A graphene-based synapse for photonic neural networks

Bicky A. Marquez, Hugh Morison, Zhimu Guo, Matthew Filipovich, and Bhavin J. Shastri
Department of Physics Engineering Physics & Astronomy, Queen's University, Kingston, ON K7L 3N6, Canada
bama@queensu.ca

***Abstract*—We propose a graphene-based synapse model that consists of an electro-absorption modulator embedded in a microring resonator that has the potential to operate >100 GHz. We found that a maximum number of 55000 synaptic weights can be represented on our photonic device.**

Keywords— graphene-based modulators, photonic neural networks, silicon photonics

I. Introduction

The representation of digital information in analog hardware is marking a milestone in the history of information processing. The benefits of analog architectures for accellerating artificial intelligence (AI) applications are considerable. Among these advantages, we highlight the possibility of designing brain-inspired physical circuits, with which we can solve AI tasks more efficiently. The digital language and the von Neumann architecture have proven inefficient in solving problems that are suited for parallel processing. For example, the training for Google's state-of-the-art large-scale language model BERT (110M parameters) requires 4 days using 16 TPUv2, which corresponds to 12,041.51W and 1,438 lbs of CO_2 emissions (equivalent to a trans-American flight) [1]. Such limitations can be overcome through the implementation of photonic analog processing on integrated circuits with small footprint, high-speed and low power consumption.

Multiple parallel processing can be achieved by using a wavelength division multiplexed (WDM) architecture, in conjunction with banks of tunable silicon microring resonators (MRRs) cascaded in series, whose demonstrated fan-in, high-gain optical-to-optical nonlinearity, and indefinite cascadability [2] make them ideal to recreate on-chip synaptic weights with small footprint. In previous approaches, synaptic weights were defined by thermally tuning the MRR resonance frequency; however it has been also associated with spectral cross-talk when cascading many MRRs in series, and a problem for repeatability and control [3].

In this work, we propose a graphene-based synapse model as a core element of a photonic neural network that can overcome problems associated to spectral cross-talk. Combining silicon photonics and waveguide-integrated graphene, we design an efficient device for analog neuromorphic computing, which has the properties of being fully-tunable and high-speed. Tunability is guaranteed through the incorporation of an electro-absorption modulator [4] on an MRR that has the potential to operate at >100 GHz. Based on numerical simulations, our study compares the graphene-based photonic synapse with state-of-the-art synapses [5, 6], and we determine its possible advantages for analog AI hardware.

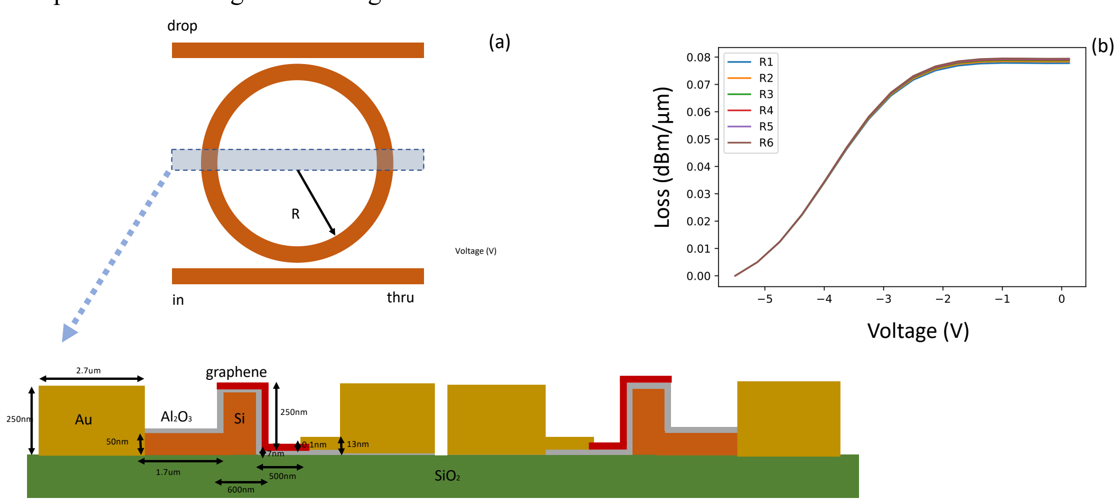

Fig. 1. Schematic diagram of the (a) add-drop MRR with the graphene-based electro-absoption modulator; and (b) the loss vs voltage for the graphene-based modulator.

II. A TUNABLE PHOTONIC SYNAPSE

Photonic synapses based on the incoherent (i.e. multi-wavelength) approach [6-8] are typically modelled by add-drop MRR designs. The representation of synaptic weights through MRRs depends on the amount of light that is trapped in the ring itself. By tuning such amount of light, it is possible to represent different weight values using a single device. Fig. 1(a) shows the schematic layout of a graphene-based optical modulator. The left side of the device is designed with a 50 nm-thick silicon layer that connects the silicon waveguide of the MRR (250 nm-thick) and both shallowly doped with boron to reduce sheet resistance. The waveguide is connected to gold contacts by a 7 nm-thick Al_2O_3 layer. Then, on the top and on the right side of the waveguide, a 0.1 nm-tick graphene monolayer was added. An additional 13 nm-thick gold layer of was extended from the right gold contact towards the bus waveguide to reduce the excess resistance of the device without disturbing the optical mode of the waveguide. The application of voltage on the gold-based contacts allows for electro-absorptive modulation through active tuning of the Fermi level of the graphene sheet.

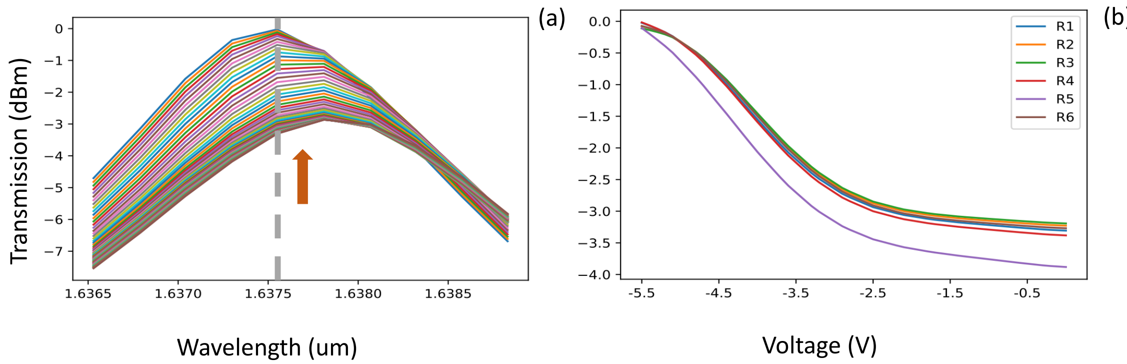

Fig. 2. (a) Transmission vs wavelength and (b) transmission vs voltage for the graphene-based MRR.

The numerical simulations were performed using Lumerical [9] HEAT and MODE were used to simulate the heater and modulator, and INTERCONNECT performed the circuit level simulations. Fig. 1(b) shows the increment of the loss (dBm/μm) in the waveguide when a voltage is applied to the gold contacts on the graphene-based MRR (R = 5 μm). In Fig. 2(a), we can see that for 65 different voltages the light that builds in the waveguide decreases progressively. As each weight can be defined in the photonic domain with the pair (voltage, transmission), the amount of weights that can ideally be represented would be 65. At λ_R = 1.6250 μm, the maximum weight value is found for a voltage equal to -5.5V, and then the rest of the weight values will be defined as the transmission decreases with incrementing voltage. We define this type of behaviour as a photonic synaptic weight that can be tuned vertically on the transmission profile. In Fig. 2(b), the transmission shows a monotonically decreasing trend with the increment of the driving voltage at λ_R= 1.6376 μm. For the six different radii (R1 = 5 μm, R2 = 6 μm, R3 = 7 μm, R4 = 8 μm, R5 = 9 μm and R6 = 10 μm), the behaviour is maintained with slight variations only.

Additionally, we developed an encoding protocol that can be followed to successfully represent information to be processed in photonics. Such protocol consists in the division of the voltage domain of the transfer functions so that each specific weight is represented by a range of transmission values. To define the weights, we divide the voltage range over which the modulator can operate into N bins of width ΔV. This binning-type of analysis is critical for the mapping of digital numbers to analog hardware in noisy environments. From this process, we found that a maximum number of 55000 synaptic weights can be represented on our device with 15.7 bits of resolution, based on the ability to control the voltage domain to a precision of 0.1 mV. In practice, this bit resolution will be limited by the performance of the control electronics, and the minimum detectable transmission change.

III. CONCLUSION

Graphene-based modulators embedded in MRRs can be used to potentially enhance the tuning efficiency of silicon photonic neural networks. These devices can overcome problems related to spectral cross-talk when cascading many MRRs in series, repeatability and control - while maintaining small footprint, high-speed and low power consumption.

REFERENCES

[1] E. Strubell, A. Ganesh, and A. McCallum, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, no. 1*, 3645 (2019).

[2] A. N. Tait et al., *Phys. Rev. Appl.* 11, 064043 (2019).

[3] C. Huang et al., *APL Photonics* 5, 040803 (2020

[4] M. Liu et al., *Nature* 474, 64 (2011).

[5] H. Jayatilleka, H. Shoman, L. Chrostowski, and S. Shekhar, *Optica* 6, 84–91 (2019).

[6] A. N. Tait et al., *Opt. Express* 26, 26422 (2018).

[7] R. Amin et al., *Appl. Opt.* 57, D130 (2018).

[8] P. R. Prucnal and B. J. Shastri. *Neuromorphic Photonics* (CRC Press, 2017).

[9] Lumerical Inc. https://www.lumerical.com/products/