Lightspeed Binary Neural Networks using Optical Phase-Change Materials

Taha Shahroodi[§], Raphael Cardoso[†], Mahdi Zahedi[§], Stephan Wong[§], Alberto Bosio[†], Ian O'Connor[†], Said Hamdioui[§]
[§]dept. Quantum and Computer Engineering, Delft University of Technology, Delft, The Netherlands
[†]Institut des Nanotechnologies de Lyon, École Centrale de Lyon, Lyon, France
Emails: {T.Shahroodi, M.Z.Zahedi, J.S.S.M.Wong, S.Hamdioui}@tudelft.nl, {raphael.cardoso, alberto.bosio, ian.oconnor}@ec-lyon.fr

***Abstract*—This paper investigates the potential of a compute-in-memory core based on optical Phase Change Materials (oPCMs) to speed up and reduce the energy consumption of the Matrix-Matrix-Multiplication operation. The paper also proposes a new data mapping for Binary Neural Networks (BNNs) tailored for our oPCM core. The preliminary results show a significant latency improvement irrespective of the evaluated network structure and size. The improvement varies from network to network and goes up to $\sim$1053$\times$.**

***Index Terms*—Binary Neural Network, Optical PCM, Computation-In-Memory**

I. Introduction

The significant progress of Neural Network (NN) applications has, unfortunately, come with extensive demand for system resources, e.g., memory and energy consumption. This high demand has already slowed down the growth of active, deployed NNs and their adoption when factoring in the costs. A recent research direction is to use simpler (operation-wise) and smaller NNs, such as BNNs. BNNs enjoy lower memory requirements, simplified arithmetic operations, and near SOTA accuracy on vision tasks [4, 14]. However, an optimized hardware implementation is still necessary for BNNs to smooth their cost-efficient adoption on our future systems.

A few works have investigated hardware realization of BNNs by introducing different mapping and data flow techniques [13, 17] or various circuitry and memristor-based crossbar structures (i.e., ReRAM or PCMs) to perform required operations [7]. Unfortunately, none of these methods exploit the full potential of underlying emerging devices for BNNs due to inefficient data mapping and the sequential nature of how they perform the necessary operations. These works are also constrained by the fact that the underlying computing cores are limited to a maximum of one single vector operation (either logical vector operation or VMM) at a given time step.

This paper presents three major contributions:

- The first hardware acceleration of BNNs utilizing oPCM instead of conventional electronic PCM.
- An efficient data flow on VMM-enabled crossbars that fully exploits the available readout circuitry and parallelism in a crossbar and particularly suits BNN operations.
- A detailed performance comparison between the proposed CMOS-compatible oPCM-based accelerator, an accelerator using the same mapping but with electronic PCM, and previous SOTA hardware accelerator for BNNs.

II. Background and Motivation

This section briefly touches on the necessary background. Please refer to [5, 6] for detailed information on these topics.

Computing Inside/Near Memory. Recent works from industry and academia [2, 3, 15] show that nanoscale emerging memory technologies are among the strongest candidates for the Computation-In-Memory (CIM) paradigm, where computation and memory units are co-located. This happens primarily due to the match between their compute capability and the computational need (i.e., VMM operation) of today's dominant data-intensive applications (i.e., NNs).

PCM-based Integrated Photonics. Phase Change Materials (PCMs for short) are currently the leading alternatives for non-volatile computation (VMM operation in particular) in silicon photonics-based platforms. This technology, commonly-known as integrated photonics based on PCM (oPCM), offers two main advantages compared to previous photonics-based platforms. First, CMOS-compatible manufacturing in contrast to diffractive computing in free-space optics [10] that cannot be integrated on-chip. Second, significantly higher speed and lower power requirements compared to interferometer-based multiplications that are CMOS-compatible but require large and power-hungry phase-shifters [16]. A CIM-enabled design exploiting oPCM also offers two additional benefits, distinguishing itself from alternative electronic emerging memories. First, oPCM allows high parallelization through wavelength division multiplexing (WDM), in which different vectors can be processed simultaneously in different parts of the frequency space. oPCM exploit WDM to effectively enable Matrix-Matrix-Multiplication (MMM), adding an extra scalability dimension to applications. Second, computing using oPCM avoids Joule heating, electromagnetic crosstalk, electronic addressing challenges for CIM-enabled designs, capacitance in custom silicon computing platforms, and resistance drift in electronic emerging memories [1, 8, 9, 11].

III. Proposal and Architecture

Fig. 1-(a) and Fig. 1-(b) present an overview of the concept and realized oPCM-based CIM-enabled accelerator with the recommended data flow. Shown in Fig. 1-(a), conceptually, this design utilizes XNOR and Popcount (❶) operations for execution of BNNs [14]. Our oPCM-based realization performs both XNOR and Popcount within 1 cycle using a VMM-enabeld crossbar by (1) re-writing XNOR as two AND and one OR operations ($A \odot B = A.B + \bar{A}\bar{B}$), and (2) re-distributing the kernels and their complements vertically in the crossbar (❷). This vertical mapping is compatible with all previous VMM-enabled crossbars (e.g., PCM-based ones) and removes the sequential, mostly digital, Popcount operation in SOTA [7]. The accelerator employs a resonator-based optical frequency comb (❸) to enable WDM. Exploiting WDM (❹), the oPCM-based accelerator provides an additional degree of computation

978-3-9819263-7-8/DATE23/© 2023 EDAA

to apply several inputs simultaneously, effectively enabling MMM. That is, oPCM performs K (3 in Fig. 1-(b)) VMM operations in one step instead of 1 VMM typically achieved by a memristor-based crossbar (❺). Current technology can comfortably support up to $K = 16$ [5].

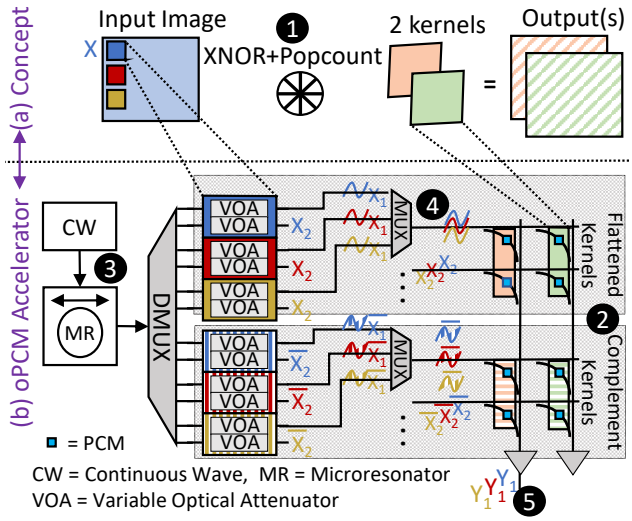

Fig. 1: Proposed Weight Mapping and oPCM Core with WDM.

IV. Evaluations

Evaluation Methodology. We build a cycle-accurate emulator using PyTorch derived from our device-aware extended circuits [18]. Due to space limitations, we only present the preliminary latency improvements here. We implemented two different configurations. (1) ePCM-Map that only considers the proposed mapping on electronic PCM-based crossbar. (2) oPCM-Core that considers the mapping but utilizes oPCMs. Our PCM and oPCM configurations are based on our previous results [5, 12]. We compare our designs against that of [7] as the SOTA hardware accelerator for BNNs (Baseline-ePCM). We evaluate all of the designs for 6 BNNs with various sizes from M1Bench [3] for handwritten digits recognition on MNIST.

Evaluation Results. Fig. 2 presents the latency improvement of proposals normalized to SOTA for the same underlying network. The y-axis uses a log-scale.

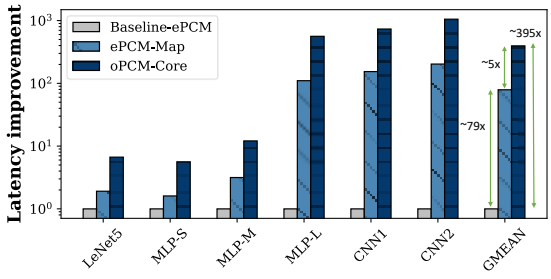

Fig. 2: Latency improvements over Baseline-ePCM.

We make four key observations.

- Both oPCM-Core and ePCM-Map improve the latency irrespective of BNN. This is because, unlike the Baseline-ePCM, these designs not only parallelize XNOR with Popcount but also parallelize both of these operations with many other sets via the proposed vertical data mapping.
- The latency improvement is network-dependent. This is directly related to the available parallelism in the operations of understudy BNN. In our BNNs, the larger the BNN is, the more parallel XNOR and Popcount operations exist. Improvements vary from ∼5.6× to ∼1053× for the evaluated BNNs.
- oPCM-Core brings on average ∼5× improvement in latency with the exact data flow compared to ePCM-Map. This happens due to the extra parallelism dimension enabled by WDM. Note that the improvement is also network-dependent.
- The highest improvement of our oPCM-Core over ePCM-Map (i.e., ∼5.18×) is still less than the available parallelism due to the new dimension via WDM (i.e., ∼16×). This is simply because the evaluated networks are not big enough to take full advantage of such an increase in the available parallelism. We will investigate very large networks using this design in our future work.

V. Conclusion

This paper proposes a CMOS-compatible oPCM-based core and an efficient data flow for BNNs. Our evaluations on latency and preliminary investigations on energy consumption suggest an enormous potential for oPCM-based cores in NNs. Hence, our work encourages further investigations of oPCM.

References

[1] S. R. Agrawal *et al.*, "A many-core architecture for in-memory data processing," in *MICRO*, 2017.

[2] A. Ankit *et al.*, "PUMA: A programmable ultra-efficient memristor-based accelerator for machine learning inference," in *ASPLOS*, 2019.

[3] P. Chi *et al.*, "PRIME: A novel processing-in-memory architecture for neural network computation in reram-based main memory," *ISCA*, 2016.

[4] M. Courbariaux *et al.*, "Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1," *arXiv:1602.02830*, 2016.

[5] J. Feldmann *et al.*, "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, 2021.

[6] S. Hamdioui *et al.*, "Memristor for computing: Myth or reality?" in *DATE*, 2017.

[7] T. Hirtzlin *et al.*, "Digital biologically plausible implementation of binarized neural networks with differential hafnium oxide resistive memory arrays," *Frontiers in neuroscience*, 2020.

[8] V. Joshi *et al.*, "Accurate deep neural network inference using computational phase-change memory," *Nature communications*, 2020.

[9] W. W. Koelmans *et al.*, "Projected phase-change memory devices," *Nature communications*, 2015.

[10] X. Lin *et al.*, "All-optical machine learning using diffractive deep neural networks," *Science*, 2018.

[11] D. A. Miller, "Attojoule optoelectronics for low-energy information processing and communications," *Journal of Lightwave Technology*, 2017.

[12] MNEMOSENE partners, "The MNEMOSENE project." http://www.mnemosene.eu/, accessed: 2022-06-02. 2020.

[13] Y.-F. Qin *et al.*, "Design of high robustness BNN inference accelerator based on binary memristors," *IEEE TED*, 2020.

[14] M. Rastegari *et al.*, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *ECCV*, 2016.

[15] T. Shahroodi *et al.*, "KrakenOnMem: a memristor-augmented HW/SW framework for taxonomic profiling," in *ICS*, 2022.

[16] A. N. Tait, "Quantifying Power in Silicon Photonic Neural Networks," *Physical Review Applied*, 2022.

[17] Y. Zhao *et al.*, "A Highly Robust Binary Neural Network Inference Accelerator Based on Binary Memristors," *Electronics*, 2021.

[18] C. Zrounba *et al.*, "Exploration of the optical behavior of phase-change materials integrated in silicon photonics platforms," in *Europe-EQEC*, 2021.