Photonic Convolution Engine Based on Phase-Change Materials and Stochastic Computing

Raphael Cardoso, Clément Zrounba, Mohab Abdalla, Paul Jimenez, Mauricio Gomes, Benoît Charbonnier, Fabio Pavanello, Ian O'Connor, Sébastien Le Beux

Photonic Convolution Engine based on Phase-Change Materials and Stochastic Computing

Raphael Cardoso[1*], Clément Zrounba[1], Mohab Abdalla[12], Paul Jimenez[1], Mauricio Gomes[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]Univ. Grenoble Alpes, CEA, LETI - F38000 Grenoble, France
[4]Department of Electrical and Computer Engineering, Concordia University - Montreal, Canada
*raphael.cardoso@ec-lyon.fr

***Abstract*—The last wave of AI developments sparked a global surge in computing resources allocated to neural network models. Even though such models solve complex problems, their mathematical foundations are simple, with the multiply-accumulate (MAC) operation standing out as one of the most important. However, improvements in traditional CMOS technologies fail to match the ever-increasing performance requirements of AI applications, therefore new technologies, as well as disruptive computing architectures must be explored. In this paper, we propose a novel in-memory implementation of a MAC operator based on stochastic computing and optical phase-change memories (oPCMs), leveraging their proven non-volatility and multi-level capabilities to achieve convolution. We show that resorting to the stochastic computing paradigm allows one to exploit the dynamic mechanisms of oPCMs to naturally compute and store MAC results with less noise sensitivity. Under similar conditions, we demonstrate an improvement of up to 64× and 10× in the applications that we evaluated.**

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

I. Introduction

We are experiencing a significant shift in computer resource allocation, with more and more hardware being dedicated to AI applications. As convolutional neural networks with billions of parameters are used to solve a growing set of problems [1], it becomes necessary to optimize their implementations. In these networks, convolution has the largest energy overhead [2], requiring intensive use of multiply-accumulate (MAC) operations. While the energy efficiency, density, and speed of MACs have significantly improved over the last decades, further developments face two fundamental limits: i) the slowdown of Dennard's scaling makes it increasingly difficult to fabricate small and efficient electronic devices, ii) traditional architectures are affected by the Von-Neumann bottleneck, as large amounts of energy are spent transferring data between processing and memory units.

In this context, new in-memory computing (IMC) architectures are interesting alternatives, with key operations such as MACs being performed within memory banks. Electronic IMC implementations typically use non-volatile memories such as resistive RAM (ReRAM), ferroelectric devices (FeFET), NAND Flash, or phase-change memories (PCMs). Among these, PCMs are the only ones that can also be manipulated optically, enabling implementation of IMC circuits also in the photonic domain. Compared to their electronic counterpart, optical phase-change memories (oPCMs) have a longer lifetime [3], coupled with improved multi-state behavior, such that devices with up to 64 different states (6 bits) were demonstrated [4], versus 8 states for electrical PCMs [5]. A

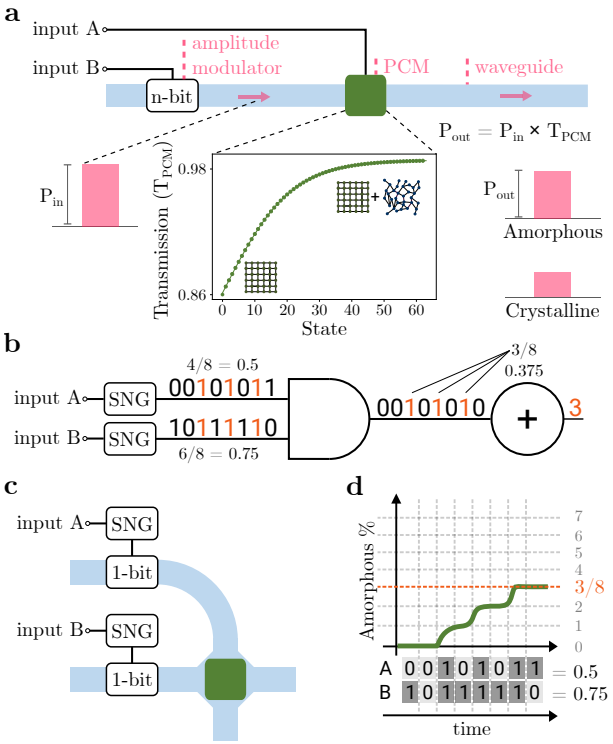

Fig. 1. (a) Scalar multiplication with amplitude-oPCM. Inset: PCM transmission characteristic based on its state, noted as the corresponding 6-bit value encoded on it, (b) multiplication in stochastic computing with digital components, (c) cell for stochastic-oPCM multiplication, (d) PCM state evolution during stochastic-oPCM operation.

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

PCM state is defined by the distribution of amorphous and crystalline phases inside the material, which directly affects its optical transmission. Changing the value stored in a PCM involves controlling its temperature in one of two specific ways: i) melting it with high intensity and quenching it as fast as possible, resulting in amorphization, and ii) annealing it by applying specific temperature profiles to recrystallize it [6].

In photonics, the interaction between the non-volatile state stored as the PCM transmission and the amplitude of an optical signal is naturally multiplicative, as illustrated in Fig. 1a. Therefore, such devices have been investigated to perform convolutions by encoding a constant multiplier into the PCM, while other operands are sent as the amplitude of light pulses with different wavelengths, simultaneously traveling in a waveguide [7–10]. This approach, which we refer to as amplitude-oPCM, performs the mathematical operation while *reading* the non-volatile memory and has been demonstrated to be highly sensitive to output perturbations in scalar multiplication [11]. Also in [11], a new cell that performs MACs while *writing* the memory (stochastic-oPCM) is proposed as a robust alternative to such perturbations.

In this work, we propose a novel optical convolution circuit based on stochastic-oPCM cells. In such cells the accumulation results are continuously updated as the PCM state, illustrated in Fig. 1c-d. To achieve multiplication, we implement the stochastic computing paradigm, in which inputs are encoded as streams of digital optical pulses. Only the combined energy of two simultaneous pulses triggers a change of the PCM state, thus it behaves as a logical AND. Therefore a single cell is capable of both the multiply and accumulate operations shown in Fig. 1b. To evaluate this approach, we propose a behavioral modelling scheme based on extrapolation from finite-difference time-domain (FDTD) simulations, and we use it to compare our solution with state of the art optical MACs over two applications: i) denoising with a convolutional filter and ii) RGB-grayscale conversion of an image, both in the presence of photodetector noise. Our results show that our approach improves upon amplitude-oPCM up to $10\times$ and $64\times$ in each respective application in terms of precision.

This paper is organized as follows: section II presents theoretical background for this work, as well as related research in the literature. In section III we describe our PCM modeling scheme, as well as the proposed convolution circuit based on stochastic-oPCM cells. In section IV we discuss the simulation results for the target applications, and section V brings the conclusions and perspectives to this work.

II. Background

In this section, we present key concepts of stochastic computing and phase-change memories in optics. Then, we discuss works that perform computing with oPCMs.

A. Stochastic Computing

Stochastic computing (SC) is a paradigm that operates on data encoded as densities of *high* states ('1') in bitstreams. Since most conventional computing paradigms rely on binary-encoded data (rather than stochastic), conversion to stochastic bitstreams is needed. This is achieved by a Stochastic Number Generator (SNG) comprised of a pseudo-Random Number Generator (RNG) and a comparator [12]. In this work, we consider maximal linear-feedback shift registers (LFSRs) as RNGs, which require as many registers as there are bits in the input data. With this, any n_b-bit value can be mapped to a *high* state probability with bitstream length (BSL) of $2^{n_b} - 1$.

To perform stochastic multiplication, let us consider two bitstreams *A* and *B* as independent, i.e., generated by two different LFSRs. Then, p_A and p_B are the probabilities of a *high* state appearing in each of these bitstreams. When they are applied as inputs of an AND gate, the probability of a *high* state at the output is calculated as $p_Q = P(A = 1 \cap B = 1) = P(A = 1) \times P(B = 1) = p_A p_B$. Finally, the output can either be used as-is, or converted to binary for storage or communication. The conversion is performed using a counter (accumulator). Fig. 1b illustrates this concept for $0.5 \times 0.75 = 0.375$ with 3-bit encoding. An important caveat of SC is that the output will often contain error, and while there are techniques that perform exact stochastic multiplication, their BSL is squared [13], rendering them impracticable for this work.

B. oPCMs

Phase-change memories are tipically based on chalcogenide glasses such as GeTe, Sb_2Te_3, or $Ge_2Sb_2Te_5$ (GST). This family of materials has been widely used in the rewritable optical storage industry since the late 90s, as they display distinct optical properties between their amorphous and crystalline phases [14]. This results in a state-dependent variation of the amount of light transmitted through a device, as shown in Fig. 1a. Furthermore, by progressively switching sub-regions of the PCM cell from crystalline to amorphous, many intermediate states can be accessed, with the most recent demonstration showcasing 64 distinct levels [4]. Assuming that the memory is initially 100% crystalline, partial amorphization can be achieved with a short, high power optical pulse that brings an inner region of the PCM above its melting point, resulting in an amorphized area [6]. Repetition of such pulses leads to the nonlinear accumulation characteristic represented in the inset of Fig. 1a, in which each discrete state corresponds to the number of amorphization events caused by input pulses [15].

C. PCM-based Optical Storage and Computing Devices

Given the possibility to optically switch PCMs and their large storage capacity, these devices have already been used as integrated optical memory cells [16, 17] allowing efficient non-volatile memory banks. The non-volatility of phase-change materials also opens up the path to in-memory computing applications, in particular for multiply-accumulate operators. The existing method for oPCM MAC involves encoding the multiplier as the state of the PCM cell and other operands as amplitudes of light pulses (amplitude-oPCM). The multiplication result is carried by the amplitudes at the output [18], and

these can in turn be summed by a photodetector. This strategy has already been used to perform matrix multiplication [7, 8] and convolution [9], with other works investigating the inclusion of negative operands [9, 10] at the cost of increased post-processing. When implementing such operators using GST-based photonic devices, there is always light loss due to absorption in the material. With the goal of allowing the construction of larger systems, different low-loss PCM materials, such as Sb_2S_3, are being investigated [19].

A major limitation of amplitude-oPCM is the non-uniform mapping between the photodectector output and its equivalent operation result [11]. This hinders result recovery after scalar multiplication, leading to heavy data corruption in the presence of electronic noise in the readout circuit. This corruption intensifies as more intermediate states are used, restricting this approach to operations with binary $(0, 1)$ or ternary $(-1, 0, 1)$ states under realistic conditions. For this reason, instead of using PCMs as *static* devices, we perform computations by changing the PCM state itself, such that the result recovery can be performed by a pulse with a fixed known amplitude. This simplifies the readout and leads to noise-resiliency, allowing reliable operation at higher bit counts. To our knowledge, only one work actively uses the *dynamic* properties of oPCMs to perform arithmetic functions [20]. Additionally, only two works apply stochastic computing in photonics [21, 22], without including phase-change memories.

III. Proposed Approach

As demonstrated in [20], a waveguide crossing with a PCM deposited on its top can function as an AND gate for amorphization. We use this effect, associated with the fact that each amorphization step accumulates over the previous, to implement the stochastic-oPCM multiply-accumulate cell. To simulate each cell in the context of a bigger convolution circuit, we developed a behavioral model described in the following.

A. oPCM Behavioral Model

Here we consider that a PCM device can be modeled purely through its transmission characteristic, i.e. we neglect temporal effects. This assumption is only valid if we respect the time t_{rest} to thermodynamic equilibrium between optical pulses, which is in the order of nanoseconds [19]. Thus, for each discrete state of amorphization, the PCM will have a different transmission, calculated from device-level simulations as shown in Fig. 2. Furthermore, for stochastic-oPCM, it is necessary to know how much energy must be injected in the waveguides to amorphize the PCM at the initial (fully crystalline) state. This energy is also calculated from device-level simulations.

For the device-level simulations, we used Ansys™ Lumerical HEAT and FDTD solutions. We employ FDTD simulations to obtain the PCM transmission in both fully crystalline (T_0) and fully amorphous (T_k) states. As we only simulate the optical characteristics at those two edge cases, we apply curve-fitting to map to produce a curve with the desired number of

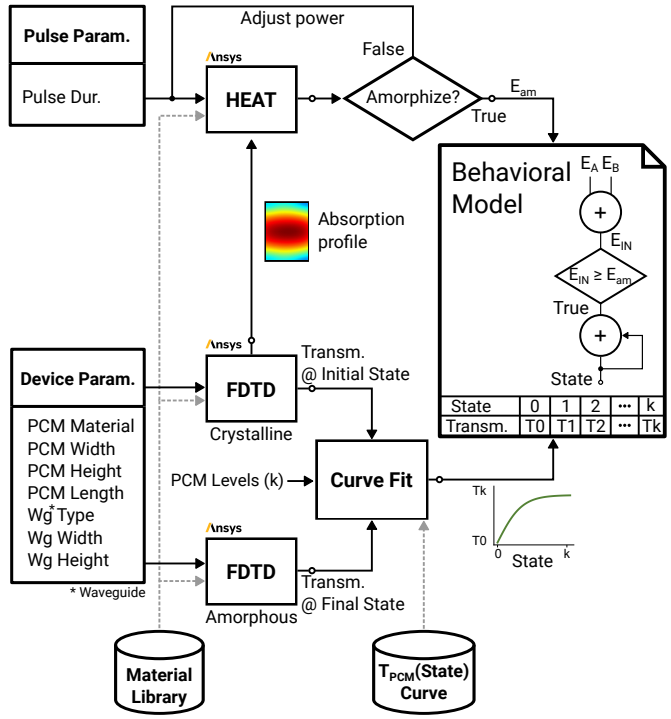

Fig. 2. Flow of necessary device simulations to construct the behavioral model used in this work.

intermediate amorphization states (k). In this step we assume the following function for transmission:

$$\mathrm{T_{PCM}} = T_0 + (T_k - T_0) \times \tanh\left(3 \times \frac{\mathrm{State}}{k}\right), \qquad (1)$$

which represents the nonlinearity of the PCM transmission as it amorphizes [4, 15]. Although the specific shape of curve is highly dependent on the device's geometry, we verified that the specific shape only marginally affects the circuit-level behavior investigated in this work.

As the phase-transition depends on thermal characteristics, we perform HEAT simulations based on the spatial profile of absorbed optical power with a determined pulse duration. From this, we find the minimum input power that will raise the temperature of a region inside the PCM above its melting point. Since we operate with nanosecond pulses and integrated devices, the melted region quickly cools down and becomes amorphous [6].

Knowing the threshold amorphization energy E_{am}, our behavioral model for the PCM dynamics works as follows: i) the energy of both input pulses is summed; ii) if the total input energy E_{IN} is equal or higher than E_{am}, the PCM amorphizes by one step incrementing its *State* variable; iii) a new optical transmission, dependent on *State*, is assigned to the device. During actual implementation, the total energy must be slightly above E_{am}, otherwise the PCM would stop amorphizing after the first pulse due to reduced absorptions.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} * \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

$$\begin{aligned} c_{11} &= a_{11}\,b_{11} + a_{12}\,b_{12} + a_{21}\,b_{21} + a_{22}\,b_{22} \\ c_{12} &= a_{12}\,b_{11} + a_{13}\,b_{12} + a_{22}\,b_{21} + a_{23}\,b_{22} \\ c_{21} &= a_{21}\,b_{11} + a_{22}\,b_{12} + a_{31}\,b_{21} + a_{32}\,b_{22} \\ c_{22} &= a_{22}\,b_{11} + a_{23}\,b_{12} + a_{32}\,b_{21} + a_{33}\,b_{22} \end{aligned}$$

$$\xrightarrow{\quad\text{time}\quad}$$

Fig. 3. Concurrent computation of $A * B = C$, with a time-multiplexed kernel.

B. Convolution with stochastic-oPCM

Given the importance and generality of convolution at the core of many algorithms for data processing, we implement it with stochastic-oPCM cells. To make use of structural parallelism, we evaluate all output coefficients concurrently, while the kernel elements are time-multiplexed as shown in Fig. 3.

In Fig. 4 we illustrate the proposed circuit for performing convolution with stochastic-oPCM. At each step, we send the converted stochastic bitstream for a kernel coefficient b, as well as all corresponding coefficients of the input matrix A, yielding a $N \times N$ PCM matrix. An appropriate power distribution must be guaranteed when routing input B such that a total energy of $E_{am}/2$ arrives in each PCM from this channel. Therefore, the first splitter must divide the pulses in N equal parts, while the directional couplers along each row must be sized to equally distribute the power, i.e., the nth cross/through ratio is $1/(N+1-n)$.

Under these assumptions, the proposed circuit can perform a sum of S successive convolutions such that the output is

$$C_{N\times N} = \sum_{i=1}^{S} A^{i}_{(N+M-1)\times(N+M-1)} * B^{i}_{M\times M}, \qquad (2)$$

where N is the number of rows/columns of the layout in Fig. 4 and M is the kernel dimension. The summing of successive convolutions is a consequence of the PCMs' non-volatility, which store the result of previous operations until reset.

Regarding the stochastic computing paradigm, only two LFSRs are necessary for convolutions of any size: one shared by all A channels, and one for B. Indeed, the PCMs do not interact with each other, and only the PCM inputs must be statistically independent from each other. Furthermore, as the accumulation of time-multiplexed inputs is not stochastic, there is no need for statistical independence between different time steps.

If we consider that the duration of operation t_{op} depends only on the optical components, we can evaluate it as scaling with the size of the operands in (2) as

$$t_{\text{op}} = SM^2(2^{n_b}-1) \times t_{\text{rest}}, \qquad (3)$$

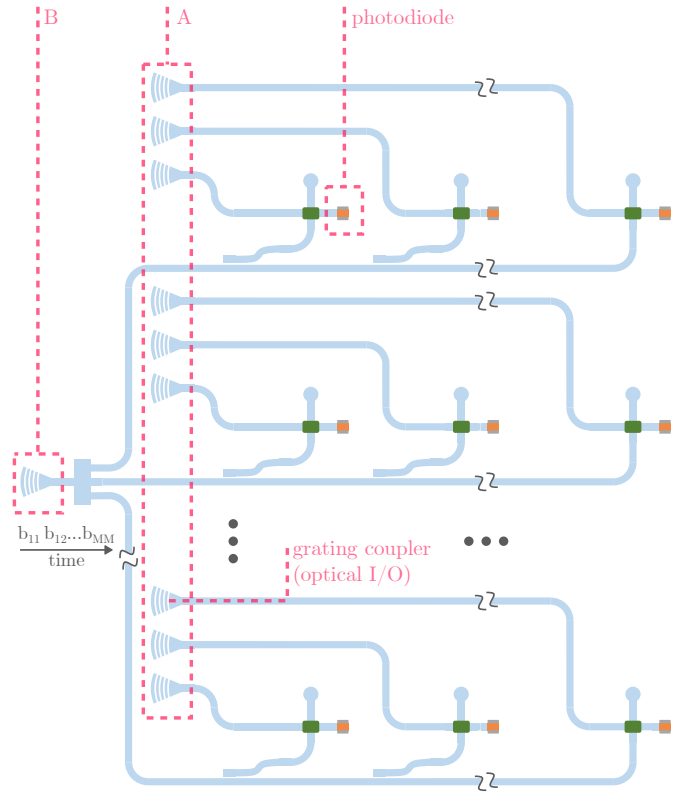

Fig. 4. Circuit for convolutions with stochastic-oPCM cells.

where S is the number of successive convolutions, , $2^{n_b}-1$ is the bitstream length of the stochastic operands and t_{rest} is the thermal relaxation time that must be respected between pulses. Additionally, the energy cost also depends on the size of the circuit such that, disregarding losses and considering the worst-case bitstream (all 1s), it is evaluated as

$$E_{\text{op}} = SM^2N^2(2^{n_b}-1) \times E_{\text{am}}. \qquad (4)$$

At the end of the operation, the output matrix is written as PCM states. Output recovery is performed by reading the PCMs with a known optical power sent through input B, that spreads it to all PCMs and their photodiodes. Each photodiode generates an electrical current that carries information about the PCM state, which can be decoded and used elsewhere. To avoid corrupting the final states, the readout power P_r must be of any value smaller than the amorphization power P_{am}. In this work, we consider $P_r = 10\%P_{am}$. Although this percentage could be larger, higher powers through the PCM may heat it up enough to cause thermo-optical effects, impacting readout resolution.

IV. Simulation Results

In this section we evaluate the proposed method, stochastic-oPCM, compared to amplitude-oPCM [9] in two applications compatible with the convolution circuit. The first is denoising through a standard averaging filter, and the second is the conversion to grayscale of an RGB image. The simulation

of stochastic-oPCM was carried out using the PCM model described in Fig. 2. In amplitude-oPCM, we assume that each nonvolatile input is correctly encoded using the same state-transmission characteristic from Fig. 1a.

A. PCM Model and Simulation Environment

We applied the proposed modelling flow on a GST patch with dimensions of 100 nm $\times$ 250 nm $\times$ 20 nm deposited on a 400 nm $\times$ 180 nm silicon waveguide. Using 500 ps pulses at $\lambda = 1550$ nm, this PCM was found to amorphize at powers above $P_{am} = 13.6$ mW, and so $E_{am} = 6.8$ pJ. For the curve-fitting step we assume that 64 levels are available ($n_b = 6$ bits), in line with the best achieved by the literature [4]. The obtained state-transmission characteristic is illustrated in the inset of Fig. 1a, with $T_0 = 0.86$ and $T_k = 0.99$. Based on these results, we determine that in stochastic-oPCM the readout step is performed with $P_\mathrm{r} = 10\% \times 13.6$ mW $= 1.36$ mW. In amplitude-oPCM, the amplitude equivalent to the highest operand (63) is encoded with P_r. This readout power is low enough to not accidentally melt the PCM, while high enough to provide up to 32 dB of signal-to-noise ratio at the output.

Then, we implemented simulated versions of the circuit in a Python environment, assuming no optical losses or phase mismatches, and output photodiodes with responsivity of 1 A/W. The only source of error was an added gaussian white noise to photodiode current, with standard deviation of 0.7 µA, equivalent to readout noise at approximately 1 GHz. In both methods, we avoid any post-processing effects by resorting to a look-up table that maps each output current during readout to the numerical operation result.

B. Denoising

In this evaluation, we consider an input grayscale image with dimensions 128×128 that was previously corrupted by gaussian noise, as seen in Fig. 5a. In this case, stochastic-oPCM had $S = 1$. The integer coefficients b of an averaging kernel are all equal and depend on its size M, such that

$$b = \mathrm{round}\left(\frac{2^{n_b} - 1}{M^2}\right). \tag{5}$$

For $M = 2$, we display the results for both methods in Fig. 5b, along with the output of an ideal digital convolution. It is noticeable that stochastic-oPCM gets similar results with respect to the ideal case, but amplitude-oPCM produces a noisier image. We also performed a sweep in the kernel size M, recording the peak signal-to-noise ratio (PNSR) achieved by all methods, as shown in Fig. 5c. In all cases, amplitude-oPCM fails to denoise the image, while stochastic-oPCM is close to the ideal filter until $M = 3$, with a PSNR 18.11 dB above the amplitude-oPCM (64.7$\times$ improvement). After $M = 3$, its behavior degrades, as more precision is needed to encode the kernel than achieved with 64 levels.

C. RGB-Grayscale Conversion

Standard RGB to grayscale conversion based on luminance requires that, for each pixel, we apply

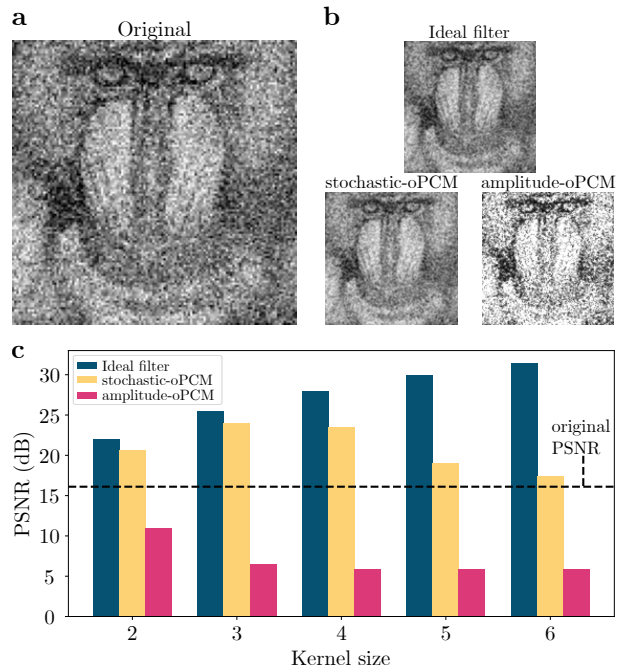

Fig. 5. (a) Input 128×128 image with added gaussian noise. (b) Results after after an averaging filter with $M = 2$ applied ideally, with stochastic-oPCM and with amplitude-oPCM. (c) PSNR results after changing the kernel size, with the dashed line representing the PSNR of the input image.

$$\mathrm{gray} = 0.2989R + 0.5870G + 0.1140B, \tag{6}$$

where R, G, and B are the color values of that pixel.

In this case, we convert the constants to 6-bit values and, for amplitude-oPCM, encode each in a PCM so one pixel is processed in a single step. Meanwhile, stochastic-oPCM has $S = 3$ and $M = 1$, such that the full matrix of each color is processed at a time and the grayscale pixels are accumulated at each PCM. We convert the same input RGB image seen in Fig. 6a, achieving the results from Fig. 6b-d.

Once again amplitude-oPCM is more affected by the output noise than stochastic-oPCM, such that our method achieves 9.8 more dB of PSNR, corresponding to an improvement of almost 10$\times$. In Fig. 6e-f we plot the absolute error of both methods with respect to the ideal case as heatmaps. It is possible to visualize that many features of the original image are corrupted by amplitude-oPCM, showing high error values, while they are better preserved by stochastic-oPCM.

D. Discussion

Our results show that, given the same conditions of operation, stochastic-oPCM leads to a more precise convolution when compared to amplitude-oPCM. The reason for it is that our method is less affected by output perturbations as verified in [11]. As the main source of error in stochastic-oPCM is the SC paradigm itself, it is expected to improve with higher bit counts, although other noise sources may become dominant. Additionally, our method uses sequences of high power pulses to perform each multiplication, thus it is expected to have a

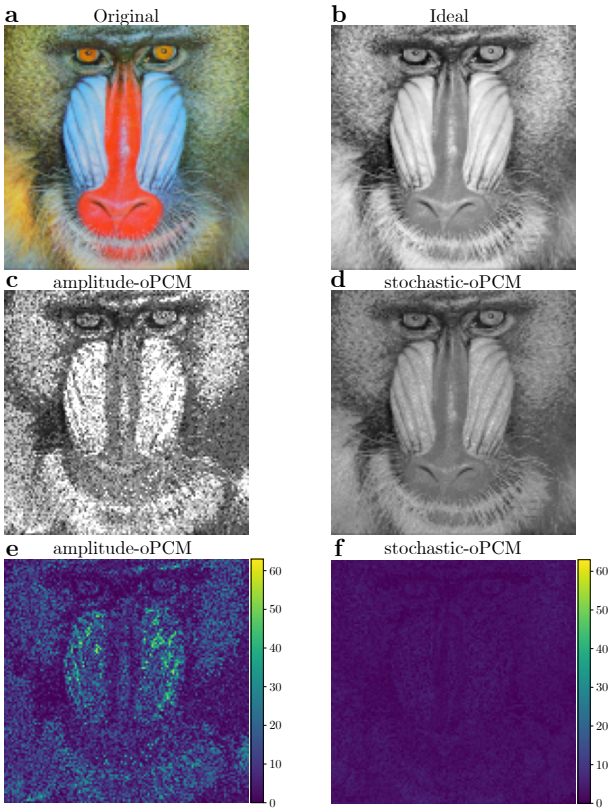

Fig. 6. (a) Original image, (b) standard conversion. Conversion with (c) amplitude-oPCM (16.8 dB PSNR), (d) stochastic-oPCM (26.6 dB PSNR). Absolute error (max pixel value = 63) between standard conversion and (e) amplitude-oPCM, (f) stochastic-oPCM.

larger energy overhead when compared to amplitude-oPCM. Nevertheless, at the end of every convolution operation, the resulting images are already stored in PCM memory. This fact allows the construction of different computing architectures in which it is not necessary to move data to memory after the operation, leading to potential advantages. In terms of speed, to avoid corrupting the written PCM state, the amplitude-oPCM method also has to wait for the thermal relaxation of the cell, expected to be on the order of nanoseconds [19], particularly if high input powers are involved. In our approach, the resting time must be respected between each pulse in a stochastic bitstream ($2^{n_b} - 1$ times). Therefore, implementations that leverage parallelization of stochastic-oPCM must be investigated.

V. Conclusions

This paper presents a novel photonic multiply-accumulate (MAC) operator using phase-change materials (PCMs). To the best of our knowledge, the proposed approach is the first to perform convolution combining the stochastic computing paradigm with optical PCMs. The synergy between the two allows the result of the operation to be stored directly in memory, and the non-volatility of the PCM allows one to read or reuse the results without timing constraints. This represents a new paradigm, in which the convolution is performed as a *memory write* instead of a *memory read*, common in other crossbar architectures. Results from denoising and RGB-grayscale conversion show that, despite the errors introduced by using stochastic computing, the proposed MAC operator still shows $64\times$ improvement in denoising and almost $10\times$ improvement in RGB-grayscale conversion in terms of PSNR. The perspectives of this work are numerous and include: experimental validation of the behavioral model, parallelization of the operator to improve its speed, and expansion of the proposed circuit to allow for signed operations.

References

[1] L. Bernstein *et al.*, "Freely scalable and reconfigurable optical hardware for deep learning," *Scientific reports*, vol. 11, no. 1, pp. 1–12, 2021.

[2] K. Guo *et al.*, "A survey of FPGA based neural network accelerator," *CoRR*, vol. abs/1712.08934, 2017.

[3] Y. Zhang *et al.*, "Myths and truths about optical phase change materials: A perspective," *Applied Physics Letters*, vol. 118, no. 21, p. 210501, 2021.

[4] 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.

[5] M. Stanisavljevic *et al.*, "Demonstration of reliable triple-level-cell (tlc) phase-change memory," in *2016 IEEE 8th International Memory Workshop (IMW)*, pp. 1–4, IEEE, 2016.

[6] C. Rios *et al.*, "Controlled switching of phase-change materials by evanescent-field coupling in integrated photonics," *Optical materials express*, vol. 8, no. 9, pp. 2455–2470, 2018.

[7] X. Li *et al.*, "On-chip phase change optical matrix multiplication core," in *2020 IEEE International Electron Devices Meeting (IEDM)*, pp. 7–5, IEEE, 2020.

[8] M. Miscuglio and V. J. Sorger, "Photonic tensor cores for machine learning," *Applied Physics Reviews*, vol. 7, no. 3, p. 031404, 2020.

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

[10] F. Brückerhoff-Plückelmann *et al.*, "Broadband photonic tensor core with integrated ultra-low crosstalk wavelength multiplexers," *Nanophotonics*, 2022.

[11] R. Cardoso *et al.*, "Towards a robust multiply-accumulate cell in photonics using phase-change materials," in *2023 Design, Automation and Test in Europe Conference (DATE)*, 2023.

[12] 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.

[13] M. H. Najafi *et al.*, "Performing stochastic computation deterministically," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 12, pp. 2925–2938, 2019.

[14] X. Li *et al.*, "Experimental investigation of silicon and silicon nitride platforms for phase-change photonic in-memory computing," *Optica*, vol. 7, no. 3, pp. 218–225, 2020.

[15] F. Ding *et al.*, "A review of compact modeling for phase change memory," *Journal of Semiconductors*, vol. 43, no. 2, p. 023101, 2022.

[16] J. Feldmann *et al.*, "Integrated 256 cell photonic phase-change memory with 512-bit capacity," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 2, pp. 1–7, 2019.

[17] A. Narayan *et al.*, "Architecting optically-controlled phase change memory," *arXiv preprint arXiv:2107.11516*, 2021.

[18] C. Rios *et al.*, "In-memory computing on a photonic platform," *Science advances*, vol. 5, no. 2, p. eaau5759, 2019.

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

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

[21] El-Derhalli *et al.*, "Stochastic computing with integrated optics," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1355–1360, IEEE, 2019.

[22] S. S. Vatsavai *et al.*, "Sconna: A stochastic computing based optical accelerator for ultra-fast, energy-efficient inference of integer-quantized cnns," *arXiv preprint arXiv:2302.07036*, 2023.