Stochastic neuron design using Conductive Bridge RAM

Giorgio Palma*, Manan Suri†, Damien Querlioz‡, Elisa Vianello*, Barbara De Salvo*

*CEA, LETI, MINATEC Campus, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France.
†Institut d’Electronique Fondamentale, Univ. Paris-Sud, CNRS, 91405, Orsay, France.
‡Institut d’Electronique Fondamentale, Univ. Paris-Sud, CNRS, 91405, Orsay, France.

Abstract—We present an original methodology to design hybrid neuron circuits (CMOS + non volatile resistive memory) with stochastic firing behaviour. In order to implement stochastic firing, we exploit unavoidable intrinsic variability occurring in emerging non-volatile resistive memory technologies. In particular, we use the variability on the ‘time-to-set’ (tₜₚ) and ‘off-state resistance’ (Rₜₒₜ) of Ag/GeS₂ based Conductive Bridge (CBRAM) memory devices. We propose a circuit and a novel self-programming technique for using CBRAM devices inside standard Integrate and Fire neurons. Our proposed solution is extremely compact with an additional area overhead of 1R-3T. The additional energy consumption to implement stochasticity in Integrate and Fire neurons is dominated by the CBRAM set-process. These results highlight the benefits of novel non memory technologies, whose impact may go far beyond traditional memory markets.

I. INTRODUCTION

Neuromorphic (or ‘neuromorphic’) hardware research has gained a lot of importance in recent years due to its promising low-power, fault-tolerant, and ultra-adaptive computing paradigms [1], [2], [3], [4], [5]. Neuromorphic computing is usually accomplished with deterministic devices and circuits. However, literature in the fields of neural networks [6], [7] and of biology [8] suggests that in many situations, actually providing a certain degree of stochastic, noisy or probabilistic behavior in their building blocks may enhance the capability and stability of neuroinspired systems. Some kind of neural networks even fundamentally rely on stochastic neurons, like Boltzmann machines [9], [10]. Finally, stochastic neurons may perform signal processing in extremely noisy environments using a phenomenon known as ‘stochastic resonance’ [11], [12].

In neuromorphic hardware, providing stochastic behavior to neurons using pseudo-random number generators will lead to significant overheads. This explains interest in developing silicon neurons with an intrinsic stochastic behavior, but which may be controlled. In previous works, different techniques to implement controlled stochasticity in hardware neural networks have proposed. It is possible to exploit the thermal noise in the CMOS but this may lead to silicon overheads and unwanted correlations [6]. Other techniques exploit CMOS circuits with using noise but have significant area overhead [13], or the noise of photons with photodetectors [14] or even special kinds of ‘noisy transistors’ [15]. Finally it was proposed to use fundamentally probabilistic nanodevices like single electron transistors [16], but which might suffer from poor CMOS compatibility and room temperature operation. In this paper, we describe an original circuit and methodology to design a neuron with stochastic firing behavior exploiting certain physical effects of emerging non-volatile resistive memory technology devices such as Conductive Bridge memory (CBRAM). There are significant advantages of our approach because of the easiness of fabrication in the Back-End-Of-Line (BEOL), CMOS compatibility [17], predicted scalability to sub-20 nm [18] and low programming voltages of CBRAM memory devices [19].

The remainder of this paper is organized as follows: in Section II, we describe the structure and the working principle of our CBRAM devices, and the stochastic effects which we exploit for designing the non-deterministic firing neuron. Section III, describes the basic concept and an example of simple circuit for obtaining a stochastic neuron. Section IV discusses transient simulations that we performed on a basic circuit, which validates our concept.

II. CBRAM TECHNOLOGY

A. Structure and working principle

Fig. 1 shows a TEM image of the CBRAM device structure used in this work. A Tungsten (W) plug, typically used as interconnect between two metal levels, is used as bottom electrode for the CBRAM. The solid electrolyte consists of a 30 nm thick GeS₂ layer deposited by radio frequency physical vapour deposition (RF-PVD). A 3 nm Ag layer is dissolved into the GeS₂ using the photo-diffusion process [20]. Then a 2nd layer of Ag is deposited to act as top electrode.
CBRAM working principle relies on the reversible formation of a conductive filament (CF) through a solid electrolyte that results in transition to low and high resistance, respectively, which are referred to as set and reset processes (Fig. 1) [21]. During the set process a positive voltage is applied to the anode which oxidizes, generating Ag$^+$ ions. The latter, under the influence of the electric field, migrate by hopping to the W cathode where they are reduced and nucleate, building-up an Ag-rich CF. Upon reversal of voltage polarity, besides an electronic current flowing in the CF, an electrochemical current gives rise to Ag$^+$ ions, inducing a collapse of the filament radius resetting the system to the high resistance state [22].

B. Stochastic effects

By cycling many times our devices a statistical distribution of the high resistive state ($R_{\text{Off}}$) was obtained. Dispersion in $R_{\text{Off}}$ may be interpreted in terms of stochastic breaking of the filament during the reset process, due to the unavoidable defects close to the filament which act as preferential sites for dissolution. In previous work [23] we showed, with the help of modeling, that a distribution in $R_{\text{Off}}$ leads to a spread in others physical quantities like the left-over filament height ($h$) and the $t_{\text{set}}$. In this work we push further the analysis by matching the modeled $t_{\text{set}}$ distribution with experimental data. In particular, we characterized the kinetic of the set operation by pulse measurements. Fig. 2 inset shows an example of the oscilloscope trace for the evolution of voltage drop across the cell ($V_c$) during a set pulse. Initially, the cell is in the high resistive state ($R_{\text{Off}} \approx 10^6 \, \Omega$) and most of the applied voltage drops on the cell. Then at time $t_{\text{set}}$ an abrupt decrease of $V_c$ is observed, revealing a sudden drop of the cell resistance corresponding to the switching from high to low resistive state. Starting from some of the measured values of $R_{\text{Off}}$ (Fig. 2(a)) we collected the spread in $t_{\text{set}}$ when the applied pulses were $V_a=3 \, \text{V}$ and $t_{\text{pulse}}=5 \, \mu\text{s}$ (Fig. 2(b)). The dotted line in Fig. 2(b), shows the simulated values of $t_{\text{set}}$. To obtain the simulated curve of $t_{\text{set}}$, first the distribution of $h$ was calculated using [24], [25]:

$$R_{\text{Off}} = \frac{\rho_{\text{on}} h + \rho_{\text{off}} (L-h)}{\pi r^2}$$  \hspace{1cm} (1)

where $\rho_{\text{on}}$ is the resistivity of the Ag-rich nanofilament, $\rho_{\text{off}}$ is the resistivity of the Ge$_2$S$_2$, $L$ is the chalcogenide thickness and $r$ is the conductive filament radius, then $t_{\text{set}}$ using:

$$t_{\text{set}} = \frac{L - h}{\alpha q E} \exp \left( \frac{-E}{k_B T} \right) \frac{1}{\sinh \left( \alpha q V - \Delta \right)}$$  \hspace{1cm} (2)

where $q$ is the elementary charge, $v_h$ is a fitting parameter for the vertical evolution velocity, $E_A$ is the activation energy, $k_B$ is the Boltzmann constant, $T$ is the temperature (300 K), $\alpha$ and $\Delta$ are fitting parameters to take into account vertical electric field dependency and the overpotential that controls the kinetic of the cathodic reaction respectively (Table I). In the following section we show how the spread in $t_{\text{set}}$ can be used to make the firing of an Integrate and Fire neuron non-deterministic.

III. STOCHASTIC NEURON DESIGN

A. Integrate and Fire Neuron

The complexity of a neuron circuit depends on the overall functionality of the neural network and of the chosen biological

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_h$</td>
<td>2 m/s</td>
<td>$E_A$</td>
<td>0.35 eV</td>
</tr>
<tr>
<td>$\rho_{\text{on}}$</td>
<td>$2.3 \times 10^{-6} , \Omega , \text{m}$</td>
<td>$\rho_{\text{off}}$</td>
<td>$10^{-3} , \Omega , \text{m}$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.08</td>
<td>$\Delta$</td>
<td>0.15 V</td>
</tr>
<tr>
<td>$r$</td>
<td>2.2 nm</td>
<td>$L$</td>
<td>30 nm</td>
</tr>
</tbody>
</table>
models. For our purpose of concept validation, we chose one of the simplest, the Integrate and Fire neuron model. Fig. 3(a) shows the concept of a simple Integrate and Fire neuron model. It constantly sums (integrates) the incoming synaptic-inputs or currents (excitatory and inhibitory) inside the neuron integration block using a capacitor. More advanced designs also work with this principle [26]. This integration leads to an increase in the membrane potential of the neuron $V_{\text{mem}}$. When the membrane potential reaches a certain threshold value $V_{\text{th}}$, the neuron generates an output spike (electrical signal). After the neuron has fired the membrane potential goes back to a resting value (initial state), through discharging of the capacitor $C_{\text{mem}}$. Usually, the output firing activity of a Integrate and Fire neuron is deterministic because the neuron fires every time the membrane potential reaches a defined threshold value.

B. Stochastic-Integrate and Fire principle and circuit

To introduce non-deterministic or stochastic behavior in Integrate and Fire neuron, we propose to connect a CBRAM device to the capacitor $C_{\text{mem}}$, such that $C_{\text{mem}}$ could only discharge through the CBRAM device by switching it to the low-resistive state (Fig. 3(b)). The anode of the CBRAM and the $V_{\text{mem}}$ net of the capacitor should be connected. The duration for which current can flow through the low-resistive CBRAM device can be controlled using a transistor. In such a configuration, the spread on the $t_{\text{set}}$ of the CBRAM would translate to a spread on the discharge-time ($t_{\text{dis}}$) of the capacitor. For consecutive neuron spikes, this would lead different initial state of $C_{\text{mem}}$, thus making the firing of the neuron stochastic. Fig. 4 illustrates conceptually the impact of four different values of $t_{\text{set}}$ (keeping constant pre-synaptic weights), on the inter-spike interval. In case (a), $t_{\text{set}}$ is very long thus the capacitor has a very weak discharge. As a consequence just few additional incoming pre-neuron spikes are required to charge back the $V_{\text{mem}}$ to the level of $V_{\text{th}}$, thus leading to an output pattern with the shortest inter-spike interval. In case (b), $t_{\text{set}}$ was the shortest, and hence the capacitor discharged the most.

Thus for this case, more incoming pre-neuron spikes are needed to recharge $V_{\text{mem}}$. Case (c) represents a deterministic Integrate and Fire situation with full $V_{\text{mem}}$ discharge. Finally, case (d) depicts a situation with different $t_{\text{set}}$ durations for consecutive output spikes. It is a possible representation of neuron inter-spike intervals for a random sequence of $t_{\text{set}}$ values that can be obtained by cycling the CBRAM device multiple times.

The circuit equivalent of the Stochastic-Integrate and Fire neuron concept shown in Fig. 3(b) is presented in Fig. 5. It consists of a current-source to simulate input currents coming from synapses and pre-neurons, a capacitor $C_{\text{mem}}$ to integrate the current and build up the neuron membrane-voltage $V_{\text{mem}}$, a nMOS transistor $M1$ to perform set operation, two nMOS transistors $M2$ and $M3$ to perform the reset operation, a comparator block, a spike-generation block, a delay-element $\Delta t$ and a CBRAM device. The delay element is used to perform the reset operation of the CBRAM device at the end of each neuron spike.

In Fig. 5, initially the CBRAM is in high-resistive state. As incoming pre-synaptic current is accumulated in $C_{\text{mem}}$, $V_{\text{mem}}$ would constantly build up at the anode of the CBRAM. During this time $M1$, $M2$ and $M3$ are off. When the neuron spikes, the spike-generation block will generate an output-spike and two additional pulsed-signals ($S1$, $S2$) going to $M1$ and $\Delta t$ respectively. $S1$ acts as a gating signal to turn on $M1$. $V_{\text{mem}}$ build-up and switching on of $M1$ will enable set-operation of the CBRAM since a positive voltage drop is established between the anode and the cathode. However during the set-operation, $M2$ and $M3$ are not turned on, as $\Delta t$ delays the signal $S2$. 

Fig. 6. Circuit used to demonstrate the concept of a S-IF effect when the CBRAM is in the set state.
At the end of the set-operation, the signal S2 will turn on M2 and M3 thus building up the voltage at the cathode to switch the CBRAM to the off-state (reset). Thus, before the next consecutive neuron spikes the CBRAM device is automatically reset and reprogrammed to a different initial R value. Note that the flow of current through the CBRAM, during the set-operation, leads to a discharge of the capacitor C_mem thus decreasing the membrane voltage V_mem. The amount of decrease in V_mem can be estimated by calculating the total duration (t_disc) for which current flows through the switched CBRAM. t_disc is the difference of the pulse-width of the signal S1 and the t_set (inset of Fig. 2). Depending on the value of t_set every time the neuron spikes, different amount of C_mem discharge will occur. Thus, in between any two firing cycles, the neuron may require different amount of incoming current to charge V_mem to the level of V_th.

IV. RESULTS AND DISCUSSION

A. Set- and reset- operation

We performed SPICE transient simulation, with Eldo simulator, to validate the proposed concept using a simplified circuit shown in Fig. 6. Transistors and capacitors sizing were not optimized with respect to a real implementation, but to give a simple proof-of-concept. Fig. 7(a) shows a simulated train of incoming pulses (excitatory currents) and the corresponding evolution of the V_mem (Fig. 7(b)) between two consecutive neuron spike-cycles. When V_mem reaches a threshold voltage V_th (V_th ≃ 3.5 V in our simulation), the CBRAM device undergoes set-operation, and C_mem begins to discharge. Fig. 7(b) shows the discharging and re-charging of C_mem for four different simulated values of t_set (in the range 300 ns - 600 ns). Fig. 7(c), shows the expected output of the neuron. Note that different number of incoming pulses are required to reach the neuron firing threshold again, since the initial V_mem value is dominated by the stochasticity in t_set. Five additional incoming pulses are needed to reach the threshold for the shortest value of t_set (300 ns). Fig. 8 shows the zoomed version of C_mem discharging for the the different simulations shown in Fig. 7. Note that the longest t_set (600 ns) corresponds to the least amount of C_mem discharge, and vice-versa. To simulate the reset operation, a pulse of 45 ns with an amplitude of 3 V was applied at M2 and M3, while keeping M1 off. Such high voltage on M3 is required to build up a voltage on V_cathode. Fig. 9 shows the time evolution of V_cathode and V_mem when the initial value of V_mem was generated by a t_set of 300 ns for two different width of M3. The actual voltage drop on the CBRAM can be increased increasing the size of the nMOS as shown in Fig. 9. Moreover, during the reset, an additional discharge of V_mem is possible depending on the size of M3, since M2, that is directly connected to V_mem, is turned on by S2 (Fig. 10(a)).

B. Parameter constraints

Due to the intrinsic physics of CBRAM device, some constraints in implementing the proposed circuit should be considered. In particular, V_th has to be greater than the minimum value of the voltage-drop required to set the CBRAM device for a given pulse-width. The amplitude of S1 should be sufficient
devoted to reset the CBRAM is negligible. For a real system, set low resistive value thus reducing \( I_{\text{set}} \) application the ratio \( R_{\text{Off}} \) (i.e. \( V_{\text{set}} \) using: \( E_{\text{set}} \)).

The stochastic response of the neuron can be controlled. The extra-energy consumption is dependent on the final CBRAM resistance obtained after the set-operation. The extra-energy consumption is dependent based on the final CBRAM resistance obtained after the set-operation.

C. Energy consumption

For the proposed S-IF, additional energy consumption per spiking cycling of the neuron will be devoted to perform set and reset operation. The extra-energy consumption is dependent on the ratio \( R_{\text{Off}}/R_{\text{On}} \); in particular on \( R_{\text{On}} \) since hundreds of \( \mu \text{A} \) can flow before \( M1 \) would be turned off, if the low resistance state is \( \approx 10^4 \Omega \), thus raising the power consumption. We estimated the energy consumption during the set operation using: \( E_{\text{set}}=V_{\text{set}} I_{\text{set}} t_{\text{set}} \). In our simulations we used \( V_{\text{set}}=3.5 \text{ V} \) (i.e. \( V_{\text{th}} \)), \( I_{\text{set}}=350 \mu \text{A} \), \( t_{\text{set}} \) in a range between 300 ns and 600 ns that gives a \( E_{\text{set}} \) energy mean value of 55 nJ. The energy devoted to reset the CBRAM is negligible. For a real system, \( E_{\text{set}} \) can be strongly reduced increasing the resistance of the low resistive value thus reducing \( t_{\text{set}} \), since for the proposed application the ratio \( R_{\text{Off}}/R_{\text{On}} \) is not a major constraint.

V. Conclusions

In this paper, we showed how CBRAM used in an unexpected fashion may allow designing stochastic neurons with low area overheads. We described how CBRAM physics naturally leads to a stochastic behavior in the programming time \( t_{\text{set}} \) which may be exploited in a circuit. SPICE simulations validate the concept on a simple Integrate and Fire neuron. The concept could be extended to more complex neuron designs like [26] and [28], paving the way for the fabrication of complex neuromorphic networks. Other emerging memory technologies might be used for the same purpose provided a certain range of stochasticity in \( t_{\text{set}} \) as reported in [29] and [30]. These results highlight the benefits of novel non memory technologies, whose stochasticity may go beyond traditional memory markets.

Acknowledgment

The PhD of M. Suri is co-financed by DGA-France. The authors would like to thank ALTIS semiconductor for providing the CBRAM devices for this study.

References


