Ultra-Efficient Modulators based on Chalcogenide and Graphene Materials for AI Hyperscalers

Jair A. de Carvalho
Daniel M. Neves
Jose C. Nascimento
Federal University of Ceara, Brazil
marchesi@fisica.ufc.br

Anderson L. Sanches
CECS-Information Engineering
Federal University of ABC
Santo Andre, Brazil
anderson.sanches@ufabc.edu.br

Alex Ferreira dos Santos
Federal University of Reconcavo of Bahia, Brazil
State University of Feira de Santana
alex.ferreira@ufrb.edu.br

Steevy J. Cordette
Directed Energy Research Center
Technology Innovation Institute
United Arab Emirates
steevy.cordette@tii.ae

Shyqyri Haxha
Department of Electronic Engineering
Royal Holloway, University of London
Egham, United Kingdom
shyqyri.haxha@rhul.ac.uk

Antonio Jurado-Navas
Thiago R. Raddo
Telecommunication Research Institute
University of Malaga, Spain
navas@ic.uma.es

***Abstract*— Driven by the global growth in AI services, the demand for efficient hardware is at an all-time high in the race for continued performance scaling of hyperscale data centers. Integrated photonics devices have emerged as a promising alternative for reducing energy consumption. We report an optical modulator design that incorporates chalcogenide and graphene materials to achieve the minimal power consumption per bit during high-speed operation. The device design and its energy consumption performance are presented. Our results pave the way for ultra-efficient modulators revealing operations at high-speed modulation bandwidth up to 104 GHz whilst reducing energy consumption to 52.3 fJ/bit.**

Keywords*— *chalcogenide, graphene, data center, energy consumption, photonics, AI, modulator, hyperscalers

I. INTRODUCTION

The extensive use of artificial intelligence (AI) applications handling vast quantities of data, underpinned by current inefficient hardware solutions, will pose a global energy consumption problem and present a significant challenge for future sustainability targets. Applications based on large language models constantly manipulate and process data in terabyte quantities, pushing energy consumption beyond expected future targets. As AI becomes more sophisticated and proliferates across industries, its escalating energy demand and associated CO_2 emissions will only grow. Google's greenhouse gas emissions rose last year by 48% (compared to the 2019 target) due to a surge in data center energy consumption and supply chain emissions [1]. Data centers and their associated transmission networks have become a primary driver of global energy consumption [2]. At present, this accounts for 3% of global consumption, emitting as much CO_2 as Brazil [2]. As energy requirements continue to grow, the current energy system, to which hyperscale data centers are connected, will face serious challenges. Although the industry is investigating novel solutions, it is clear that efficient hardware design in data communication and computing will be required to cope with the demand of the upcoming years, while supporting similar or lower energy consumption. The future of data centres, the backbone of the modern digital economy and AI workloads, will be shaped by the adoption of green technology and eventually new materials with better power consumption. Substantial improvements in the energy efficiency of information and communication technologies (ICT) devices for reduction of CO_2 emissions by means of greener technologies has become of major importance for the AI industry [3]. Greener technologies will play a pivotal role in the coming generation of ICT and AI hyperscalers solutions [3]. New materials such as chalcogenides and graphene will have significant importance in next-generation device solutions, which might result in a substantial increase in their market share over the next decade. From hyperscale data centers [4] to ICT networks [5], [6], energy-efficient optical devices such as modulators will play a key role in scaling performance progress while reducing power consumption. Accordingly, modulator devices account for key functionalities in many solutions, regardless of the application. In most telecommunication and hyperscaler systems, a signal must be modulated with data before its transmission or computational manipulation. In hyperscale data centers, optical modulators, typically based on silicon photonics, are indispensable for communication among switches, servers, racks, and clusters. Silicon photonics technology has contributed to new solutions with the capability to achieve efficient power consumption, higher-speed data transfer, and lower latency. Nevertheless, existing solutions cannot sustainably cope with the insatiable surge in hardware demand at hyperscalers as AI technology continues to be rapidly adopted around the world, pushing power consumption to unprecedented levels. The hyperscalers used to train and operate large language models require vast amounts of electricity. Optical modulators can be implemented using a combination of photonics and materials such as barium titanate [7], lead zirconate titanate [8], chalcogenide [9], graphene [10], or a combination of these. Graphene, a CMOS-compatible 2-D material with a single layer of carbon atoms, has many unique characteristics such as high thermal stability, tunable conductivity, and ultra-high carrier mobility. Optical modulators based on graphene have outperformed traditional modulators in terms of bandwidth provision, low energy consumption, and lightweight [10]. Chalcogenides have similar electro-optic properties to graphene (but very different thermal conductivity) and an ambipolar electro-optic response. It has been shown that the judicious design of graphene- and chalcogenide-based modulators can deliver unprecedented advancements [11].

In this paper, we introduce a new design for optical modulators and investigate its performance in terms of energy consumption. The OFF and ON chemical potential states of the proposed modulator are obtained. We address the performance of the new device, aiming to provide maximum efficiency in terms of energy consumption per bit for a given bandwidth and footprint. Preliminary results shown that it is possible to significantly reduce energy consumption by orders of magnitude without sacrificing the bandwidth performance of the device. The device achieved high-speed modulations up

This work was partially supported by the UEFS-AUXPPG 2024 and CAPES-PROAP 2024 grants

to 104 GHz, showing a power consumption as low as 52.3 fJ/bit with a 22.1 μm^2 footprint.

II. Device Design and Description

The modulator is composed of a graphene, alumina, tungsten diselenide (WSe_2) layers with distinct thickness, and a bus waveguide as well as a ring resonator both based on silicon. Examples of different transition metal dichalcogenide (TMD) materials that can be integrated with silicon photonics are shown in Fig. 1(a). The novel device uses WSe_2 due to the ambipolar nature of its electro-optic effect [11]. It is worth point out that this novel structure has a small formfactor, being favorable for the implementation of an array structure of various devices, as illustrated in Fig. 1(b). The design of the novel modulator is shown in Fig. 2(a). The bus waveguide has width $W_1 = 340$ nm coupled to a ring-shaped silicon resonator with width $W_2 = 600$ nm with a thickness $hS_i = 220$ nm. Bear in mind that the bus waveguide supports single mode operations. The ring-shaped silicon resonator has three thin layers on its top. The first layer is composed of graphene and has thickness $hg = 0.34$ nm, the second layer is composed of alumina with thickness $ha = 7$ nm, and finally the last layer is made of a chalcogenide material with thickness $hWSe_2 = 0.65$ nm. A dielectric ring resonator is formed from the graphene and WSe_2. The fundamental trade-off between device length and optical loss can be alleviated by the unique combination of graphene and chalcogenide materials. The device structure is surrounded by silica substrate. Lastly, there are two metal contacts with each connected to the graphene and the chalcogenide layers forming a capacitor. The electro-optic properties of both graphene and chalcogenide materials can be modified by the application of a voltage across their layers.

An optical signal goes through the bus waveguide, couples to the ring resonator, and exists via the bus waveguide. This renders a phase modification to the transmitted optical signal. The latter can be used for efficient high-speed data bit modulation in AI hyperscalers. The usage of up to hundreds of individual ring resonators can be processed on a single wafer to scale information modulation output to optimally match specific application requirements in AI hyperscalers. Noteworthy, the values for the bus waveguide and ring resonator width are optimized in order to maximize their efficiency in terms of gap considering the fabrication feasibility and minimal bending losses for resonators with external radius of 2.08 μm whilst still preserving a fundamental mode resonant structure. Finally, it is more advantageous to cover the entire resonator area with graphene and chalcogenide to increase the bandwidth modulation performance of the device. Nevertheless, this may result in increased losses and power consumption. Unique performance advantages (power consumption, bandwidth, and losses) can be achieved by optimizing the amount of material (graphene and chalcogenide) used in the design of the modulators, as previously investigated in [10].

III. Results and Discussion

The physical understanding of graphene's optoelectronic properties are of great interest for AI hyperscalers applications. Graphene allows its electrons close to the k point to have zero effective mass by presenting an energy curve of linear dispersion. This renders the material with high charge mobility and extremely sensitive to its optical properties variations, which can be achieved with small variations in the level of their chemical potential, whether achieved via chemical interaction with other molecules or atoms, or via application of an external electric field, changing their dielectric and optical absorption properties. The energy levels of various 2-D materials are shown in Fig. 1(a). The chemical potential ON-OFF state variation as a function of the wavelength in then addressed. The results are plotted in Fig. 2(b). These values give important evidences on the modulator's behavior. Indeed, the chemical potential obtained to the OFF state is key parameter to yield distinct characteristics in the device's behavior and consequently operation. The simulation considers that a quasi-TE fundamental mode is excited at the input port 1 of the bus waveguide as it provides a better optical confinement than does the fundamental quasi-TM for the adopted cross-sectional dimension. Furthermore, since the optical and electrical response of graphene is much faster than in doped silicon, it reduces the bandwidth limitation related to the rise and relaxation lifetime of the free carrier concentrations [10]. It is worth mentioning that the loss related to the ring's curvature as well as the graphene optical absorption are taken into account by considering gap values satisfying the modulator's critical coupling condition. Also, the graphene's permittivity is calculated via the Kubo's formula. The values for the refractive indexes of silicon, silica and alumina are obtained from experiments [10].

An approach is followed that the ratio of the index change (Δn_{eff}) to the loss change (Δk_{eff}) has to be unity approximately ($\Delta n_{eff} / \Delta k_{eff} \approx 1$) [11]. Accordingly, the novel device design obeys this condition and operates near the critically coupled regime. The device's ring coverage with the chalcogenide material is 100%, which satisfies the near critically coupled

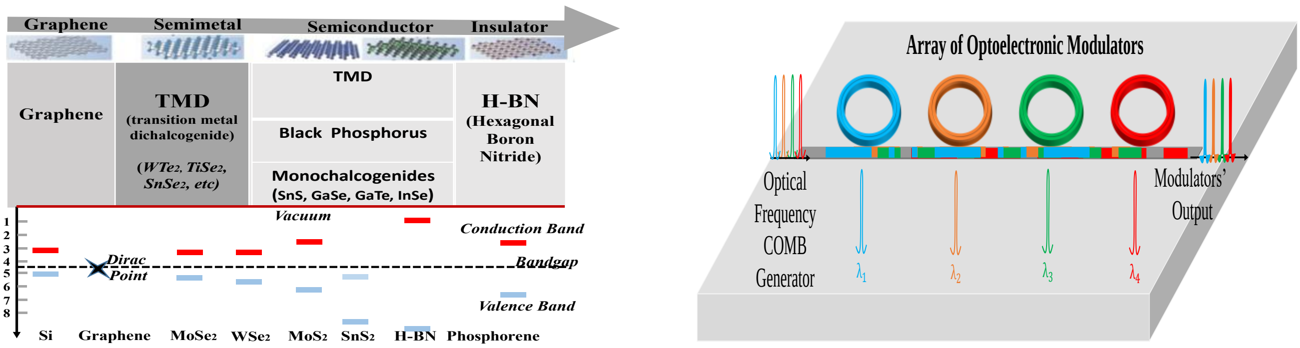

Fig. 1. (a) left: A wide range of 2-D materials from conductors to insulators and their energy levels. (b) right: Array structure of modulators comprising up to various individual devices. TMD: transition metal dichalcogenide.

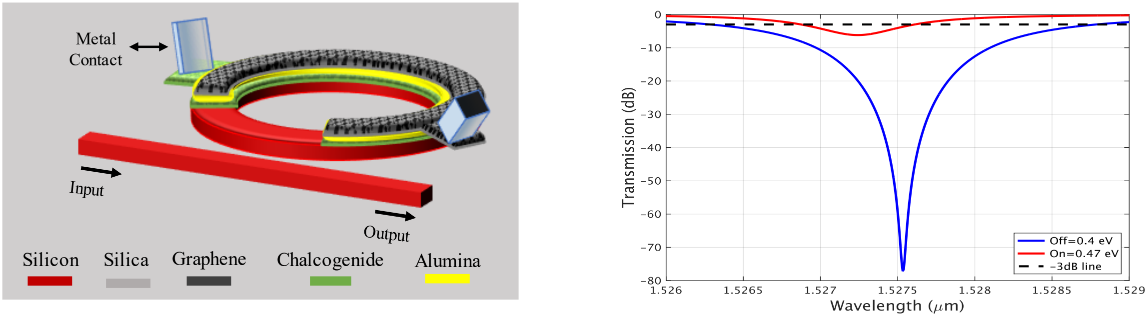

Fig. 2. (a) left: Design of the new graphene chalcogenide ring-resonator modulator coupled with a bus waveguide. (b) right: The modulator's transmission curves versus the wavelength considering both OFF and ON states.

regime and consequently supports a large phase shift with low insertion loss. In this instance, an alteration to Δn_{eff} of the resonator causes a resonance detuning effect, whereas a modification to the Δk_{eff} of the resonator ring results in a detuning of the resonator's critically coupled condition, thereby modifying the all-pass transmission profile, which leads to an overly-coupled state. The idea here is to obtain the optimum combination of two materials (graphene and chalcogenide) with distinct phase and loss variation profile to vary the imaginary and real parts of the waveguide's effective index. It is important to note that the modulator design should be optimized to achieve the maximum phase variation with the minimum possible loss variation. The variation function of the real and imaginary parts of the effective index was addressed for the proposed device. A chemical potential equal 0.47 eV (ON state) and 0.4 eV (OFF state) were achieved, as can be seen in Fig. 2(b). The initial chemical potential considered was 0.4 eV. The WSe_2 material is used in such a way that low losses and efficient phase variation response for the C-band range allow this condition to be achieved in an efficient fashion. This can be particularly interesting for performance metrics such as power consumption and insertion loss. Accordingly, for a resonance wavelength of 1.5275 um, an insertion loss as low as -3.71 dB is achieved for a surprisingly low power consumption of ~52.3 fJ/bit. The latter represents a remarkably close to 40% energy consumption reduction if compared to classical graphene's double layer devices (which only achieves the same $\Delta n_{eff} = \Delta k_{eff}$ condition for an OFF state chemical potential equal to 0.5 eV) [10]. Such performance can be further improved by fine-tuning the device by reducing the active area or by using different amounts of material [10], leading to a proportional reduction in energy consumption with a small cost in terms of a reduction in total bandwidth. However, even without considering a fine-tuning of this device, it reaches a total, RC, and optical bandwidths [10] equal to 104 GHz, 110 GHz, and 319.6 GHz, respectively. These results pave the way for the future realization of emerging applications with high integration and scalability, such as arrays of graphene chalcogenide modulators comprising hundreds of individual devices to increase the data bit modulation requirements in AI hyperscalers.

IV. Conclusions

In this paper, we have proposed a novel modulator design based on the combination of two distinctive and unique materials: chalcogenide and graphene. They combination represent a significant advance in energy efficiency. The modulator is comprised of a layer of chalcogenide and graphene. The critical coupling condition representing the OFF-state, and the 3-dB transmission level representing the ON-state, are obtained. The results indicated that the novel modulator is capable of achieving d.c. energy consumption as low as nanowatt order of magnitude [11], along with ~52.3 fJ/bit, whilst operating at high-speed rates (104 GHz) with a minimal footprint of just 22.1 μm^2. The combination of different materials remains of key importance for energy consumption reduction. The use and optimization of new materials enable unprecedented performance levels of optoelectronic devices, resulting in energy reduction. In future work, the effects of varying quantities of graphene and chalcogenide on performance will be examined in order to further enhance performance. The full analysis of the newly designed modulator and its complete performance will be published elsewhere. The proposed modulator design, based on graphene and chalcogenide materials, is capable of addressing the issue of power consumption while maintaining a high-speed bandwidth. The development of such an array of optical modulators, comprising up to several individual sources, can lead to substantially increased requirements in data modulation and power consumption for applications in AI hyperscalers. Energy efficiency remains a primary concern for hyperscalers. As these facilities continue to expand or newly emerge to meet the demands of the digital age, it is of the utmost importance to optimise power usage.

References

[1] Google's Environmental Report, 2024: https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf.

[2] https://www.forbes.com/sites/arielcohen/2024/05/23/ai-is-pushing-the-world-towards-an-energy-crisis/

[3] C. A. Silva *et al.*, "A review on the decarbonization of high-performance computing centers," *Renewable and Sustainable Energy Reviews*, 2024.

[4] K.-I. Sato, "Optical switching will innovate intra data center networks," *Journal of Optical Communications and Networking*, vol. 16, no. 1, pp. A1-A23, 2024.

[5] D. Neves *et al.*, "Beyond 5G fronthaul based on FSO using spread spectrum codes and graphene modulators," *Sensors*, vol. 23, no. 8, p. 3791, 2023.

[6] T. R. Raddo *et al.*, "Transition technologies towards 6G networks," *Journal on Wireless Communications and Networking*, 100, 1-22, 2021.

[7] L. Czornomaz *et al.*, "BTO-enhanced silicon photonics - a scalable PIC platform with ultra-efficient electro-optical modulation," *OFC*, 2022.

[8] G. F. Feutmba *et al.*, "High frequency characterization of PZT thin-films deposited by chemical solution deposition on SOI for integrated high speed electro-optic modulators," *Opical Materials Express*, 13, 2023.

[9] I. Datta *et al.*, "Low-loss composite photonic platform based on 2D semiconductor monolayers," *Nature Photonics*, 14, 256-262, 2020.

[10] D. Neves *et al.*, "Power consumption analysis of an optical modulator based on different amounts of graphene," *Optics Continuum*, 1, 2077-2090, 2022.

[11] I. Datta *et al.*, "2D material platform for overcoming the amplitude-phase tradeoff in ring resonators," *Optica*, 11, 48-57, 2024.