

Tunneling-based CMOS Floating Gate Synapse for Low Power Spike Timing Dependent Plasticity

Michele Mastella^{1,2}, IEEE Student Member, Fabio Toso¹, Giuseppe Sciortino¹,
 Enrico Prati³, Giorgio Ferrari¹, IEEE Member

Abstract—We propose a CMOS architecture for spiking neural networks with permanent memory and online learning. It uses a three-transistors synapse with a floating node that stores the synaptic weight, programmed by using only Fowler-Nordheim tunneling current in the pA range for ultra-low power operation. A neuron with a conditioning circuit programs the floating gate synapse following the spike timing dependent plasticity rule. Simulations using a standard 150 nm CMOS process show the online learning capabilities of the architecture.

Index Terms—VLSI, floating gate, STDP, spiking, synapse

I. INTRODUCTION

Several hardware approaches are being implemented for machine learning, ranging from rate neurons on Von Neumann-Zuse computer architecture [1], [2], FPGA [3] and ASICs [4] from one side, to alternative approaches such as neuromorphic hardware [5]–[7] and quantum computers [8] for quantum machine learning [9], on the other side. Among the most promising for what concerns applications requiring low power consumption or readiness for brain-machine interface, circuits of spiking neurons [10] occupy a prominent role. A spiking neural network (SNN) transmits information along with the network, through spikes in place of finite digits. This coding method mimics that of biological neurons, with great efficiency in energy management [11]. In the past, low power consumption has been addressed by designing essential neurons or synapses [12]–[15] or by elaborating complex networks [16], [17]. We achieved such target by designing a circuit fully compatible with commercial CMOS technology and able to store multiple weights. The device has been designed to store permanently the inter-neuron connections and yet modifying them during its lifetime with learning algorithms, in our case as the Spike Time Dependent Plasticity (STDP). The latter is a famous method used to modify the strength of synapses depending on the relative times of spiking of the involved neurons [18]. The memory element is a floating gate that stores quasi-permanently a charge and it is one of the main candidates for neuromorphic circuits [19]–[21] thanks to the full compatibility with the current CMOS technology. Differently from previously reported floating gate synapses

This work has been partially supported by the EU H2020 projects ICT-STREAMS grant No. 688172 and Neutouch grant No. 813713.

¹Dipartimento di elettronica, informazione e bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, I-20133 Milano, Italy

²Faculty of Technology and Cognitive Interaction Technology Center of Excellence (CITEC), Universität Bielefeld, Universitätsstraße 25, 33615 Bielefeld, Germany

³Istituto di Fotonica e Nanotecnologie, Consiglio Nazionale delle Ricerche, Piazza Leonardo da Vinci 32, I-20133 Milano, Italy

[14], [20]–[22], the stored charge is modified through Fowler-Nordheim tunneling effect [23] avoiding the large current required by hot-carrier injection [24]. In the following, the architecture of the neuron, the synapse and the conditioning circuits are discussed and simulations are shown to validate the design.

II. NEURON

The schematic of the compact and biologically plausible neuron is shown in Fig. 1. The structure was first reported by Sourikopoulos et al. [25] and here is modified to be used with the chosen technology and to add a refractory period, as required by our implementation of the STDP rule. The membrane capacitor C_{mem} is charged by the synaptic current I_{syn} and by a positive feedback current from the sodium channel implemented by M_{Na} and discharged by a negative feedback current from the potassium channel (M_K). The positive feedback, triggered by the commutation of the first inverter (M_{n1}, M_{p1}), creates a spike by clamping the voltage V_{mem} to V_{dd} . The negative feedback, triggered by the second inverter (M_{n2}, M_{p2}), brings V_{mem} to ground for a refractory period of tens of milliseconds. Transistors were sized to operate the neuron with the very low supply voltage of $V_{dd} = 0.4$ V. This gives us the possibility to embrace low currents from the synapses ($I_{syn} \approx 100$ fA – 10 pA) since the voltage threshold of the first inverter is low. The transistor M_{ST} has been added to define a refractory period longer than the time required by the update of the synaptic weight, in order to avoid a spurious injection of charge in C_{mem} , as will be discussed later.

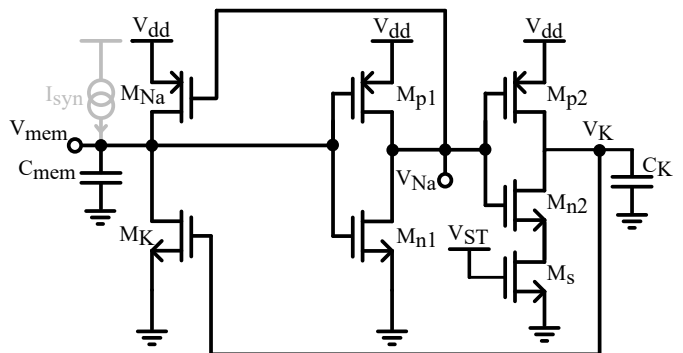


Fig. 1. Schematic of the CMOS neuron based on two inverters that control the positive and the negative feedback with different commutation times.

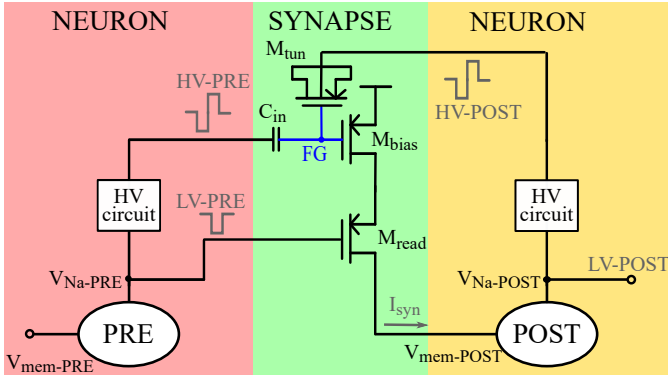


Fig. 2. Scheme of the synaptic architecture. The PRE and POST neurons are connected to the three-transistors synapse with a low voltage path (LV-PRE and $V_{mem-POST}$) for the reading of the synaptic weight and a high voltage path (HV-PRE, HV-POST) for programming the synaptic weight.

III. SYNAPSE

For every chain of two neurons, we can identify the one that is sending the spike, called PRE, and the one that is receiving it, called POST. The synapse is the element responsible for connecting two neurons through a programmable weight, as shown in Figure 2. The design of the synapse followed the optimization of power consumption and area footprint, aiming to replicate the device many times in the neural network. The floating gate, used to store the synaptic weight, is obtained by keeping electrically insulated the polysilicon gate of standard MOS transistors. The Fowler-Nordheim tunneling effect [23] is used for injecting and removing the charge in the floating node. Figure 2 shows the proposed STDP-compatible synapse designed using LFoundry 150 nm CMOS technology that provides transistors operating at 1.8 V, 3.3 V, 5 V along with high voltage LDMOS transistors. The synapse uses four elements:

- A thick oxide P-type 3.3V transistor M_{tun} in a capacitor configuration. It is designed to achieve tunneling between the substrate (N-well) and the gate, allowing the modification of the charge in the floating node. The need for long time retention prevented the use of a 1.8 V transistor, since its thin oxide (< 3 nm) has a significant tunneling rate even at low voltages.
- A MIM capacitor C_{in} that connects the floating node to the previous PRE neuron.
- A thick oxide P-type 3.3V transistor M_{bias} . It converts the charge of the floating node into a current injected in the POST neuron.
- A P-type 1.8V transistor M_{read} used as a switch to connect the synapse at the POST neuron only during the reading phase.

The synapse has two operation modes. The first one is the *reading phase* of the synaptic weight. When a PRE neuron's spike arrives, the node LV-PRE signal is decreased from 0.4 V to 0 V for one millisecond, switching on the transistor M_{read} and leading to an injection of a current I_{syn} into the POST neuron. The current value, i.e. the synaptic weight, is controlled by M_{bias} that is biased by the floating gate node.

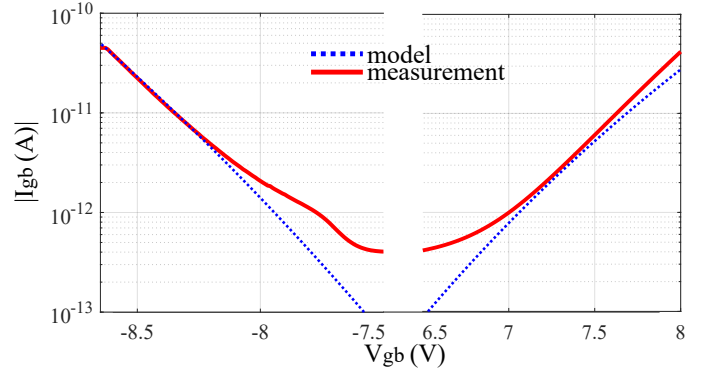


Fig. 3. Tunneling current between gate and bulk of a 3.3 V device in 150 nm CMOS technology.

The second operation mode is the *programming phase*, used for updating the synaptic weight accordingly to the STDP rule. Once a neuron's spike is sensed, a high voltage signal is sent to all the synapses connected to the neuron. The synapses connected at the input of the neuron receive the high voltage signal at the substrate terminal of the M_{tun} transistor (HV-POST in Fig. 2). On the contrary, the synapses at the output of the neuron receive the same high voltage signal at the capacitor C_{in} (HV-PRE). In this way, each synapse receives two high voltages, one coming from the neuron PRE and one from the neuron POST, whose temporal combination allows a positive or a negative tunneling through the gate oxide of M_{tun} .

IV. STDP CIRCUITS

Figure 3 shows the experimental tunneling current of a 3.3 V pMOS transistor. Data are well fitted by the empirical model reported in [24] and suggest a voltage higher than 8 V to achieve a significant current in the floating gate node. Around 6 V the data vary from the model because of stray currents in the setup.

In order to obtain the STDP learning rule, we generate the HV voltages with a negative and a positive swing of ± 4.5 V, as shown in Fig. 4. The time separation between the HV-PRE and HV-POST signals defines the behavior of the synapse. If the two neurons spike far in time (≥ 8 ms), the maximum voltage across the oxide of M_{tun} is ± 4.5 V resulting in a negligible tunneling current. When the PRE neuron's spike arrives less than 8 ms after the PRE neuron's spike (Fig. 4, left), the overall voltage across M_{tun} , given by difference between HV-PRE and HV-POST, reaches 9 V triggering a tunneling current that reduces the voltage of the floating node and increases the synaptic weight. On the contrary, if the PRE neuron's spike comes less than 8 ms after the POST neuron's one (Fig. 4, right), the voltage across the M_{tun} reaches -9 V producing a tunneling current that reduces the synaptic weight.

The HV-PRE signal is generated after the reading phase to avoid cross-talk between the programming and the reading parts. However, a HV-POST signal could occur during the reading phase causing a change of the synaptic current due to the capacitive coupling of HV-POST and the floating node.

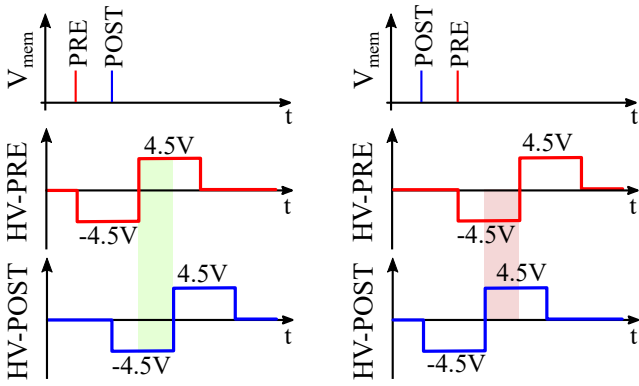


Fig. 4. Timing waveforms of the membrane voltage of the PRE and POST neurons and of the HV signals applied to the synapse in the case of a potentiation (left) and in the case of a depression (right).

In order to prevent a spurious charging of the membrane capacitor, the refractory period of the neuron is longer than the duration of the HV-POST signal.

The high voltages involved require a conditioning circuit to shift the 0.4 V at the output of the neuron to ± 4.5 V. The designed circuit, included in each neuron and shared by all the synapses connected to it, is shown in Figure 5.

The first stage is a 0.8 V digital circuit, responsible for creating the right timing of the positive and negative signals starting from the neuron's spike that is detected by monitoring the V_{Na} voltage. A temporal overlap of these two signals could result in a conductive path from the high voltage power supply to ground in the HV output stage, causing strong power dissipation. To avoid it, the timing circuit has been realized with edge detectors that activate the second rectangular signal only when the first one is in the falling edge.

The second stage shifts the 0.8 V digital signals to voltages high enough to control the HV-MOSFET of the output stage. A standard digital voltage shifter [26] was implemented using 5 V transistors.

Finally, the signals generated by the voltage shifters are fed to the output stage. The HV voltage is obtained using two LDMOS transistors, HV-PMOS and HV-NMOS, able to stand up to 40 V voltage difference between the drain and the source. There are three different configurations of the output stage controlled by the outputs of the voltage shifters (Fig. 5, right):

- IDLE: When no spikes are sensed the floating gate shouldn't experience any change in the stored charge. This is assured by an output voltage at 0 V, obtained by keeping the HV-NMOS conductive (G_{NMOS} high) with the source (S_{NMOS}) at 0 V;
- DW: After $\approx 100 \mu s$ from a spike, the HV output signal is moved to -4.5 V, half of the voltage needed for tunneling, by forcing the source of HV-NMOS at -4.5 V and the gate-source voltage above the threshold voltage;
- UP: When the negative signal is terminated, the positive part is triggered. The HV-NMOS is switched off

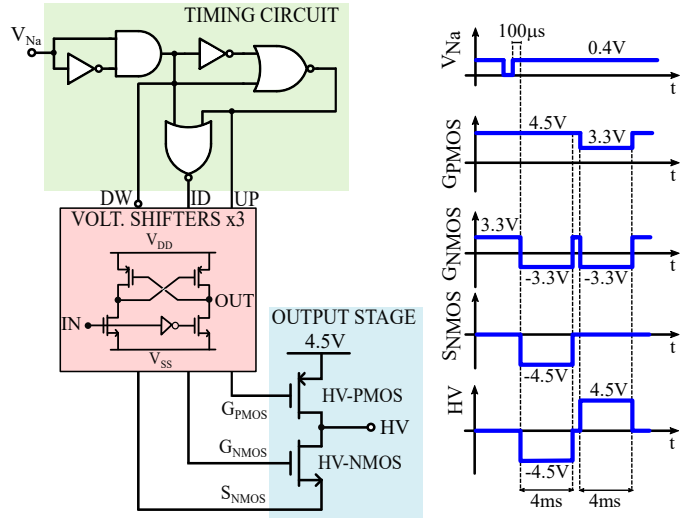


Fig. 5. Left: conditioning circuit to convert a low voltage spike in the high voltage symmetric HV signal. Right: timing diagram of the signals used to create the high voltage HV applied to the synapses.

($S_{NMOS} = 0$ V and $G_{NMOS} = -3.3$ V) and the HV-PMOS is switched on ($G_{PMOS} = 3.3$ V), shunting the output to 4.5 V.

The duration of the DW and UP signals is 4 ms. They are separated by a short IDLE phase ($10 \mu s$) to avoid a cross-conduction current given by the simultaneous activation of HV-PMOS and HV-NMOS.

V. SIMULATION RESULTS

The STDP learning mechanism was demonstrated by simulations at transistor level of the simple chain neuron - synapse - neuron shown in Fig. 2. To simulate the tunneling current of M_{tun} , we developed a Verilog-A component based on the model reported in [24] with the parameters extracted by the experimental measurements in Fig. 3. To study the learning behavior of the architecture, we stimulated the PRE and POST neurons with two bias current, I_{inPRE} and I_{inPOST} respectively, injected at the V_{mem} node. The simulation has been carried for 0.2 s with $I_{inPRE} = 1$ pA from 20 ms to 200 ms and $I_{inPOST} = 1.5$ pA from 70 ms to 140 ms. The results of the simulation are reported in Fig. 6 where we can identify four different conditions based on the values of I_{inPRE} and I_{inPOST} .

In the first time slot, from 0 s to 20 ms (called IDLE in Figure 6) no input current is injected and no neurons spike. As well no high voltage signals (PRE-HV and POST-HV in Fig. 6b) are generated by the conditioning circuits. The voltage of the floating gate node (Fig. 6c-d) is steady at a predefined value. In this phase, negligible power is consumed.

The ENFORCE phase starts when $I_{inPRE} = 1$ pA is fed to the first neuron. The latter starts to spike and, since it is connected with the second neuron through the synapse, it also begins to charge the membrane capacitance increasing the voltage $V_{mem,POST}$. When a spike of the PRE neuron is able

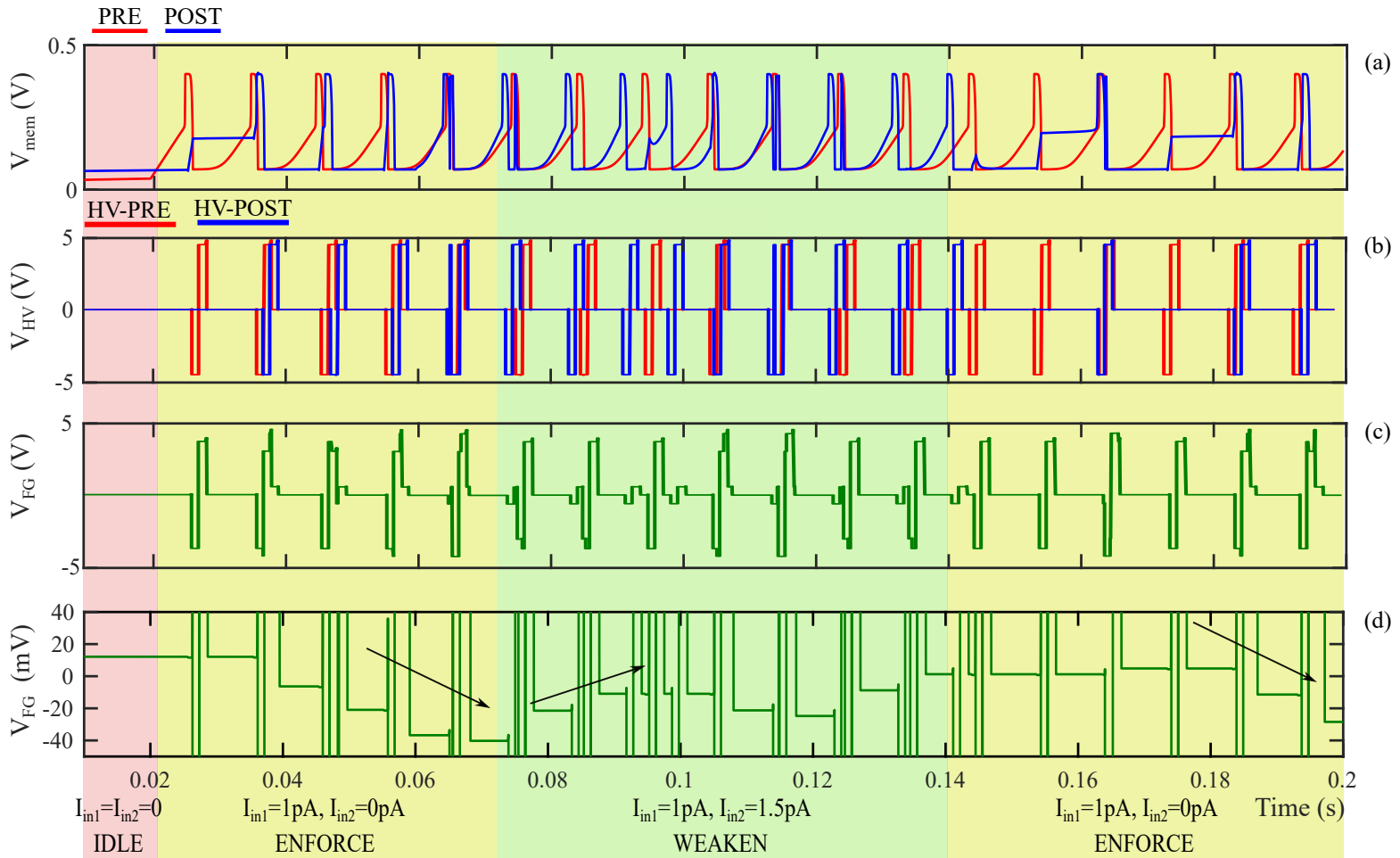


Fig. 6. Membrane voltages (a), HV-PRE and HV-POST signals (b), and voltage of the floating gate on two different scales (c, d) in response to an excitation of the two neurons with different currents: $I_{in1} = 0$ pA, $I_{in2} = 0$ pA (IDLE); $I_{in1} = 1$ pA, $I_{in2} = 0$ pA (ENFORCE); $I_{in1} = 1$ pA, $I_{in2} = 1.5$ pA (WEAKEN). The three arrows highlight the learning trend.

to induce a spike of the POST neuron, the floating gate bias voltage is correctly updated to a lower voltage. This causes a strengthen in the connection between the two neurons so that for each spike of the PRE neuron, we observe a spike of the POST neuron and a further decrease of the floating gate voltage. Starting from $t = 70$ ms the POST neuron is stimulated with a constant current of $I_{inPOST} = 1.5$ pA. Due to this current, the POST neuron has additional spikes uncorrelated with the the spikes of the PRE neuron. In agreement with the STDP rule, the floating gate voltage increases and the connections between the two neurons are weakened. Finally, at $t = 140$ ms the I_{inPOST} is removed. Because of the previous weaken phase, the second neuron initially cannot follow the first neuron. However, the system learns again the connection.

VI. CONCLUSIONS

A VLSI solution for an efficient spiking neural network with permanent memory and STDP is proposed by using a floating gate device obtained in a standard CMOS process. The charge of the floating node is modified by using *Fowler-Nordheim* tunneling only. The energy consumption for each writing (E/Event

TABLE I
COMPARISON BETWEEN FG IMPLEMENTATIONS

	[17] (0.35 μm)	[27] (65 nm)	This (0.15 μm)
E/Spike N	10 pJ	290 fJ (est.)	21 fJ(N), 20 pJ(HV)
E/Event S_{read}	10 pJ	40 fJ (est.)	30 fJ
E/Event S_{write}	4.5 pJ	(volatile)	4 fJ
CMOS-ready	Yes	Yes	Yes
Area (μm^2)	n.a. (N), 133(S)	n.a. (N), 49(S)	168(N), 1240(HV), 27(S)

S_{write}) and reading (E/Event S_{read}) is reduced, as shown in Table I, thanks to the removal of hot carrier injection processes and the introduction of a new architecture, respectively. The low synaptic current provides also the opportunity to improve the consumption of the neuron itself (E/spike N). The need for high voltage is handled through a conditioning circuit that currently is limiting the energy consumption and the area footprint performances. However, the conditioning circuit is shared by all the synapses linked to the same neuron reducing the overall consumption in highly connected neural networks.

REFERENCES

- [1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller, "Playing atari with deep reinforcement learning," *CoRR*, vol. abs/1312.5602, 2013. [Online]. Available: <http://arxiv.org/abs/1312.5602>
- [2] R. Porotti, D. Tamascelli, M. Restelli, and E. Prati, "Coherent transport of quantum states by deep reinforcement learning," *Communications Physics*, vol. 2, no. 1, p. 61, 2019.
- [3] C. Farabet, Y. LeCun, K. Kavukcuoglu, E. Culurciello, B. Martini, P. Akseelrod, and S. Talay, "Large-scale fpga-based convolutional networks," *Scaling up Machine Learning: Parallel and Distributed Approaches*, pp. 399–419, 2011.
- [4] E. Nurvitadhi, D. Sheffield, J. Sim, A. Mishra, G. Venkatesh, and D. Marr, "Accelerating binarized neural networks: Comparison of fpga, cpu, gpu, and asic," in *2016 International Conference on Field-Programmable Technology (FPT)*. IEEE, 2016, pp. 77–84.
- [5] N. Zheng and P. Mazumder, "Hardware-friendly actor-critic reinforcement learning through modulation of spike-timing-dependent plasticity," *IEEE Transactions on Computers*, vol. 66, no. 2, pp. 299–311, 2016.
- [6] E. Prati, "Atomic scale nanoelectronics for quantum neuromorphic devices: comparing different materials," *International Journal of Nanotechnology*, vol. 13, no. 7, pp. 509–523, 2016.
- [7] 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 *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.
- [8] D. Rotta, F. Sebastiano, E. Charbon, and E. Prati, "Quantum information density scaling and qubit operation time constraints of cmos silicon-based quantum computer architectures," *npj Quantum Information*, vol. 3, no. 1, p. 26, 2017.
- [9] M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz, "Quantum-assisted learning of hardware-embedded probabilistic graphical models," *Physical Review X*, vol. 7, no. 4, p. 041052, 2017.
- [10] N. Zheng and P. Mazumder, "Learning in memristor crossbar-based spiking neural networks through modulation of weight-dependent spike-timing-dependent plasticity," *IEEE Transactions on Nanotechnology*, vol. 17, no. 3, pp. 520–532, 2018.
- [11] W. B. Levy and R. A. Baxter, "Energy efficient neural codes," *Neural computation*, vol. 8, no. 3, pp. 531–543, 1996.
- [12] J. M. Cruz-Albrecht, M. W. Yung, and N. Srinivasa, "Energy-efficient neuron, synapse and STDP integrated circuits," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 6, no. 3, pp. 246–256, 2012.
- [13] J. L. Molin, A. Eisape, C. S. Thakur, V. Varghese, C. Brandli, and R. Etienne-Cummings, "Low-power, low-mismatch, highly-dense array of vlsi mihalas-niebur neurons," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2017, pp. 1–4.
- [14] V. Kornijcuk, H. Lim, I. Kim, J. K. Park, W. S. Lee, J. H. Choi, B. J. Choi, and D. S. Jeong, "Scalable excitatory synaptic circuit design using floating gate based leaky integrators," *Scientific Reports*, vol. 7, no. 1, pp. 1–13, 2017. [Online]. Available: <http://dx.doi.org/10.1038/s41598-017-17889-8>
- [15] W. Wang, G. Pedretti, V. Milo, R. Carboni, A. Calderoni, N. Ramaswamy, A. S. Spinelli, and D. Ielmini, "Learning of spatiotemporal patterns in a spiking neural network with resistive switching synapses," *Science Advances*, vol. 4, no. 9, 2018. [Online]. Available: <https://advances.sciencemag.org/content/4/9/eaat4752>
- [16] 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, 2014.
- [17] S. Brink, S. Nease, P. Hasler, S. Ramakrishnan, R. Wunderlich, A. Basu, and B. Degnan, "A learning-enabled neuron array ic based upon transistor channel models of biological phenomena," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 7, no. 1, pp. 71–81, Feb 2013.
- [18] M. R. Azghadi, N. Iannella, S. F. Al-Sarawi, G. Indiveri, and D. Abbott, "Spike-based synaptic plasticity in silicon: Design, implementation, application, and challenges," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 717–737, 2014.
- [19] A. Basu, J. Acharya, T. Karnik, H. Liu, H. Li, J. S. Seo, and C. Song, "Low-Power, Adaptive Neuromorphic Systems: Recent Progress and Future Directions," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 1, pp. 6–27, 2018.
- [20] S. Ramakrishnan, P. Hasler, and C. Gordon, "Floating gate synapses with spike time dependent plasticity," in *2010 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2010, pp. 369–372.
- [21] M. Pankaala, M. Laiho, and P. Hasler, "Compact floating-gate learning array with STDP," *Proceedings of the International Joint Conference on Neural Networks*, pp. 2409–2415, 2009.
- [22] S. Nease and E. Chicca, "Floating-gate-based intrinsic plasticity with low-voltage rate control," *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2016-July, pp. 2507–2510, 2016.
- [23] M. Lenzlinger and E. H. Snow, "Fowler-nordheim tunneling into thermally grown SiO_2 ," *Journal of Applied Physics*, vol. 40, no. 1, pp. 278–283, 1969.
- [24] K. Rahimi, C. Diorio, C. Hernandez, and M. D. Brockhausen, "A simulation model for floating-gate mos synapse transistors," in *2002 IEEE International Symposium on Circuits and Systems. Proceedings (Cat. No. 02CH37353)*, vol. 2. IEEE, 2002, pp. II–II.
- [25] I. Sourikopoulos, S. Hedayat, C. Loyez, F. Danneville, V. Hoel, E. Mercier, and A. Cappy, "A 4-fJ/spike artificial neuron in 65 nm CMOS technology," *Frontiers in Neuroscience*, vol. 11, no. MAR, pp. 1–14, 2017.
- [26] S. Hosseini, M. Saberi, and R. Lotfi, "A low-power subthreshold to above-threshold voltage level shifter," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 61, pp. 753–757, 10 2014.
- [27] V. Kornijcuk, H. Lim, J. Y. Seok, G. Kim, S. K. Kim, I. Kim, B. J. Choi, and D. S. Jeong, "Leaky integrate-and-fire neuron circuit based on floating-gate integrator," *Frontiers in Neuroscience*, vol. 10, p. 212, 2016. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fnins.2016.00212>