

Predictive Performance of Photonic SRAM-based In-Memory Computing for Tensor Decomposition

Sasindu Wijeratne[†], Sugeet Sunder^{*}, Md Abdullah-Al Kaiser[‡], Akhilesh Jaiswal[‡]
Clynn Mathew^{*}, Ajey P. Jacob^{*}, Viktor Prasanna[†]

[†]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California

^{*}Information Sciences Institute (ISI), University of Southern California

[‡]Electrical and Computer Engineering, University of Wisconsin-Madison

Email: {kangaram, prasanna}@usc.edu, {sunder, cmathew, ajey}@isi.edu, {mkaiser8, akhilesh.jaiswal}@wisc.edu

Abstract—Photronics-based in-memory computing systems have demonstrated a significant speedup over traditional transistor-based systems because of their ultra-fast operating frequencies and high data bandwidths. Photonic static random access memory (pSRAM) is a crucial component for achieving the objective of ultra-fast photonic in-memory computing systems. In this work, we model and evaluate the performance of a novel photonic SRAM array architecture in development. Additionally, we examine hyperspectral operation through wavelength division multiplexing (WDM) to enhance the throughput of the pSRAM array. We map Matricized Tensor Times Khatri-Rao Product (MTTKRP), a computational kernel commonly used in tensor decomposition, to the proposed pSRAM array architecture. We also develop a predictive performance model to estimate the sustained performance of different configurations of the pSRAM array. Using the predictive performance model, we demonstrate that the pSRAM array achieves 17 PetaOps while performing MTTKRP in a practical hardware configuration.

Index Terms—Photonic Computing, MTTKRP, Tensor Decomposition

I. INTRODUCTION

Recent advancements in analyzing large datasets have led to information being inherently represented as higher-order data structures known as tensors. Tensor decomposition converts input tensors into a reduced latent space, which can then be utilized to identify important features of the underlying data distribution. Tensor decomposition has been effectively used in various fields, such as machine learning, signal processing, and network analysis [1], [2], [3]. Additionally, tensor decomposition has been instrumental in improving the interpretability of complex models by breaking down multi-dimensional data into simpler, more manageable components. This decomposition technique has also facilitated advancements in areas such as bioinformatics [4], where it aids in the analysis of multi-modal biological data, and in computer vision [5], where it enhances image and video processing tasks. The flexibility and robustness of tensor decomposition methods continue to drive innovation across a wide range of scientific and engineering disciplines.

Canonical Polyadic Decomposition (CPD) is arguably the most widely used method to decompose a tensor into a low-rank tensor decomposition model [6], [7]. It has become the standard tool for unsupervised multi-way data analysis. The Matricized Tensor Times Khatri-Rao Product (MTTKRP) [8]

is recognized as the most time-consuming computational kernel in CPD. Due to the irregular shapes of the real-world tensors, specialized hardware accelerators are increasingly popular to enhance the efficiency of sparse tensor computations

In recent years, digital electronics have significantly advanced the power and performance metrics of digital computing systems [9]. The 6-transistor electrical SRAM has become the standard for on-chip memory storage. However, these electrical SRAMs fall short in terms of computing speed, throughput, and power efficiency for new data-intensive applications such as machine learning, signal processing, and large-scale simulations [10]. The rapid increase in the energy delay product associated with data movement and memory access for compute operations underscores the memory bottleneck that affects modern digital computing systems [9]. Solutions involving parallel, near-memory, and in-memory processors based on electronic physical-state variables do not completely address the issue due to constraints related to scalar computing and the latency of long metal interconnections.

Recently, photonic SRAMs have emerged as a promising alternative to electronic charge-based SRAMs [11]. Numerous photonic SRAM implementations have been investigated in the past [12], [13], [14], [15], [16], [17], [18], [19]. However, developing a photonic SRAM technology that is compatible with current foundry manufacturing processes and offers ultra-high speed and low energy consumption remains a significant challenge.

In this work, we introduce a novel photonic SRAM array embedded in a scalable optical in-memory compute engine, designed using existing foundry processes. The photonic SRAM uses available GF45SPCLO photodiodes and ring resonators, which are upgrades over our previous designs [11] and utilized models of photonic devices reported in the literature [20], [21]. Our approach combines the high-speed and bandwidth advantages of photonic technology with the proven reliability of SRAM while addressing the challenges of integrating optical components with standard CMOS processes to create a scalable and efficient in-memory computing solution.

The contributions of our paper are as follows:

- A novel embedded photonic SRAM (pSRAM) array is designed and implemented in a scalable optical in-memory

computing engine, operating in the O-band. The pSRAM is reconfigurable at speeds exceeding 20 GHz; the write speed of the RAM. The overall performance of the pSRAM array is determined by the speed of the optical components that constitute the system architecture.

- We map the compute primitives of MTTKRP to the pSRAM array architecture.
- We develop a predictive performance model to evaluate the sustained performance of the proposed photonic SRAM memory architecture on MTTKRP.
- The predictive performance model shows that the pSRAM array can achieve a sustained performance of 17 PetaOps with 8-bit precision under practical pSRAM array configuration.

II. BACKGROUND

A tensor is a generalization of an array in multiple dimensions. In TD, the number of dimensions of an input tensor is commonly called the number of tensor modes. A real-valued N mode tensor is denoted by $\mathcal{X} \in \mathbb{R}^{I_0 \times \dots \times I_{N-1}}$. Further, $\mathcal{X}_{(n)}$ denotes the mode- n matricization or the unfolding [22] of the matrix \mathcal{X} . $\mathcal{X}_{(n)}$ is defined as the matrix $\mathcal{X}_{(n)} \in \mathbb{R}^{I_n \times (I_0 \dots I_{n-1} I_{n+1} \dots I_{N-1})}$ where parenthetical ordering indicates that the mode- n column vectors are arranged by sweeping all the other mode indices through their ranges.

Canonical Polyadic Decomposition (CPD) decomposes \mathcal{X} into a sum of single-mode tensors (i.e., arrays), which best approximates \mathcal{X} . For example, given 3-mode tensor $\mathcal{X} \in \mathbb{R}^{I_0 \times I_1 \times I_2}$, our goal is to approximate the original tensor as

$$\mathcal{X} \approx \sum_{r=0}^{R-1} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \quad (1)$$

where R is a positive integer and $\mathbf{a}_r \in \mathbb{R}^{I_0}$, $\mathbf{b}_r \in \mathbb{R}^{I_1}$, and $\mathbf{c}_r \in \mathbb{R}^{I_2}$. For a thorough review of CPD, refer to [6].

In the rest of Section II, we assume that the number of modes is three for illustration purposes.

Algorithm 1: CP-ALS for a 3-mode tensor

- 1 Input: A tensor $\mathcal{X} \in \mathbb{R}^{I_0 \times I_1 \times I_2}$, the rank $R \in \mathbb{Z}^+$
 - 2 Output: CP decomposition $[[\mathbf{A}, \mathbf{B}, \mathbf{C}]]$, $\mathbf{A} \in \mathbb{R}^{I_0 \times R}$, $\mathbf{B} \in \mathbb{R}^{I_1 \times R}$, $\mathbf{C} \in \mathbb{R}^{I_2 \times R}$
 - 3 **while** stopping criterion not met **do**
 - 4 // Matricization of \mathcal{X} is different for each factor matrix computation
 - 5 $\mathbf{A} \leftarrow \text{spMTTKRP}(\mathcal{X}_{(0)}, \mathbf{B}, \mathbf{C})$
 - 6 $\mathbf{B} \leftarrow \text{spMTTKRP}(\mathcal{X}_{(1)}, \mathbf{A}, \mathbf{C})$
 - 7 $\mathbf{C} \leftarrow \text{spMTTKRP}(\mathcal{X}_{(2)}, \mathbf{A}, \mathbf{B})$
 - 8 Normalize $\mathbf{A}, \mathbf{B}, \mathbf{C}$
-

The alternating least squares (ALS) method is used to compute CPD. Algorithm 1 shows the ALS method for CPD (i.e., CP-ALS) where Matricized Tensor-Times Khatri-Rao product (MTTKRP) is iteratively performed on all the Matricizations of \mathcal{X} , iteratively. In this paper, performing MTTKRP on all the Matricizations of an input tensor is called computing

MTTKRP along all the modes. The outputs \mathbf{A} , \mathbf{B} , and \mathbf{C} are the factor matrices that approximate \mathcal{X} . \mathbf{a}_r , \mathbf{b}_r , and \mathbf{c}_r in Equation 1 refers to the r^{th} column of \mathbf{A} , \mathbf{B} , and \mathbf{C} , respectively.

III. OPTICAL SRAM ARCHITECTURE

A. Input Encoding:

One significant advantage of employing photonic devices in computing architectures is leveraging hyperspectral computing using wavelength division multiplexing (WDM) [23]. WDM allows a single optical channel to carry multiple data streams by simultaneously transmitting signals at different wavelengths, maximizing data transmission capacity without interference. Optical frequency combs, generated using microresonators, produce a precise series of narrow spectral lines (comb lines) spanning a broad range of wavelengths [24].

The proposed device operates in the O-band, offering 52 wavelength channels (based on Global Foundries 45SPCLO PDK) with sub-nanometer spacing for efficient data transmission. In this system, we envision an intensity encoded input data, with each discrete power level corresponding to a specific value represented by an 8-bit word. To modulate multiple wavelength channels simultaneously with varying intensity levels, we employ comb shapers—optical devices designed to manipulate the spectral properties of an optical frequency comb. High-speed Electro-optic modulators are a common type of comb shapers that selectively attenuate or enhance specific comb lines, allowing for precise shaping of the comb spectrum for various applications.

B. Bitcell:

Conventional electrical SRAMs face significant speed and power consumption bottlenecks due to the large bitline/wordline capacitance and high interconnect resistance due to technology scaling [25]. In contrast, our proposed photonic SRAM (pSRAM) exhibits ultra-low energy consumption and high-speed read/write operation [11]. To construct the optical latch structure, the pSRAM bitcell employs cross-coupled microring resonators (MRR) and photodiodes (PD) illustrated in Figure 1. The through port of MRR R1 (R2) drives the photodiode P2 (P1), which controls the resonance state of the other MRR R2 (R1). Hence, the cross-coupled structure ensures the storing of differential optical data inside the latch. The pSRAM bitcell is projected to operate at a 20 GHz frequency while consuming ~ 1.04 pJ/bit (~ 16.7 aJ/bit) switching (static) energy [11]. The proposed pSRAM is designed as a 2D crossbar array of memory bitcells, with each cell connected to a pair of bit lines and a word line. Typically, a word line is one word and controls the activation of read/write operations for the corresponding bitcells. While the read speed of pSRAM is faster, constrained by the time constant of ring resonators. The write speed of pSRAM is currently at 20 GHz, which determines the reconfigurability rate of pSRAM.

Furthermore, the fabrication-friendly architecture of our pSRAM subsystem, which is based on a pSRAM prototype designed on GF 45SPCLO PDK, has been sent to the fab for

tapeout, enabling seamless integration alongside the electrical subsystem.

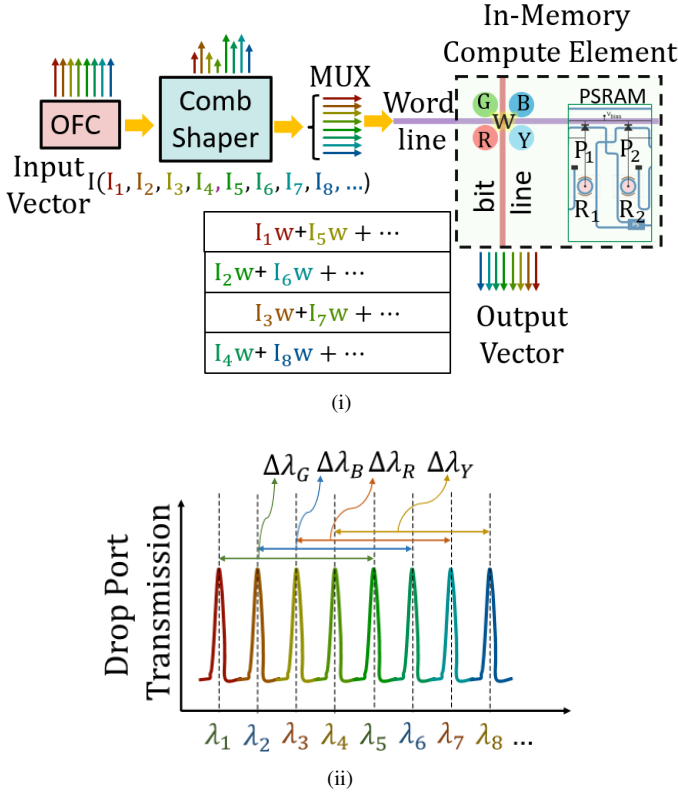


Fig. 1: (i) Schematic of the proposed computing engine. Optical frequency combs (OFC) are used to generate precise wavelength channels, which are then modulated using high-speed comb-shapers. The input, encoded across multiple independent wavelength channels, is sent into the word-line and multiplied with a memory bit. Different ring modulators (G/B/R/Y) are employed to handle different sets of wavelengths, with the resonances of other three resonators spaced within the FSR of the one. An analog output is received on the bit-line for further processing. (ii) The drop port transmission characteristics of the compute ring modulators indicates the spacing of wavelength channels used for WDM.

C. Output Encoding:

As bitcell can only store binary data, appropriate intensity scaling depending on the bit-position is required for the compute operation. The dot product of the 8-bit intensity encoded input and a 8-bit binary word stored in the pSRAM array results in 8 analog optical outputs. Each optical output obtained is inherently scaled according to its corresponding bit significance, with the maximum optical power delivered to the bit representing the most significant digit (MSB) and appropriately scaled power sent into the least significant bit (LSB). These optical outputs are converted and accumulated through the photocurrents of the photodetectors. The analog accumulated photocurrent values can be converted into digital

electrical bitstreams through high-speed on-chip analog-to-digital converters (ADC). Integrating on-chip ADC facilitates the seamless conversion of analog optical signals to digital electrical form, which enables on-chip CMOS hardware/accelerator for further processing in the electrical domain.

IV. MAPPING MTTKRP COMPUTATIONAL PRIMITIVES TO PSRAM ARRAY

A. Grid Representation of pSRAM Array

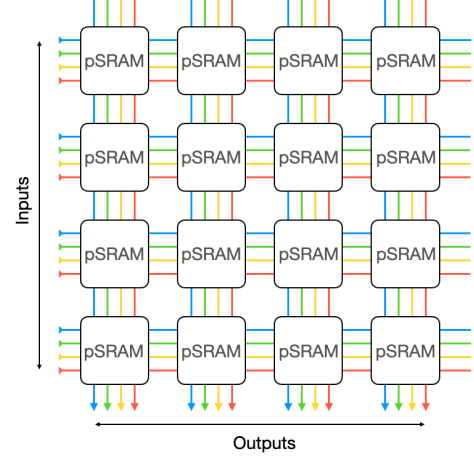


Fig. 2: Grid representation of pSRAM array.

Figure 2 illustrates the 2D grid representation of the pSRAM array. In this figure, each pSRAM word (group of pSRAM cells) is shown as a square, while a wavelength is shown as a line. Since the design supports hyperspectral encoding, different color lines in Figure 2 represent different wavelengths. For demonstration purposes, Figure 2 displays a 2D grid containing 4×4 pSRAM words and 4 distinct wavelengths. Each pSRAM is capable of multiplying the values stored within the word by the inputs from the wavelengths. The addition is conducted along each column by summing the intensity of identical wavelengths.

B. Computational Primitives

We identified 3 computational primitives (CPs) that the pSRAM array should support to perform MTTKRP.

In this Section, we describe the computational primitives using a 3-mode tensor \mathcal{X} and its factor matrices \mathbf{A} , \mathbf{B} , and \mathbf{C} . We consider an example in which MTTKRP is performed on the tensor \mathcal{X} to generate the factor matrix \mathbf{A} . Note that the computational primitives introduced in this Section can be extended to any tensor with any number of modes.

C. Hadamard Product of Factor Matrix Rows (CP 1)

This primitive involves computing the Hadamard Product, which is an elementwise multiplication of corresponding elements of two vectors. For example, given the rows of the factor matrix b_j and c_k , the Hadamard product is represented as $b_j \circ c_k$.

As illustrated in Figure 3, a row \mathbf{b}_i (i.e., $i = 0, 1, 2, \dots$) of the factor matrix \mathbf{B} is loaded and stored in each column

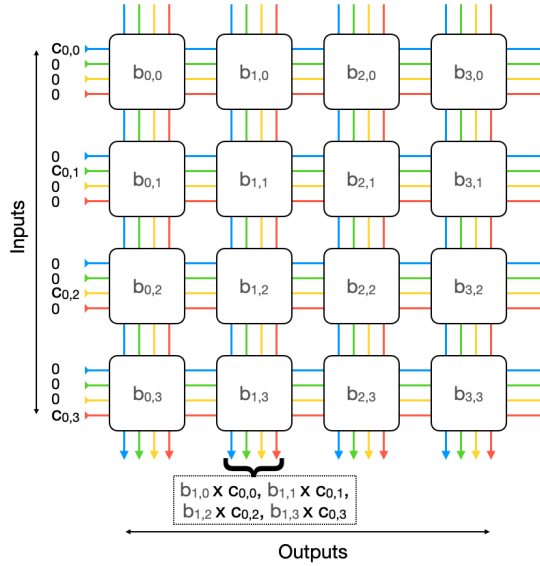


Fig. 3: Mapping CP 1 to pSRAM array.

of the pSRAM array. Subsequently, each row of the factor matrix \mathbf{C} (denoted as c_i) is loaded, and each element in c_i is multiplied by the corresponding element in b_i . We interleave the wavelengths while feeding the inputs to avoid addition among column values, as shown in Figure 3. The resulting product from each column of the pSRAM array is the Hadamard Product of the respective rows of the factor matrices \mathbf{B} and \mathbf{C} . Figure 3 shows only the output from a single column of the pSRAM array. Note that all columns in the grid generate outputs simultaneously.

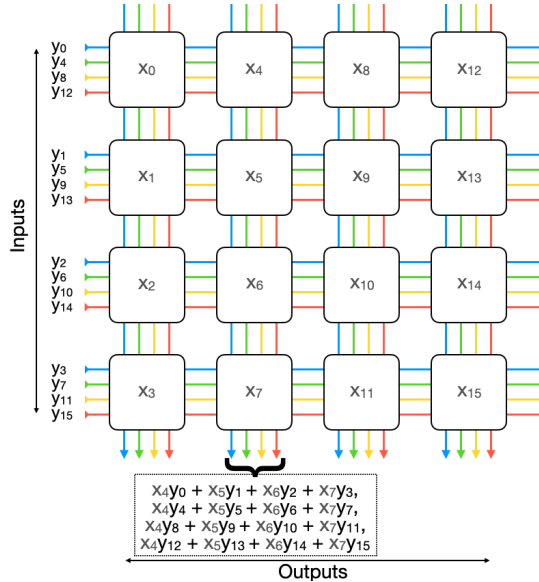


Fig. 4: Mapping CP2 and CP3 to pSRAM array.

D. Scaling with a Tensor Element (CP 2)

Using the Hadamard Product (results from CP 1), the second computational primitive multiplies the resulting vector by the respective tensor element x_i . Formally, this can be written as $x_i \cdot (B_{j_0} \circ C_{k_0})$.

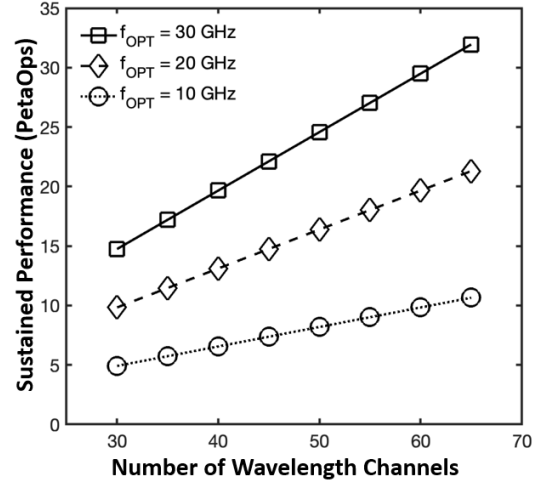
E. Elementwise Vector Addition to Generate the Factor Matrix (CP 3)

The third primitive adds the scaled vectors produced in CP 2 to the corresponding row of the factor matrix, \mathbf{A} through vector addition. This operation is given by $A_{i_0} + x \cdot (B_{j_0} \circ C_{k_0})$, where A_{i_0} is a row from the factor matrix \mathbf{A} .

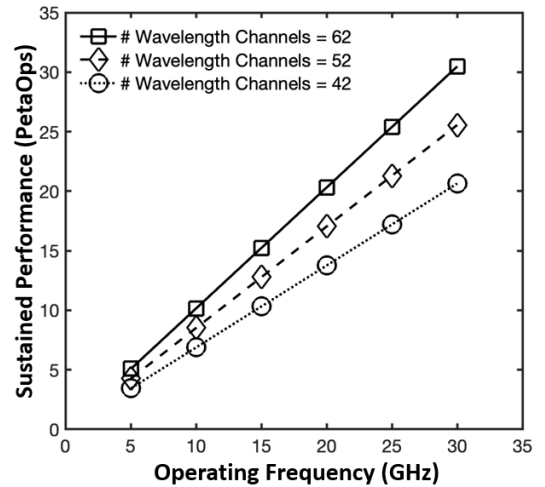
As shown in Figure 4, CP 2 and CP 3 are mapped to pSRAM to produce the final output of $A_{i_0} + x_i \cdot (B_{j_0} \circ C_{k_0})$. Tensor elements (indicated as x_i) are loaded and stored inside the pSRAM words, and Hadamard products of the matrix columns of factor matrices (shown as y_i) are loaded as input using different wavelengths. Similar to Figure 3, a single output of a column of the pSRAM array is shown in Figure 4.

V. EVALUATION

A. Experiments Setup



(i)



(ii)

Fig. 5: (i) Impact of wavelength channels. (ii) Impact of operating frequency.

The proposed pSRAM array has 256×256 bits. In each row of the pSRAM array, 8 bits are collected together as a word to support 8-bit precision, creating an array of 256×32 words.

Using performance modeling, we evaluated the sustained performance of the pSRAM array while executing MTTKRP on very large tensors (e.g., a 3-mode dense tensor with 1 million indices in each mode).

B. Overall Performance

Through hardware simulations and performance modeling, we identified that the operating frequency and the number of wavelength channels are the most critical hardware parameters that impact the sustained performance of the proposed architecture. As shown in Figure 5, the sustained performance of MTTKRP on the proposed pSRAM array linearly increases as the operating frequency and the number of wavelength channels increase. Our initial hardware implementation and simulation results show that the proposed pSRAM array can support 52 wavelength channels while operating at 20 GHz. Under these conditions, the proposed optical array achieves a sustained performance of 17 PetaOps while performing MTTKRP.

VI. CONCLUSION

In this work, we evaluate the performance of a novel photonic SRAM in-memory compute array architecture employing a predictive performance model. We mapped the compute primitives of MTTKRP to the pSRAM array architecture and demonstrated that the architecture can achieve 17 PetaOps in a practical hardware configuration. This work demonstrates the usefulness of photonic in-memory scalar and hyperspectral computing systems that can accelerate complex data-intensive tasks such as MTTKRP.

ACKNOWLEDGEMENT

This work is supported by DARPA under grant N660012424003 and the National Science Foundation (NSF) under grant CNS-2009057.

REFERENCES

- [1] M. Mondelli and A. Montanari, "On the connection between learning two-layer neural networks and tensor decomposition," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1051–1060.
- [2] Z. Cheng, B. Li, Y. Fan, and Y. Bao, "A novel rank selection scheme in tensor ring decomposition based on reinforcement learning for deep neural networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3292–3296.
- [3] F. Wen, H. C. So, and H. Wymeersch, "Tensor decomposition-based beamspace esprit algorithm for multidimensional harmonic retrieval," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4572–4576.
- [4] Y. Taguchi, "Tensor decomposition based unsupervised feature extraction applied to bioinformatics," in *Application of Omics, AI and Blockchain in Bioinformatics Research*. World Scientific, 2020, pp. 159–187.
- [5] Y. Panagakos, J. Kossaifi, G. G. Chrysos, J. Oldfield, M. A. Nicolaou, A. Anandkumar, and S. Zafeiriou, "Tensor methods in computer vision and deep learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 863–890, 2021.

- [6] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.
- [7] K. S. Aggour, *Intelligent and Scalable Algorithms for Canonical Polyadic Decomposition*. Rensselaer Polytechnic Institute, 2019.
- [8] I. Nisa, J. Li, A. Sukumaran-Rajam, R. Vuduc, and P. Sadayappan, "Load-balanced sparse mttkrp on gpus," in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2019, pp. 123–133.
- [9] H. N. Khan, D. A. Hounshell, and E. R. Fuchs, "Science and research policy at the end of moore's law," *Nature Electronics*, vol. 1, no. 1, pp. 14–21, 2018.
- [10] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, 2017.
- [11] R. Kudalippallyalil, S. Chandran, A. P. Jacob, and A. Jaiswal, "Towards scalable, energy-efficient and ultra-fast optical sram," *arXiv preprint arXiv:2111.13682*, 2021.
- [12] N. Pleros, D. Apostolopoulos, D. Petrantonakis, C. Stamatidis, and H. Avramopoulos, "Optical static RAM cell," *IEEE Photonics Technology Letters*, vol. 21, no. 2, pp. 73–75, 2008.
- [13] A. Tsakyridis, T. Alexoudi, A. Miliou, N. Pleros, and C. Vagionas, "10 Gb/s optical random access memory (RAM) cell," *Optics Letters*, vol. 44, no. 7, pp. 1821–1824, 2019.
- [14] B. Dong, H. Cai, Y. Gu, Z. Yang, Y. Jin, Y. Hao, D. Kwong, and A. Liu, "Nano-optomechanical static random access memory (SRAM)," in *2015 28th IEEE International Conference on Micro Electro Mechanical Systems (MEMS)*. IEEE, 2015, pp. 49–52.
- [15] B. Li, M. I. Memon, G. Mezosi, Z. Wang, M. Sorel, and S. Yu, "Optical static random access memory cell using an integrated semiconductor ring laser," in *2009 International Conference on Photonics in Switching*. IEEE, 2009, pp. 1–2.
- [16] T. Alexoudi, D. Fitsios, A. Bazin, P. Monnier, R. Raj, A. Miliou, G. T. Kanellos, N. Pleros, and F. Raineri, "III–V-on-Si photonic crystal nanocavity laser technology for optical static random access memories," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 22, no. 6, pp. 295–304, 2016.
- [17] S. Pitris, C. Vagionas, T. Tekin, R. Broeke, G. Kanellos, and N. Pleros, "WDM-enabled optical RAM at 5 Gb/s using a monolithic inp flip-flop chip," *IEEE Photonics Journal*, vol. 8, no. 2, pp. 1–7, 2016.
- [18] Y. Liu, R. McDougall, M. Hill, G. Maxwell, S. Zhang, R. Harmon, F. Huijskens, L. Rivers, H. Dorren, and A. Poustie, "Packaged and hybrid integrated all-optical flip-flop memory," *Electronics Letters*, vol. 42, no. 24, pp. 1399–1400, 2006.
- [19] A. Trita, G. Mezosi, M. Zanola, M. Sorel, P. Ghelfi, A. Bogoni, and G. Giuliani, "Monolithic all-optical set-reset flip-flop operating at 10 Gb/s," *IEEE Photonics Technology Letters*, vol. 25, no. 24, pp. 2408–2411, 2013.
- [20] A. Jacob and A. Jaiswal, "Non-volatile electro-optical high-bandwidth ultra-fast large-scale memory architecture," Jan. 18 2024, uS Patent App. 18/030,380.
- [21] A. P. Jacob, A. R. Jaiswal, R. Kudalippallyalil, and S. Chandran, "Electro-optical high bandwidth ultrafast differential ram," May 23 2024, uS Patent App. 18/281,662.
- [22] G. Favier and A. L. de Almeida, "Overview of constrained parafac models," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, pp. 1–25, 2014.
- [23] X. Xu, W. Han, M. Tan, Y. Sun, Y. Li, J. Wu, R. Morandotti, A. Mitchell, K. Xu, and D. J. Moss, "Neuromorphic computing based on wavelength-division multiplexing," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 29, no. 2: Optical Computing, pp. 1–12, 2022.
- [24] L. Chang, S. Liu, and J. E. Bowers, "Integrated optical frequency comb technologies," *Nature Photonics*, vol. 16, no. 2, pp. 95–108, 2022.
- [25] K. Cho, H. Choi, I. J. Jung, J. Oh, T. W. Oh, K. Kim, G. Kim, T. Choi, C. Sim, T. Song *et al.*, "Sram write-and performance-assist cells for reducing interconnect resistance effects increased with technology scaling," *IEEE Journal of Solid-State Circuits*, vol. 57, no. 4, pp. 1039–1048, 2022.