Programmable Photonic Neural Networks for advanced Machine Learning tasks

Angelina Totović,[1,2,*] Apostolos Tsakyridis,[2] George Giamougiannis,[2] Miltiadis Moralis-Pegios,[2] Anastasios Tefas[2] and Nikos Pleros[2]

[1] *Celestial AI, 3001 Tasman Drive, Santa Clara, CA 95054, USA*
[2] *Department of Informatics, Center for Interdisciplinary Research & Innovation, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece*

[*] *atotovic@celestial.ai*

Abstract: Photonics holds the promise of reshaping Machine Learning and High-Performance Computing hardware landscape, stripping it of unnecessary signal conversion overhead, complying with strict power dissipation envelopes while unlocking unrivaled compute and bandwidth capacity.

1. Cloud hardware ecosystem of tomorrow

While still discovering new applications of Artificial Intelligence (AI) and Machine Learning (ML), enabling services that were unimaginable a few decades ago - from Natural Language Processing and Generation (NLP/NLG) [1], autonomous vehicles, medical diagnostics, to predictive modeling and recommendation engines [2, 3], the underlying models are exploding in terms of parameter count (hitting 1 trillion in Meta's DLRM22 [3]), data volume, bandwidth requirements and complexity (reaching 10^9 petaFLOPs for large-scale model training [4]). Long relied-upon workhorses of compute - CPUs and GPUs, combined with pluggable optics for intra- and inter-Data Center Interconnection (DCI) do not suffice anymore. Increasing compute demand translates into growing hardware, making Processing Units (xPUs, x = C, G, T) reticle-limited and enforcing transition from monolithic chips to chiplets and multi-chip modules, calling for improved interconnection bandwidth. At the same time, compute resource underutilization and limited beachfront for the I/O requires a more flexible interconnection fabric, ready to support resource disaggregation [5], which would also aid in relieving power density challenges, in the view of recent total dissipated power projections of kW per socket [3]. The main takeaway from the current state of AI/ML and High-Performance Computing (HPC) hardware is that sub-systems cannot be designed/optimized independently from software - or one another - anymore; the whole system needs to be treated as a unique entity, following lean design principles, carefully pruning any excess operation in the chain from software down to chip.

Figure 1 depicts a vision of a future cloud ecosystem, relying on three paradigms: (*i*) software-hardware co-design, (*ii*) heterogeneous computing, and (*iii*) memory and compute resource disaggregation. Both the hosts and the chiplets require decentralized interconnection approach, based on the connectivity fabric [3, 5], which, if implemented optically, could offer higher bandwidth, lower latency, and, if co-packaged, significantly improved power efficiency. Chiplets themselves should support heterogeneous computing [6], having a variety of processing, memory and control units, tailored to a particular class of tasks. Technology will also most likely be diverse, without enforcing one-fits-all solution, but harnessing the benefits of each platform depending on the task at hand. Finally, the software will need to be backward compatible with old platforms and support implementation of pre-

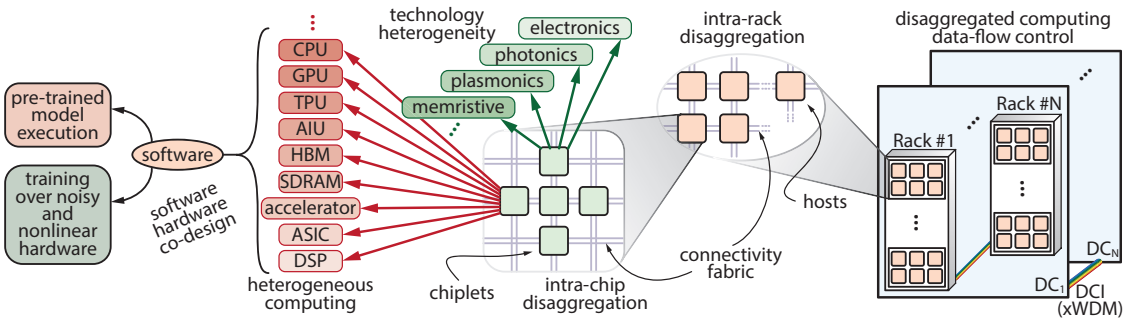


Fig. 1. Abstraction of a future hardware-software ecosystem for AI/ML/HPC.

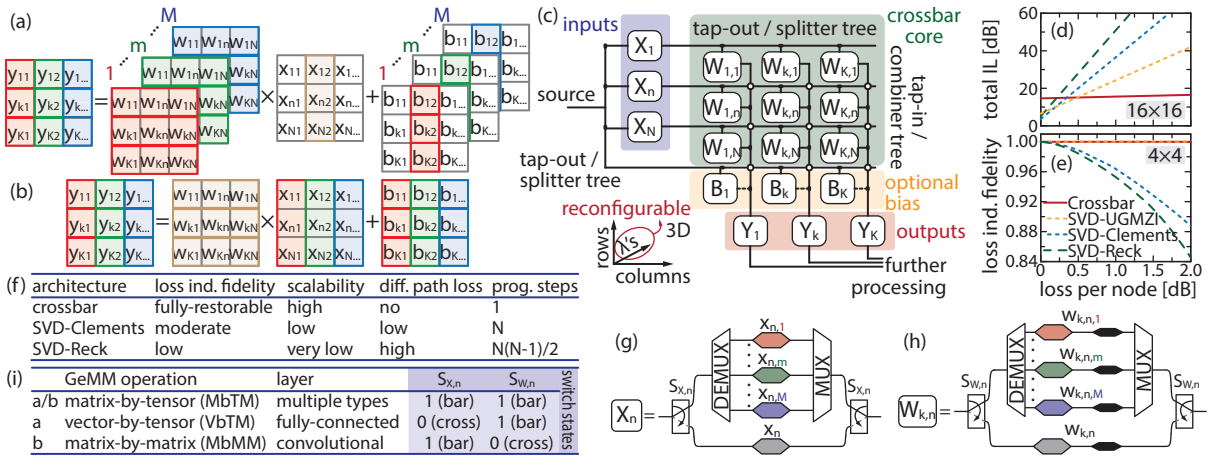
trained models, minimizing the energy and time cost of retraining. At the same time, new models should be trained directly on the hardware they will be executed on. In case of photonics, noisy, nonlinear and analog signal nature will imply lower bit precision, but embracing this may result in more resilient models, as demonstrated in [7, 8].

2. Programmable Photonic GeMM engines

Introducing photonic compute hardware into hosts could alleviate excessive O/E/O conversion in fetch-compute-transmit processes, while supporting large-scale General Matrix Multiplication (GeMM) operations, frequently encountered in AI/ML models that are particularly "heavy" for the electronic xPUs. Modern ML models, such as Deep Learning Recommendation Models (DLRMs), leverage both large embedding tables for tackling sparse data and Multi-Layer Perceptrons (MLPs) [2], the latter being a sequence of Fully-Connected Layers (FCLs) interleaved with activations, which can be represented by Vector-by-Tensor Multiplication (VbTM), Fig. 2(a), [9]. Similarly, NLPs/NLGs, which are transformer-type networks, rely on encoder-decoder blocks (a multitude of FCLs, calling for VbTM) with scaled dot-product attention implemented via Matrix-by-Matrix Multiplication (MbMM), Fig. 2(b), for increased efficiency [1]. Finally, Convolutional Neural Networks (CNNs) still remain the standard in image classification, video processing and object tracking [10]. If approached to in parallel fashion (filtering multiple input vectors by the same kernel), convolutional layer can be represented by MbMM [9].

The consensus among academic community shows that the most suitable paradigm for arbitrary matrix multiplication in photonic platform is the crossbar [12–17], schematically depicted in Fig. 2(c). Comparing to its most-well-known competitor in coherent domain, based on Singular Value Decomposition (SVD) and Mach-Zehnder Interferometer (MZI)-mesh unitary matrix implementation, crossbar offers numerous benefits for arbitrary matrices, summarized in Fig. 2(f) [11, 12], falling behind only in the special case of unitary matrices, where MZI-mesh power conservation features come into the spotlight, making it suitable for quantum applications [18], but not quite as much for AI/ML ones. The main advantage of the crossbar is bijective mapping of weight elements to hardware nodes, which reduces the programming steps to 1, drives down the total Insertion Loss (IL), Fig. 2(d), and almost diminishes differential path loss, enables straightforward fault detection and, more importantly, full loss-induced fidelity restoration, Fig. 2(e) [12]. Lower total IL and full fidelity restoration open the possibility of migrating from ultra-low-loss node weighting technologies, mandatory in SVD application, to lossy ones that offer higher bandwidth and unlock the possibility not only for multi-10 GHz inference [7, 8], but also training over photonic hardware [13]. Yet another benefit of lower overall IL is the high scalability potential, where photonic crossbar engine now becomes limited by available laser power and reticle size instead of fidelity and loss.

As long as the size of the matrix does not exceed the size of the crossbar, it supports time-of-flight multiplication in a single compute step; otherwise, tiled-matrix multiplication can be adopted [19]. Going a step further in supporting higher dimensionality calls for introducing another degree of freedom, with wavelength being the most straightforward choice for the coherent crossbar [9, 20]. The true potential of multichannel operation can be seen not when both input and weight are λ-selective, but when one of the two is, while the other uses common modulator for all channels, executing VbTM, Fig. 2(a), or MbMM, Fig. 2(b). Moreover, offering two possible routes



(f)

architecture	loss ind. fidelity	scalability	diff. path loss	prog. steps
crossbar	fully-restorable	high	no	1
SVD-Clements	moderate	low	low	N
SVD-Reck	low	very low	high	N(N-1)/2

(i)

	GeMM operation	layer	$S_{X,n}$	$S_{W,n}$	switch states
a/b	matrix-by-tensor (MbTM)	multiple types	1 (bar)	1 (bar)	
a	vector-by-tensor (VbTM)	fully-connected	0 (cross)	1 (bar)	
b	matrix-by-matrix (MbMM)	convolutional	1 (bar)	0 (cross)	

Fig. 2. (a) Vector-by-Tensor and (b) Matrix-by-Matrix multiplication, both with optional bias. (c) Programmable WDM-enhanced coherent photonic crossbar [9], with its (d) total IL and (e) loss induced fidelity dependence on loss per node and (f) performance summary [11, 12]. (g,h) Reconfigurable (g) input and (h) weight banks producing the GeMM/layer as per table (i).

for the signal (λ-selective and common), that can be reconfigured on-demand, Fig. 2(g), (h), seamlessly supports switching between the most common GeMM operations, as summarized in Fig. 2(i), accelerates computation and brings significant power-savings [9].

Programmable crossbar easily turns into a Programmable Photonic NN (PPNN) by connecting one layer's outputs to the next layer's inputs [20]. As it does not rely on photodetection for summation, it is particularly suitable for all-optical activations, eliminating O/E/O conversion. The layers can be placed on physically separate circuits, or can be a part of a single crossbar, where its GeMM mode of operation would be reconfigured from one time-step to another. Using PPNN in a hybrid software-photonic NN, where the last two layers were implemented in photonics (using VPI Transmission Maker™) forming a fully-convolutional network, has shown accuracy degradation of only 2% comparing to software-only in MNIST digit parity identification, and offered guidelines for choosing the optimal input/weight resolution (down to 4 bits) and tolerable modulator extinction ratio ($\sim 8-10$ dB) [20].

3. Conclusion

The exciting times we are in are inclusive of a multitude of technologies, demand co-design on different platforms, a tight software-hardware development framework and heterogeneous approach to compute and data movement, allowing for the most efficient technologies to be gradually filtered out and implemented as native in future systems [6]. Photonics certainly has a bright future in advanced ML hardware, both in connectivity and compute domains - there is undoubtedly many more useful system features that photonics can enable or advance. Some of them have already been demonstrated, such as physical layer DDoS attack identification [19], real-time analog signal processing [21] etc., and some are yet to be envisioned.

References

1. A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems,* vol. 30 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc., 2017).
2. M. Naumov *et al.*, "Deep learning recommendation model for personalization and recommendation systems," arXiv preprint arXiv:1906.00091 (2019).
3. A. Bjorlin, "Infrastructure for large scale AI: Empowering open," OCP Global Summit (2022).
4. J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos, "Compute trends across three eras of machine learning," in *2022 International Joint Conference on Neural Networks (IJCNN),* (2022), pp. 1–8.
5. M. Glick *et al.*, "PINE: Photonic integrated networked energy efficient datacenters (ENLITENED program)," J. Opt. Commun. Netw. **12**, 443–456 (2020).
6. M. Zahran, "Heterogeneous computing: Here to stay," Commun. ACM **60**, 42–45 (2017).
7. G. Mourgias-Alexandris *et al.*, "Noise-resilient and high-speed deep learning with coherent silicon photonics," Nat. Commun. **13**, 5572 (2022).
8. M. Moralis-Pegios *et al.*, "Neuromorphic silicon photonics and hardware-aware deep learning for high-speed inference," J. Light. Technol. **40**, 3243–3254 (2022).
9. A. Totovic, G. Giamougiannis, A. Tsakyridis, D. Lazovsky, and N. Pleros, "Programmable photonic neural networks combining WDM with coherent linear optics," Sci. Reports **12**, 5605 (2022).
10. J. Gu *et al.*, "Recent advances in convolutional neural networks," Pattern Recognit. **77**, 354–377 (2018).
11. G. Giamougiannis *et al.*, "Universal linear optics revisited: New perspectives for neuromorphic computing with silicon photonics," IEEE J. Sel. Top. Quantum Electron. (2022). Submitted.
12. G. Giamougiannis, A. Tsakyridis, Y. Ma, A. Totovic, D. Lazovsky, and N. Pleros, "Coherent photonic crossbar as a universal linear operator," arXiv preprint arXiv:2208.12033 (2022).
13. A. Tsakyridis *et al.*, "Universal linear optics for ultra-fast neuromorphic silicon photonics towards fJ/MAC and TMAC/sec/mm^2 engines," IEEE J. Sel. Top. Quantum Electron. **28**, 1–15 (2022).
14. S. Xu *et al.*, "Parallel optical coherent dot-product architecture for large-scale matrix multiplication with compatibility for diverse phase shifters," Opt. Express **30**, 42057–42068 (2022).
15. R. Tang *et al.*, "Two-layer integrated photonic architectures with multiport photodetectors for high-fidelity and energy-efficient matrix multiplications," Opt. Express **30**, 33940–33954 (2022).
16. M. van Niekerk *et al.*, "Massively scalable wavelength diverse integrated photonic linear neuron," Neuromorphic Comput. Eng. **2**, 034012 (2022).
17. N. Youngblood, "Coherent photonic crossbar arrays for large-scale matrix-matrix multiplication," IEEE J. Sel. Top. Quantum Electron. pp. 1–1 (2022).
18. C. Taballione *et al.*, "20-mode universal quantum photonic processor," arXiv preprint arXiv:2203.01801 (2022).
19. A. Tsakyridis *et al.*, "Neuromorphic silicon photonics with 50 GHz tiled matrix multiplication for deep learning applications," Adv. Photonics (2022). Submitted.
20. A. Totovic *et al.*, "WDM equipped universal linear optics for programmable neuromorphic photonic processors," Neuromorphic Comput. Eng. **2**, 024010 (2022).
21. C. Huang *et al.*, "A silicon photonic–electronic neural network for fibre nonlinearity compensation," Nat. Electron. **4**, 837–844 (2021).