Mathematical-Write: a counterpoint of analog computing crossbars exemplified by photonics

Raphael Cardoso[1*], **Clément Zrounba**[1], **Mohab Abdalla**[12], **Paul Jimenez**[1], **Mauricio Gomes de Queiroz**[1], **Fabio Pavanello**[1], **Ian O'Connor**[1], **Sébastien Le Beux**[3]

1. Univ. Lyon - CNRS, Ecole Centrale de Lyon, INSA Lyon, Université Claude Bernard Lyon 1, CPE Lyon - INL, UMR5270 - Écully, F-69134, France

2. Integrated Photonics and Applications Centre (InPAC), School of Engineering, RMIT University, Melbourne, VIC 3000, Australia

3. Department of Electrical and Computer Engineering, Concordia University - Montreal, Canada

**raphael.cardoso@ec-lyon.fr*

Summary. In this work, we propose a new paradigm as alternative to analog computing crossbars for neuromorphic applications, in which the output result is written to memory as part of the operation. This leads to energy savings in terms of data movement, and further allows systems that integrate both learning (writing) and inferencing (reading) in the same circuit. Our results indicate that *mathematical-Write* can lead to a 33% energy reduction when performing convolution in memory.

As new AI models surpass the count of one billion parameters being processed by sequences of multiplications and accumulations (MACs), their energy consumption grows exponentially due to the costly data movement between processor and memory. Consequently, new analog computing architectures have been explored to accelerate and increase the energy efficiency of MACs while using non-volatile memory technologies (FeFET, PCM, ReRAM) [1]. Equivalents of electronic crossbars can also be seen in photonics, with its basic cell illustrated by Fig. 1a. In such cells, multiplication is performed by encoding one operand as the intensity of an optical pulse, while the other is the optical transmission through a phase-change material (PCM) [2]. This multiplication is analogous to a memory readout, where a pulse with fixed known power is used, and by verifying the output power it is possible to infer the value stored in memory. In an analog crossbar, the input power carries extra information, so the output will represent the product between input and stored values. Such method, to which we refer as *mathematical-Read* $(m - R)$, allows high-speed MACs with low energy consumption if the input stored in the non-volatile device is constant. However, if the operation results are to be used elsewhere, they will be written to RAM, spending more energy to move and write data.

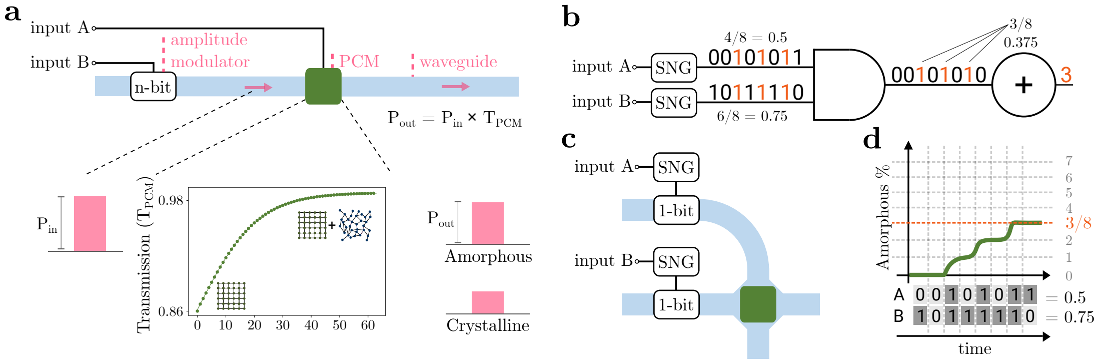
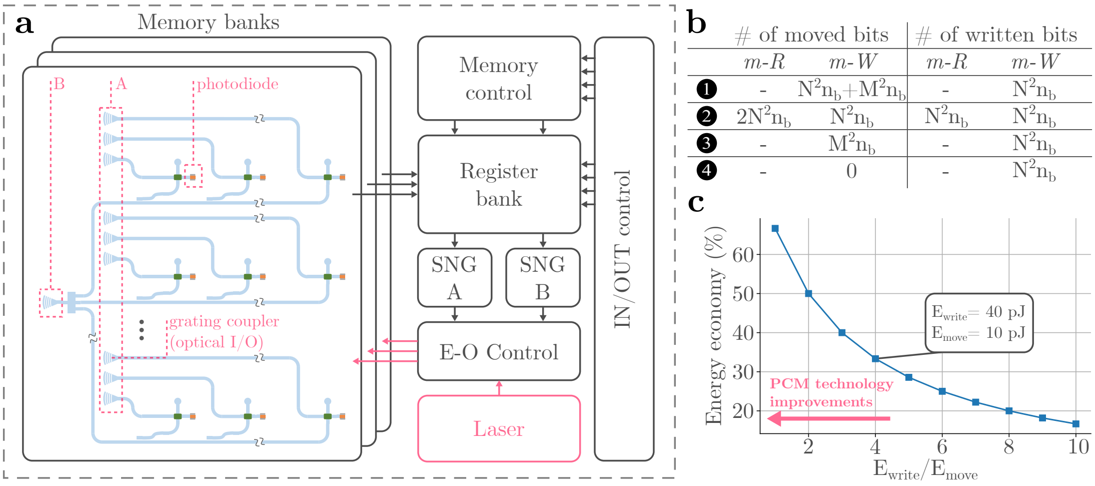


Figure 1: (a) *mathematical-Read* cell in photonics based on PCMs [2]; (b) primitive for stochastic multiplication, in which the output probability is the product of the input probabilities; (c) *mathematical-Write* cell in photonics based on PCMs [3]; (d) incremental accumulation between PCM state changes.

In this paper, we propose a different paradigm: *mathematical-Write* $(m - W)$, where both inputs are sent to a memory cell that performs the operation and writes the result at the same time. One possible implementation was proposed in [4], treating optical PCMs as an abacus. Alternatively, we recently proposed a cell based on stochastic computing [3], as illustrated in Fig. 1b-d. In this case, the optical pulses have only two levels (*low* and *high*), and the pulse power must be carefully chosen such that a state change will be triggered in the PCM only when two inputs are *high* simultaneously (logical AND). Furthermore, each new state change accumulates on the previous, as shown by Fig. 1d, for a maximum of $2^{n_b} - 1$ increments per product, where n_b is the bit encoding. This cell can be included in circuits capable of convolution, as shown in Fig. 2a.



	# of moved bits		# of written bits	
	m-R	*m-W*	*m-R*	*m-W*
❶	-	$N^2n_b+M^2n_b$	-	N^2n_b
❷	$2N^2n_b$	N^2n_b	N^2n_b	N^2n_b
❸	-	M^2n_b	-	N^2n_b
❹	-	0	-	N^2n_b

Figure 2: (a) Augmented optical PCM memory bank, capable of convolution between a matrix A and a kernel B; (b) number of bits moved and written for $m-R$ and $m-W$ as function of the operand sizes and bit encoding; (c) total energy saved by $m-W$ in its best case as the cost of write is reduced.

Our idea is therefore to augment optical PCM memory banks, such as the one in [5], with $m-W$, resulting in the architecture shown in Fig. 2a. To assess the possible gains of such architecture, we compare $m-W$ with $m-R$ in photonics. The number of data movements and memory writes in both methods is shown in Fig. 2b in a convolution between an $M \times M$ matrix and an $N \times N$ kernel using n_b bits encoding for different cases: ❶ none of the inputs are stored in memory; ❷ only the kernel is previously stored (this is always the case in $m-R$); ❸ only the input matrix is previously stored; ❹ both inputs are previously stored. A unique characteristic of $m-W$ is that it can bring the data movement between memory and processor (or accelerator) down to zero in case ❹.

After including the energy costs of E_{write} to write a bit and E_{move} to move a bit, using an encoding of $n_b = 4$, we realized that the total energy cost is independent of the operand sizes, depending only on the ratio between E_{move} and E_{write}. This analysis assumes: i) E_{move} and E_{write} dominate the energy consumption per bit; ii) both methods use the same underlying memory technology, iii) each increment in Fig. 1d costs $n_b E_{write}/(2^{n_b}-1)$, amounting to E_{write} per written bit of information, iv) the convolution result remains in $n_b = 4$ encoding. Under these assumptions, using realistic energy costs of $E_{write} = 40$ pJ [5] and $E_{move} = 10$ pJ, it is possible to achieve a 33% reduction in the best case (❹) for $m-W$ as shown in Fig. 2c. Even larger reductions are possible as E_{write} is improved from better PCM technology.

These results illustrate the biggest advantage of *mathematical-Write*: after performing the operation, there is no need to move the data and store it somewhere else, as it is automatically stored in the PCM cells. The proposed method also represents a change of paradigm, allowing the creation of new hybrid architectures that may be used in writing or reading mode as necessary, that could be useful for on-chip training. Lastly, the proposed strategy is not restricted to PCMs in photonics, but can also be applied to other non-volatile electronic technologies.

References

[1] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 14–26, 2016.

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

[3] 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. *In press.*

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

[5] A. Narayan, Y. Thonnart, P. Vivet, A. Coskun, and A. Joshi, "Architecting optically controlled phase change memory," *ACM Transactions on Architecture and Code Optimization*, vol. 19, no. 4, pp. 1–26, 2022.