Hyperspectral In-Memory Computing

Mostafa Honari Latifpour,[1,2,†] Byoung Jun Park,[1,3,†] Yoshihisa Yamamoto,[1] and Myoung-Gyun Suh[1,*]

[1] *Physics & Informatics Laboratories, NTT Research, Inc., Sunnyvale, CA 94085, USA*
[2] *The Graduate Center, City University of New York, New York, NY 10016, USA*
[3] *KU-KIST Graduate School of Converging Science and Technology, Korea University, Seoul, Republic of Korea*
[†] *These authors contributed equally to this work.*
[*] *myoung-gyun.suh@ntt-research.com*

Abstract: We propose and demonstrate hyperspectral in-memory computing systems that harness both frequency and space dimensions, utilizing optical frequency combs and programmable optical memories. This approach offers the potential for energy-efficient optical information processing beyond PetaOPS-level performance.

1. Introduction

Advances in machine learning, exemplified by models like ChatGPT, have revolutionized various industries, driving demand for extensive matrix-vector multiplication (MVM) [1]. This growing computational need poses challenges for traditional von Neumann computing architectures, prompting researchers to explore alternative approaches like in-memory computing [2]. By incorporating non-volatile memory elements within processors, in-memory computing systems overcome bottlenecks arising from memory and processing unit separation. This results in more efficient data handling, reduced power consumption, and support for highly parallel computations. Numerous in-memory computing systems have been demonstrated using analog circuits based on resistive switching [3].

Simultaneously, there is renewed interest in utilizing optics for energy-efficient MVM due to optics' inherent suitability for parallel mathematical operations [4]. Various optical MVM systems have been proposed and demonstrated over the past few decades [5–7]. Among these, three-dimensional optical systems utilizing free-space optics hold significant potential, benefiting from scalable display, imaging, and camera technology. However, most of the systems demonstrated to date rely primarily on space multiplexing, leaving the frequency dimension largely untapped. Here, we propose and demonstrate a hyperspectral in-memory computing architecture that integrates space multiplexing with frequency multiplexing of optical frequency combs and uses spatial light modulators (SLMs) as a programmable optical memory, thereby boosting the computational throughput and the energy efficiency [8] (See Figure 1a). In our system, computations are primarily executed on a two-dimensional spatial light modulator, where the weight matrix either remains static or undergoes slow updates. This configuration enables optics to handle energy-efficient parallel data processing, while electronics ensure programmability.

2. Experimental Details

In the experiment, we demonstrated optical matrix-matrix multiplication (MMM) operations using hyperspectral encoding, where each SLM pixel encodes a matrix weight across multiple comb lines. This is essentially a batch processing of matrix-vector multiplication using wavelength-division multiplexing. We conducted numerous MMM tests, and the results aligned with theoretical predictions, including the multiplication of the NTT logo with the identity matrix, as depicted in Figure 1b.

As illustrated in Figure 1b, the input source is a fiber OFC in the optical C-band with a 250 MHz pulse repetition rate, coarsely filtered using line-by-line waveshaping [9]. The input optical source, with an average power of approximately 1 mW, is introduced into the system. The coarsely-filtered comb lines are spatially separated by a grating, fanned out vertically using a cylindrical lens, and then focused onto SLM 1, where the input matrix is encoded. Subsequently, the comb lines are recombined using additional gratings, fanned out horizontally using a cylindrical lens, and focused onto SLM 2, where the second matrix is encoded. After being fanned-in vertically using a cylindrical lens, the comb lines are vertically sorted by color using a grating, completing the hyperspectral multiply-accumulate operation. In this setup, a linear polarizer is used to operate the phase-only SLM for intensity modulation, and the system's non-uniformity is calibrated by adjusting the phase of the SLM pixels.

To illustrate the hyperspectral operation, we presented an example test with a hyperspectral factor of 5, where each SLM pixel encodes a matrix weight over 5 comb lines (see Figure 2a). With slight adjustments to our current

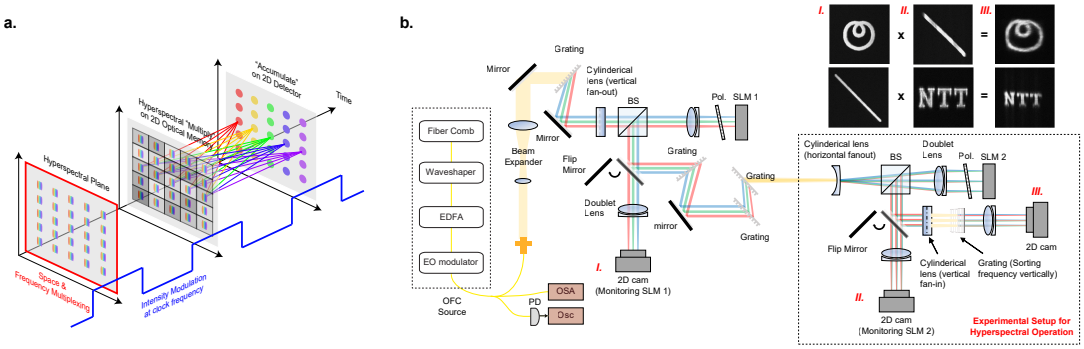

Fig. 1. **Hyperspectral In-Memory Computing.** (a) Concept: Boosting computational throughput by combining space and frequency multiplexing with intensity modulation at the computational clock frequency. (b) Experimental Setup: The setup outlined in the dotted box enables matrix-matrix multiplication (MMM) through the hyperspectral multiply-accumulate operation. Matrices are encoded onto SLM 1 and SLM 2, and the resultant matrix is captured by a 2D camera. As examples, encoding the NTT logo and identity matrix (I and II), each with an approximate size of 100 by 100, yields an output matrix displaying the NTT logo (III).

system, achieving a hyperspectral factor of 10 or even higher was possible (see Figure 2b). To assess the computational accuracy of our system, we analyzed the error distribution for every potential MAC value. Here, the matrix was encoded using only non-negative weights with 4 bits. We conducted 400 measurements for each target MAC value, ranging from 0 to 150 (see Figure 2c). As the target MAC values increased, the standard deviation of the error tended to grow, reaching a saturation point after a certain value (as shown in Figure 2d, top and middle panels). Accordingly, the relative error, defined as |measured MAC value - target MAC value|/(target MAC value), saw its standard deviation decrease to below 5 percent as the target MAC value rose. These errors can be attributed to factors like intensity fluctuations of the input OFC source, crosstalk between adjacent pixels, and optical alignment errors. We expect the standard deviation of the relative error to remain consistent at a similar level, even when the system operates with a larger matrix size or at higher bit encoding. It's worth noting that noise up to a certain threshold might not significantly impact the computational results. We verified this by analyzing the classification of MNIST data, a basic yet illustrative example, under different noise levels (see Figure 2e). Furthermore, in certain optimization tasks, noise can even help the system escape from a local optimum solution.

In our experiment, the system operated in an open-loop, where the encoding of the input matrix and the readout of output MAC results were executed independently using commercial digital electronic circuits. For high-throughput computation in open-loop operations, fast external modulation and readout are crucial. However, when the system operates in closed-loop mode with a desired nonlinear operation, it can serve as a physical solver for optimization tasks without the need for rapid external modulation and readout. Here, all computations occur in an analog mode, while only the initial problem loading and the final solution readout are managed digitally. This hybrid architecture merges the rapid processing of analog computing with the flexibility of a digital interface [10]. It's worth noting that for enabling fast pixel-by-pixel parallel modulation in the closed-loop system, it may be essential to introduce a novel device that would directly connect each photodetector pixel to its corresponding modulator pixel, eliminating the need for a camera and SLM connected via a serial bus. Such a direct link ensures continuous parallel processing for both data transfer and computation, avoiding the typical computational slowdowns caused by converting parallel information to a serial format.

3. Discussions and Conclusions

In the current work, we utilized optical sources and other components for the optical C-band due to the equipment's availability in our laboratory. If the demonstrated system operates at wavelength ranges below 1.1 microns, faster cameras with a larger number of pixels become readily available thanks to advanced CMOS technology. Moreover, shorter wavelengths and diffraction-limited optics will enhance the alignment precision and system resolution.

Looking ahead to the near-term system designed to function in a closed-loop setting, improvements such as better alignment and wider spectral bandwidth that allows for larger matrix sizes (300 x 300), a greater hyperspectral factor (x30), and increased modulation speed (1 GHz) are expected to enable a performance of up to 2.7 PetaOPS with an estimated power efficiency of 10.26 W/PetaOPS. In the long run, with a hyperspectral enhancement of x100 and a matrix size of 1000 x 1000, we project that the future system could achieve 100 PetaOPS with a

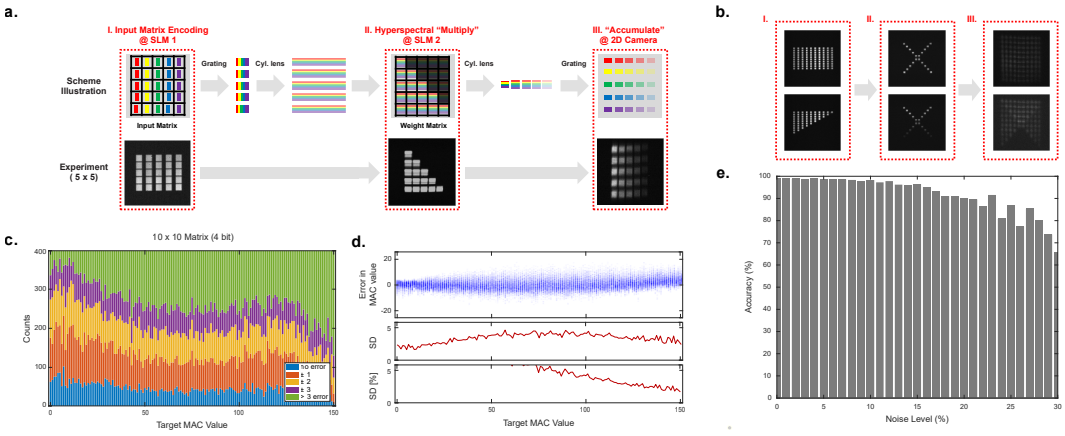

Fig. 2. **Hyperspectral Multiply-Accumulate** (a) Illustrations of MMM performed through hyperspectral operations are paired with experimental images of a 5-by-5 matrix captured by a 2D camera. (b) Additional test experiments were conducted using 10-by-10 matrices. (c) Error distribution for each possible MAC value is shown. For each MAC value, 400 MAC operations are performed for analysis. (d) Absolute error at each target MAC value is shown in the top panel, the standard deviation (SD) of the error distribution in the middle panel, and the error as a percentage in the bottom panel. (e) The classification accuracy of MNIST data is assessed under various noise levels.

power efficiency close to 2 W/PetaOPS. This represents a two-orders-of-magnitude efficiency boost compared to state-of-the-art electronic GPUs.

Our hyperspectral in-memory computing design combines frequency, space, and time dimensions, prioritizing scalability and leveraging advancements in SLM and OFC technologies [11, 12]. The modular approach ensures adaptability, and incorporation of technologies like metalenses [13] and chip-integrated OFCs [14] suggests potential for miniaturization. As components continue to evolve, this system could reshape energy-efficient optical information processing, offering performance beyond the PetaOPS level in future cloud computing.

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436–444 (2015).
2. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. & Eleftheriou, E. Memory devices and applications for in-memory computing. *Nature nanotechnology* **15**, 529–544 (2020).
3. Ielmini, D. & Wong, H.-S. P. In-memory computing with resistive switching devices. *Nature electronics* **1**, 333–343 (2018).
4. Caulfield, H. J. & Dolev, S. Why future supercomputing requires optics. *Nature Photonics* **4**, 261–263 (2010).
5. Feldmann, J. *et al.* Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
6. Wang, T. *et al.* An optical neural network using less than 1 photon per multiplication. *Nature Communications* **13**, 123 (2022).
7. Zhou, T. *et al.* Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nature Photonics* **15**, 367–373 (2021).
8. Latifpour, M. H., Park, B. J., Yamamoto, Y. & Suh, M.-G. Hyperspectral In-Memory Computing with Optical Frequency Combs and Programmable Optical Memories. *arXiv preprint arXiv:2310.11014* (2023).
9. Weiner, A. M. Femtosecond pulse shaping using spatial light modulators. *Review of scientific instruments* **71**, 1929–1960 (2000).
10. Mourgias-Alexandris, G. *et al.* Analog Iterative Machine (AIM): using light to solve quadratic optimization problems with mixed variables. *arXiv preprint arXiv:2304.12594* (2023).
11. Panuski, C. L. *et al.* A full degree-of-freedom spatiotemporal light modulator. *Nature Photonics* **16**, 834–842 (2022).
12. Diddams, S. A., Vahala, K. & Udem, T. Optical frequency combs: Coherently uniting the electromagnetic spectrum. *Science* **369**, eaay3676 (2020).
13. Chen, W. T. *et al.* A broadband achromatic metalens for focusing and imaging in the visible. *Nature nanotechnology* **13**, 220–226 (2018).
14. Xiang, C. *et al.* Laser soliton microcombs heterogeneously integrated on silicon. *Science* **373**, 99–103 (2021).