Towards a Robust Multiply-Accumulate Cell in Photonics using Phase-Change Materials

Raphael Cardoso[1*], Clément Zrounba[1], Mohab Abdalla[12], Paul Jimenez[1], Mauricio Gomes de Queiroz[1], Benoît Charbonnier[3], Fabio Pavanello[1], Ian O'Connor[1], Sébastien Le Beux[4]
[1]Univ. Lyon - CNRS, Ecole Centrale de Lyon, INSA Lyon, Université Claude Bernard Lyon 1, CPE Lyon - INL, UMR5270 - Écully, F-69134, France
[2]School of Engineering, RMIT University, Melbourne, VIC 3000, Australia
[3]CEA LETI - Grenoble, France
[4]Department of Electrical and Computer Engineering, Concordia University - Montreal, Canada
*raphael.cardoso@ec-lyon.fr

***Abstract*—In this paper we propose a novel approach to multiply-accumulate (MAC) operations in photonics. This approach is based on stochastic computing and on the dynamic behavior of phase-change materials (PCMs), leading to the unique characteristic of automatically storing the result in non-volatile memory. We demonstrate that, even with perfect look-up tables, the standard approach to PCM scalar multiplication is highly susceptible to perturbations as small as 0.1% of the input power, causing repetitive peaks of 600% relative error. In the same operating conditions, the proposed method achieves an average of $7\times$ improvement in precision.**

***Index Terms*—photonic computing, phase-change memories, stochastic computing, computation-in-memory**

I. Introduction

Following the needs of electronic crossbars for multiply-accumulate (MAC) operations in AI applications, photonic approaches are being investigated with the promise of improvements in operation speed due to the unmatched parallelism achieved with light [1]. Amongst photonic MAC alternatives, most are interference-based, which utilize large, power-hungry phase-shifters to compute [1]. On the other hand, optical phase-change materials (PCMs) offer a small footprint, non-volatile alternative [2]. As shown in Fig. 1a, this method is based on the multiplicative interaction between the state of a PCM, i.e. the distribution of amorphous/crystalline phase, and the amplitude of an optical pulse. Thus we refer to it as amplitude-oPCM.

An important parameter for computing with PCMs is the number of intermediate states that it allows. Currently, up to 64 states (6 bits) have been demonstrated [3]. As new devices with higher bit counts become available, it is intuitive to expect better precision from them. However, in this paper we show that amplitude-oPCM actually *degrades* in accuracy as more bits are available in the presence of small perturbations. Besides, this approach further requires: i) precise adjustments of input intensity, ii) precise writing of values to the multi-level memory, iii) multi-wavelength light sources for MAC implementation. In this context, we propose stochastic-oPCM, a single-wavelength approach that requires only 1-bit modulators, and thus does not depend on the precise adjustment of either the light pulses or the memory state. Our approach operates by continuously updating the accumulated values in the PCM state, therefore the results are automatically written in memory. To achieve multiplication, we rely on stochastic computing, which also includes error in the system, thus we compare the accuracy of both methods in the presence of small output perturbations, that could be caused by receiver electronics noise or temperature fluctuations.

This work was funded by Agence Nationale de la Recherche (ANR) under grant no. ANR-20-CE24-0019 (OCTANE).

II. Proposed Approach

A. Stochastic-oPCM Cell

In stochastic computing, data is converted to pseudo-random bitstreams using stochastic number generators (SNGs). Multiplication is performed by feeding these bitstreams to an AND gate and an accumulator, as seen in Fig. 1b. If the inputs are uncorrelated, the accumulated *1*s will represent the product between the inputs, with additional error due to the approach's randomness [4]. In photonics, a PCM deposited on top of a waveguide crossing can act as both the AND gate and the accumulator. To do so, we split a short, high power amorphizing pulse between the two branches of the crossing as seen in Fig. 1c, each carrying a stochastic bistream encoded by a one-bit optical modulator. Therefore, if two *1*s arrive simultaneously, the PCM will partially amorphize, performing the AND operation [5]. This operation is cumulative, so subsequent pulses will amorphize other parts of the material as illustrated in Fig. 1d. A disadvantage of this approach is that thermodynamic equilibrium must be reached between melting events, incurring a nanosecond delay between pulses [6]. Finally, the transmission through an optical PCM depends on its state (inset of Fig.1a), therefore we send a non-melting pulse with known amplitude to recover the result.

B. Simulation Environment and Methodology

We implemented a behavioral simulation environment in Python based on results from multiphysics simulations of a $Ge_2Sb_2Te_5$ thin film of $100\text{ nm}\times250\text{ nm}\times20\text{ nm}$ deposited on a $400\text{ nm}\times180\text{ nm}$ silicon waveguide. This PCM requires an

978-3-9819263-7-8/DATE23/© 2023 EDAA

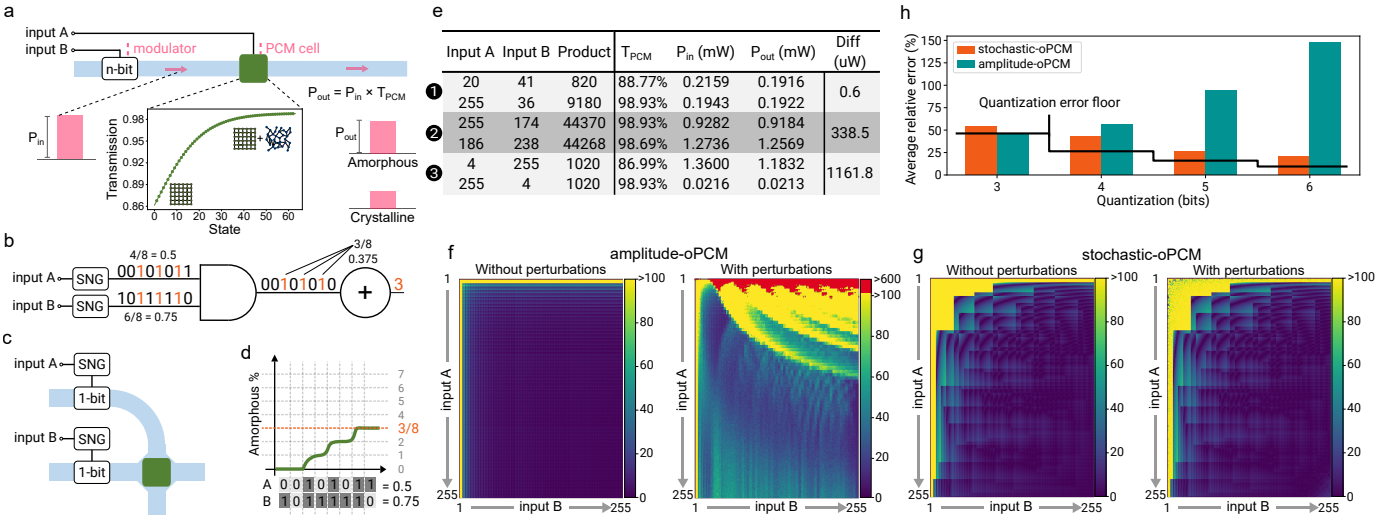

Fig. 1. (a) Scalar multiplication with amplitude-oPCM. Inset: PCM transmission characteristic based on its state, (b) multiplication in stochastic computing with digital components, (c) cell for stochastic-oPCM multiplication, (d) PCM state evolution during stochastic-oPCM operation, (e) amplitude-oPCM operation for three cases, (f) relative error (%) of all operations in amplitude-oPCM multiplication, (g) relative error (%) for all operations in stochastic-oPCM, (h) average relative error for both methods with different quantizations.

optical pulse of 13.6 mW and 500 ps to amorphize one step. For stochastic-oPCM, each *1* in a bitstream is encoded with 6.8 mW and thus we assume that the PCM state behaves as a discrete counter of amorphizing events with the transmission characteristic from Fig. 1a. To avoid corrupting the memory state at the end of operation, we use a non-melting readout pulse of 1.36 mW.

For amplitude-oPCM, we model the PCM as an attenuator with input-dependent transmission corresponding to the graphic in Fig. 1a. Input B is encoded linearly as an optical pulse between 0 and 1.36 mW. Optical losses are not considered in the simulations as they do not affect the precision, but rather the required laser power. The results are recovered from a look-up table obtained from unperturbed simulations.

III. Simulation Results and Discussion

All tests in this section represent the average of 100 simulations of scalar multiplication considering 8-bit values, quantized down to 6 bits. Initially, in Fig. 1e, we demonstrate the sensitivity of amplitude-oPCM with three examples without perturbations: (❶) two largely different products result in a similar output power, differing only by 0.6 µW, (❷) two similar products produce largely different outputs, (❸) multiplication is not commutative. When all the powers are exact, the correct output for any input can be recovered from the look-up table, as shown in Fig. 1f, in which the error comes only from quantization. However, as soon as we include a normally distributed perturbation with standard deviation of 1.36 µW, 0.1% of the readout power, high relative error peaks appear, surpassing 600%.

We also performed the same simulations for stochastic-oPCM, as shown in Fig. 1g. We observe that it also produces error, especially for small inputs and almost entirely due to the stochastic paradigm, as the behavior barely changes when perturbations are included. Lastly, we verify the average error of all operations for different bit quantizations, i.e. using PCMs with fewer levels, with results shown in Fig. 1h. In this case, we note that amplitude-oPCM achieves minimum error for 3 bits, due to quantization alone, but quickly degrades to 145% average error with 6 bits. On the other hand, stochastic-oPCM improves with more levels due to longer bitstreams [4], reaching 21% average error with 6 bits.

IV. Conclusions

We propose a novel approach to achieve MAC operations in photonics using PCMs that requires less complex control than the state-of-the-art. We show that our method's accuracy is more robust to small perturbations. The proposed approach also includes error in the operation due to the nature of stochastic computing, and reducing it requires longer bitstreams, and thus PCMs with higher level counts. In future work, we intend to validate our assumptions with experimental data. Furthermore, we shall investigate how the stochastic-oPCM cell can be used in Computation-In-Memory architectures, taking advantage of its unique characteristic of automatically writing the result to memory.

References

[1] N. Peserico *et al.*, "Integrated photonic tensor processing unit for a matrix multiply: a review," *arXiv preprint arXiv:2211.01476*, 2022.

[2] J. Feldmann *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.

[3] C. Wu *et al.*, "Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network," *Nature communications*, vol. 12, no. 1, pp. 1–8, 2021.

[4] A. Alaghi and J. P. Hayes, "Survey of stochastic computing," *ACM Transactions on Embedded computing systems (TECS)*, vol. 12, no. 2s, pp. 1–19, 2013.

[5] J. Feldmann *et al.*, "Calculating with light using a chip-scale all-optical abacus," *Nature communications*, vol. 8, no. 1, pp. 1–8, 2017.

[6] T. Y. Teo *et al.*, "Programmable chalcogenide-based all-optical deep neural networks," *Nanophotonics*, 2022.