Graphene Nanoribbon-based Synapses with Versatile Plasticity

H. Wang, N. Cucu Laurenciu, Y. Jiang, S.D. Cotofana
Computer Engineering Laboratory,
Delft University of Technology, The Netherlands.
{H.Wang-13, N.CucuLaurenciu, Yande.Jiang, S.D.Cotofana}@tudelft.nl

Abstract—Designing and implementing artificial systems that can be interfaced with the human brain or that can provide computational ability akin to brain’s processing information efficient style is crucial for understanding human brain fundamental operating principles and to unleashing the full potential of brain-inspired computing. As basic neural network components, responsible for information transfer between neurons, artificial synapses able to emulate analog biological synaptic behaviour are of particular interest. State of the art CMOS and memristor-based synapses suffer from scalability drawbacks (large energy consumption and area footprint), variability-induced instability, and are not bio-compatible. In this paper, we propose a generic Graphene Nanoribbon (GNR) based synapse structure and demonstrate that by changing GNR geometry and external bias voltages it can emulate different synaptic plasticity behaviours, i.e., Spike Timing Dependent Plasticity and Long-Term Depression and Potentiation, and that both excitatory and inhibitory synaptic behavior can be obtained with the same GNR geometry. To demonstrate biologically plausible operation, we make use of low voltage bias, i.e., 0.1 V, 0.2 V, and consider inputs consistent with measured brain synapses data, i.e., −50 mV to 50 mV pre- and post-synaptic spikes voltage range, and −60 ms to 60 ms time range. The simulations indicate that by changing the GNR shape we can enrich the plasticity behaviour (potentially beyond the considered cases) and the plasticity change of 100% provided by natural synapses can be achieved. Our investigation clearly suggests that the proposed GNR synapse structure is a promising candidate for large-scale neuromorphic systems integration, which might potentially bring novel insight on brain neurophysiology, as it requires a small footprint, is energy effective, biocompatible, and versatile from the synaptic behaviour point of view.

Index Terms—Neuromorphic Computing, STDP, Artificial Synapse, Graphene, GNR.

I. INTRODUCTION

The human brain, comprising approximately 86 billion neurons connected through trillions of synapses, is the natural high performance computing system. Its unique capabilities, e.g., low power consumption, robustness, massively parallel information processing, suitability for complex tasks, inspired Carver Mead in the late 1980s [1] in coining a disruptive computing paradigm, the Neuromorphic Computing (NC). Over the last decades NC gained substantial momentum, provided valuable inside into brain’s complex functionality, novel brain-inspired computation paradigms have been introduced [2], and biologically-inspired neuromorphic systems fabricated [3].

Synapses are the most ubiquitous neural network components which are ensuring information interchange between neurons. Their essential property is Synaptic Plasticity (SP), manifested either by the strengthening or the weakening of the transmitted signals, is the brain learning and memory enabler. Spike Timing Dependant Plasticity (STDP), which enables synaptic transmission strength changes according to the relative timing of pre- and post-synaptic spikes, Long-Term Potentiation (LTP) and Long-Term Depression (LTD), which are persistent synaptic strengthening and weakening, respectively, are the essential functionalities an artificial synapse ought to be endowed with [4].

Given that at any brain inspired system crux of the matter reside artificial synapses able to emulate the biological ones, their design and fabrication received massive attention. In most of today’s neuromorphic computing systems, artificial synapses are typically implemented using dozens of CMOS devices [5], [6]. However, CMOS technology inadvertently imposes restrictions on both functionality and neuromorphic system implementations, foremost in terms of energy efficiency, scalability, and integration density. Furthermore, CMOS devices cannot truly convey the analog behaviour associated with biological synapses. Alternatively, emerging resistive switching memory devices [7] based synapses have been also proposed [8], [9] and exhibit promising characteristics, e.g., simple (a single or a few memristors) structure, inherently analog conductance, and good scalability potential. However, they suffer from temporal (cycle-to-cycle) and spatial (device-to-device) variability of the resistive state even under the same applied signals, as well as undesired nonlinearities, which may cause the instability of the entire neuromorphic system.

Graphene, one of the prominent post-Si forerunners, owing to its outstanding properties, e.g., fast switching speed, low energy, thermal stability, ultimate thinness, flexibility, and biocompatibility [10], [11], has emerged as a potent material [12] and previous work demonstrated that graphene-based artificial synapse can emulate plasticity. In [13], by changing the back-gate voltage, the authors obtained various synaptic plasticities within the same device. However, quite large back-gate voltages (20 V, 40 V) and input pulses (2 V) are utilized, which are faraway larger than the electric potentials measured in natural neurons, and are negatively affecting the power consumption. Moreover, the obtained synaptic weight change is relatively small (≈ 10%) when compared with (100%) in biological counterparts, and the provided synaptic plasticity is restricted. In [14], a fabricated graphene-based electrochemical synapse is reported, whose conductance is modulated by changing the Li ion concentration between the graphene layers. This synapse enables low-power switching, exhibits low variability, and is potentially suitable for large-scale integration. However, the reported STDP conductance change of ≈ 2% and timing difference around 1000 ms are in
a different range than the one of natural synapses, i.e., ≈100% and 100 ms, respectively.

In this paper, we propose a generic Graphene Nanoribbon (GNR) based synapse structure consisting of a GNR monolayer placed above an insulator and a doped substrate. Two top contacts are utilized to bias the GNR and its conduction modulated by means of electrostatic interaction via top and back gates reflects the synaptic weight. Specifically, we consider 2 fundamental synapse functionalities, i.e., Spike Timing Dependant Plasticity (STDP) and Long-Term Plasticity (LTP), which are known to underlie learning and memory in brain. By carving the GNR synapse geometry and changing the bias and back-gate voltages we successfully emulate: Balanced Hebbian STDP, Potentiation Dominated Hebbian STDP, Potentiation Dominated Anti-Hebbian STDP, and Long-Term Plasticity. Furthermore, we demonstrate that both excitatory and inhibitory synaptic behaviours can be obtained with the same GNR synapse, simply by changing the bias back-gate voltage. To demonstrate biologically plausible operation, we bias the GNRs at low voltage (0.1 V, 0.2 V) and consider inputs consistent with measured brain synapses data, i.e., −50 mV to +50 mV pre- and post-synaptic spikes voltage range, and −60 ms to 60 ms time range.

The simulations indicate that by changing the GNR shape we can enrich the plasticity behaviour (potentially beyond the 4 considered cases) and the plasticity change of 100% provided by natural synapses can be achieved. Our investigation clearly suggests that the proposed GNR synapse structure is a promising candidate for large-scale neuromorphic systems integration, which might potentially bring novel insight on brain neurophysiology, as it is small (e.g., 38 nm²), energy effective, biocompatible, and versatile from the synaptic behaviour point of view.

The remaining of this paper is organized as follows: Section II outlines basic synapse and plasticity concepts, and introduces the proposed GNR-based synapse. In Section III we describe the simulation model employed for simulating the GNR electronic properties, and the overall simulation setup and methodology. In Section IV we present the obtained simulation results, and in Section V we conclude the paper.

II. SYNAPTIC PLASTICITY AND GRAPHENE-BASED SYNAPSE

In this section, we briefly present the synaptic plasticity underlying concepts and then introduce the proposed graphene-based synapse.

To explain the role of a synapse in the neuron information interchange, Figure 1 depicts a very small network composed of two neurons, N_j and N_k, connected via synapses to a third neuron, N_i. Neuron N_i collects input signals from the two pre-synaptic neurons, N_j and N_k, and when their cumulated signals effects exceed a certain neuron-specific firing threshold, neuron N_i generates an output signal (spike) which then propagates through all its terminations. From the synapse perspective, (consider for instance the synapse between neurons N_i and N_j), there are (i) two input spikes: the pre-synaptic spike S_j, which comes from neuron N_j, and the post-synaptic spike S_i, which is generated by neuron N_i, and (ii) one output spike S^out, which will be transmitted to neuron N_i. In general, the synaptic transmission efficiency quantified through the synaptic weights W - is variable, either weakening or strengthening the magnitude of the signals transmitted via the synapse. This property is known as synaptic plasticity and it is believed to hold a crucial role in learning and memory in brain. Spike Timing Dependent Plasticity (STDP) [15] is a widely utilized Hebbian synaptic learning rule, for which the synaptic weight changes based on the relative timing between the pre- and post-synaptic spikes, as follows: (i) when the pre-synaptic spike arrives shortly prior to the post-synaptic one, the synaptic weight increases and this may lead to a persistent weight increase (Long-Term Potentiation (LTP)); otherwise, the synaptic weight decreases and may lead to a persistent weight decrease (Long-Term Depression (LTD)), and (ii) if the pre- and post-synaptic spikes arrive very close to each other, a large synaptic weight change occurs. We denote by ∆W, the synaptic weight change, and by ∆t = t_post − t_pre, the arrival time difference of the pre- and post-synaptic spikes. Figure 2 graphically illustrates ∆W(∆t) according to a biological synapse measured data [16]. Even though the data exhibit stochasticity, a widely accepted interpolating model is the following:

\[
\Delta W(\Delta t) = \begin{cases} 
A_+ \cdot exp(-\Delta t/\tau_+), & \text{for } \Delta t > 0 \\
-A_- \cdot exp(\Delta t/\tau_-), & \text{for } \Delta t < 0, \end{cases}
\]

where A_+ and A_- are parameters determining the magnitude of synaptic potentiation and depression, while \(\tau_+\) and \(\tau_-\) are time constants in the order of 10 ms fitted by experimental data reflecting the temporal range over which the synaptic strengthening and weakening occurs.

Figure 3 illustrates the proposed graphene-based generic synapse structure. It consists of a monolayer Graphene Nanoribbon (GNR) located above an insulating layer and
III. SIMULATION FRAMEWORK

In this section, we present the model we used for deriving the electronic properties of the graphene synapse, and describe the simulation setup and the employed methodology to calculate the synaptic weight change (the GNR conductance change) and to obtain the plasticity behaviour.

A. Simulation Model

For the electronic transport computation, we used an NEGF-based hysteresis-aware simulation model [17]. Tight Binding (TB) Hamiltonian matrix \( H = H_0 + U \) is used in this model to model the interaction between carbon atoms (via \( H_0 \)) and external potentials (via \( U \)). The interaction matrix is calculated as follows:

\[
H_0 = \sum_{i,j} t_{i,j} \mid i \rangle \langle j \mid ,
\]

where \( t_{i,j} = \begin{cases} \tau, & \text{if atoms } i \text{ and } j \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases} \)

We account for the first nearest-neighbor interaction with \( \tau = -2.7 eV \). The potential distribution \( U \) is calculated by solving a 3D Poisson equation self-consistently with finite difference method. As the interface traps cause an equivalent shift of the gate voltage, denoted \( \Delta V_{gs} \), we update the potential profile with \( V_g + \Delta V_{gs} \) while solving the Poisson equation. The interface trap charge can be calculated with an accumulation equation:

\[
Q_{it} = \sum \alpha_i \cdot Q_q \cdot \exp(- (t - t_{arrival}) / t_{trap}),
\]

where \( t_{arrival} \) is the input spike arrival time. The value of \( \alpha \) can be calculated as a function of the interface traps capacitance \( C_{it} \):

\[
\alpha(C_{it}) = \frac{C_{it} \cdot (V_g C_{ox} - Q_q)}{Q_q \cdot (C_{it} + C_{ox})},
\]

where \( V_g \) is the top-gate voltage, \( Q_q \) is quantum capacitor charge, \( C_{ox} \) is the oxide capacitance, and \( C_{it} \) is the interface traps capacitance, expressed as a function of the traps density as \( C_{it} = q^2 \cdot D_{it} \).

The source and drain contacts along the end sides of graphene channel, which sustain the conduction, can be modeled with self-energy matrices \( \Sigma_1 \) and \( \Sigma_2 \), respectively. Thus, the transmission function \( T(E) \) that models the possibility of one electron being transmitted between the source and drain contacts, can be derived as a function of energy:

\[
T(E) = \text{Trace} \left[ \Gamma_1 G_R \Gamma_2 G_R^\dagger \right],
\]

where

\[
G_R(E) = [EI - H - \Sigma_1 - \Sigma_2]^{-1},
\]

\[
\Gamma_{1,2} = i[\Sigma_{1,2} - \Sigma_{1,2}^\dagger].
\]
The current through the graphene channel is calculated by the Landauer formula:

\[
I = \frac{q}{\hbar} \int_{-\infty}^{+\infty} T(E) \cdot (f_0(E - \mu_1) - f_0(E - \mu_2)) \, dE,
\]

where \(f_0(E)\) denotes the Fermi-Dirac distribution function at temperature \(T\), and \(\mu_{1,2}\) represents the source and drain contacts electrochemical potential. A block-by-block algorithm [18] for computing matrix inversion is used to speedup the calculation of \(G_R\). Then the conductance of GNR device is calculated as:

\[
G = \frac{I}{V_d - V_s}.
\]

B. Simulation Setup and Methodology

In order to apply the input spikes to the graphene synapse, we employ a single-input scheme, as exemplified in Figure 4. The signal applied as input to the synapse is computed as a superposition of the pre- and post-synaptic spikes (i.e., the voltage difference between the two spikes). We define \(T_{\text{overlap}}\) as the arrival time of the secondly arriving spike. To perform biologically plausible simulations, we considered data consistent with measured data from brain synapses: \(-50 \text{ mV}\) to \(+50 \text{ mV}\) pre- and post-synaptic spikes voltage range, and \(-60 \text{ ms}\) to \(60 \text{ ms}\) \(\Delta t\) range (which covers the general time range for biological LTP and LTD) [15], [16].

As concerns the GNR, we define its topology in Figure 5. In particular, \(W\) and \(L\) represent the GNR width and length. \(P_{V_g}\) signifies the distance between the top-gate and the drain contact, and \(W_{V_g}\) denotes the width of the top-gate contact. In our simulation, we considered multiple non-rectangular GNRs with different shapes but the same overall \(W = 39a\) and \(L = 28\sqrt{3}a\). For the top-gate contact we set \(P_{V_g} = 8\sqrt{3}a\) and \(W_{V_g} = 6\sqrt{3}a\), where \(a = 0.142\text{ nm}\). Concerning the traps induced hysteresis, we assume a density of interface traps of \(2.5 \times 10^{-12} \text{ cm}^{-2}(eV)^{-1}\), and we set a trapping/detrapping time constant of \(20 \text{ ms}\). Subsequently, we present the overall design and simulation methodology. For a desired plasticity behaviour, we first determine a potentially appropriate GNR geometry and drain-to-source and back-gate voltages. Subsequently, we subject the graphene synapse to a train of spikes applied via the top-gate, one spike for each \(\Delta t\) in the considered range. Corresponding to each input spike, we then measured the synaptic weight change \(\Delta W\) (the difference between the GNR conductance values at two consecutive time moments, i.e., \(T_{\text{overlap}}\) and the immediately previous time moment), and assess its compliance with the desired \(\Delta W(\Delta t)\) plasticity curve. If results are not according with the desired plasticity we change the GNR geometry and bias voltages.

IV. Simulation Results

To evaluate the capabilities of proposed graphene synapse, we target 3 common plasticity types underlying balanced and potentiation dominated learning [19], [20]: Hebbian STDP with balanced LTD and LTP (Figure 6 (a)), LTP-biased Hebbian STDP (Figure 6 (d)), and LTP-biased Anti-Hebbian STDP (Figure 6 (g)).

Figure 6 (b) depicts the GNR synapse shape we obtained for the Hebbian STDP with balanced LTD and LTP scenario, biased at \(V_d = -0.2 \text{ V}\) and back-gate voltage \(V_{\text{back}} = -0.2 \text{ V}\). The simulated synaptic weight change (conductance change) (Figure 6 (c)) indicates a good resemblance with the Hebbian STDP with balanced LTD and LTP weight change trend. In biological models, there is a certain randomness in the synapse reaction. We seek a synaptic reaction tendency closer to the plasticity models. When fitting the simulated conductance change with the canonical model in Equation 1, we obtained \(\tau_+ = 23 \text{ ms}\) and \(\tau_- = 37 \text{ ms}\). Since for a biologically plausible input, we obtain an amplitude of the conductance change around 100%, which is consistent with biological synapse measured data shown in Figure 2, the proposed graphene synapse can enable potentially biologically plausible implementations (artificial synapses which can be interfaced with biological neurons in the context of, e.g., neural prosthetics).

Figure 6 (e) and (h) illustrate the obtained GNR synapse shapes for LTP-biased Hebbian STDP and LTP-biased Anti-Hebbian STDP, respectively. The drain voltage \(V_d\) is set to 0.1 V for both shapes, while the applied back-gate voltage is 0 V and \(-0.5 \text{ V}\), respectively. The simulated synaptic weight change (conductance change) in Figure 6 (f) and (i) is temporally asymmetric, being dominated by (LTP) potentiation for both graphene synapses. When fitted with the model in Equation 1, we obtained \(\tau_+ = 21 \text{ ms}\) and \(\tau_- = 10 \text{ ms}\) for the
A synapse can either be excitatory (i.e., potentiation for pre-before-post synaptic spikes arrival) or inhibitory (i.e., depression for pre-before-post synaptic spikes arrival). Traditionally, 2 artificial synapses are employed, but we are able to obtain both excitatory and inhibitory behaviours with a single synapse, which is beneficial from the area and energy standpoints for large-scale integrations. For instance, the GNR synapse shape illustrated in Figure 6 (b) exhibits an excitatory behaviour but by simply changing the biasing gate voltage $V_{\text{back}}$ from 0.2 V to 0.5 V, while the other GNR applied voltages ($V_d$ and $V_g$) are identical the inhibitory counterpart is obtained, as depicted in Figure 7.

Apart from STDP, Long-Term Plasticity is a fundamental synaptic functionality, dominant for how the brain stores information, which is obtained when applying an identical spike consecutively. In our experiments we considered the GNR synapse shape from Figure 6 (h) and applied 50 mV input spikes with an intermission period between the spikes of 1 s. For each spike, we measured the GNR drain to source current, which represents the current of the output spike generated.
Timing Dependent Plasticity and Long-Term Plasticity, which fully emulated two fundamental synapse functionalities: Spike excitatory and inhibitory synaptic behaviours. We successfully emulated two fundamental synapse functionalities: Spike Timing Dependent Plasticity and Long-Term Plasticity, which underlie learning and memory in brain. We demonstrated that the plasticity can be tuned by changing the GNR synapse shape and topology, thus even though only three STDP types have been considered the presented GNR synapse design methodology is generic and can be utilized for the design of synapses able to provide other plasticity types. All the simulations have been performed with biologically plausible settings, which indicates that GNR based biologically compatible synapses can be designed, fabricated, and eventually interfaced with biological neurons. The proposed synapses have a small area footprint (order of 10 nm²) and operate at low operating voltages (order of 100 mV), which makes them strong candidates for the potential implementation of large-scale energy effective artificial neural networks.

References