90	442	100	40p	0.4
[2]	65	35	26×10^3	105p	0.004
This work	65	31	15.7×10^3	30p	0.002

Let us describe the performance reported in Table I. In [5] the capacitances are in the order of few hundred fF and voltages are higher than 1V. In [6], the architecture corresponds to a leaky integrate and fire AN, which adds up an extra circuit (comparator) to set the firing threshold AN, the latter contributing to extra power consumption. Moreover, the membrane capacitance is equal to 500 fF and the AN operates at supply voltage VDD around 1 V. In [7], the architecture of the AN is based on a transconductance amplifier designed with more than 10 transistors, and operates under VDD equal to 0.6 V. It is obvious that for these AN, a down-scaling on the membrane capacitance nor on the supply voltage VDD was not carried out, leading to a non-optimized energy efficiency.
In light of the comparison between the ML [2] and the AH neurons, the following comments can be addressed: (i) the DC power is three times lower for the AH. As already stated, the main reason lies in the fact that both the membrane and output voltages are 0 V when the neuron is not excited. When not excited, the membrane voltage is of few tens of mV for the ML, making the sodium (MP_{Na}) and potassium (MN_K) transistors (Fig. 1) dissipating a DC power, (ii) the energy efficiency per spike (including the DC power) is better for the AH: the benefit comes from the lower DC power and also the use of o*nly* one capacitance.

It is to be noted that the emulation of bursting mode (of interest in a robotic context) or the highlight of stochastic resonance, which has been shown for the ML [10], could also be obtained with the AH.

In conclusion, we believe that the ULP AH neuron presented in this work constitutes a serious candidate for SNNs design, if combined with synapses such the one described in [12], to address the energy dissipation challenge to come due to the growing AI applications. Further works need to address the temperature sensitivity of such a subthreshold circuit.

Acknowledgements

The authors acknowledge IRCICA and SATT NORD for the financial support. The personnel of the "Centrale de Caractérisation" of IEMN for its contribution to the electrical measurements of the chips. The authors also acknowledge the contribution of CEA Leti staff for its permanent support to this work.

References

[1] Waldrop M.M. The chips are down for Moore's law. Nature 2016;530:144-7.
[2] Sourikopoulos I., Hedayat S., Loyez C., Danneville F., Hoel V., Mercier E., Cappy A. A 4-fJ/Spike Artificial Neuron in 65 nm CMOS Technology. Front. Neurosci. 2017;11(123):1-14.
[3] Indiveri G., Linares-Barranco B., Hamilton T.J., van Schaik A., Etienne-Cummings R., Delbruck T. et al. Neuromorphic Silicon Neuron Circuits. Front. Neurosci. 2011;5(73):1-23.
[4] Mead CA. Analog VLSI and Neural Systems. Reading. 1989. MA: Addison-Wesley.
[5] Basu A., Hasler PE. Nullcline-Based Design of a Silicon Neuron. IEEE Transactions on Circuits and Systems I 2010;57(11):2938-47.
[6] Joubert A., Belhadj B., Temam O., Héliot, R. Hardware spiking neurons design: Analog or digital?. 2012 International Joint Conference on Neural Networks (IJCNN).
[7] Cruz-Albrecht JM., Yung MW, Srinivasa N. Energy-Efficient Neuron, Synapse and STDP Integrated Circuits. IEEE Transactions on Biomedical Circuits and Systems 2012;6(3):246-56.
[8] Yao E., Basu A. VLSI Extreme Learning Machine: A Design Space Exploration. IEEE Tran. on VLSI systems. 2017;25(1):60-74.
[9] Dokic B., Pajkanovic A. Subthreshold Operated CMOS Analytic Model, IX Symposium Industrial Electronics INDEL 2012, Banja Luka.
[10] Danneville F., Sourikopoulos I., Hedayat S., Loyez C., Hoel V., Cappy A. Ultra low power analog design and technology for artificial neurons. Proc. of 31st Annual IEEE Bipolar/BiCMOS Circuits and Technology Meeting, BCTM 2017.
[11] Degnan B., Hasler J. On the temperature dependence of subthreshold currents in MOS electron inversion layers, revisited, Proc. of 2016 IEEE International Symposium on Circuits and Systems (ISCAS).
[12] Bartolozzi C., Indiveri G. Synaptic dynamics in analog VLSI. Neural Comput. 2007;10:2581-603.