nature photonics

Supplementary information

https://doi.org/10.1038/s41566-026-01987-z

Photonic accelerators for AI across cloud and edge platforms

In the format provided by the authors and unedited

Photonic Accelerators for AI Across Cloud and Edge Platforms

Supplementary Information

Dmitry Nechipurenko, Abhishek Gautam, Maxim R. Shcherbakov*

Department of Electrical Engineering and Computer Science,
University of California, Irvine, CA 92697, USA

* E-mail: maxim.shcherbakov@uci.edu

Supplementary Note 1. Optical Energy Overheads in Single-pass Architectures

In coherent MZM meshes, such as Clements and Reck meshes, the optical depth increases with matrix size as O(N), so the insertion loss and phase noise accumulate with N, reducing the received optical power per channel and collapsing bit depth. Even with an optimistic power budget per chip, with current MZM losses, an optical mesh with $N \gtrsim 30$ is fundamentally unachievable.

In contrast, crossbar-like architectures [S1] maintain fixed optical depth per channel at O(1), so scaling is primarily constrained by (1) optical power distribution and (2) detector/trans-impedance amplifier (TIA) noise rather than cascading photonic loss.

In case of silicon-based components, the system with $N = 1000$ inputs (and 10^6 modulators for weight encoding) and a realistic set of parameters (symbol refresh frequency $f_s = 10$ GS/s, detector $\mathrm{NEP_{Ge}} = 5 \times 10^{-12}\ \mathrm{W}/\sqrt{\mathrm{Hz}}$ [S2], and bit depth at the detector $n = 5$ bits leading to $\mathrm{SNR_{dB}} \approx 6.02n + 1.76$), will have the following optical power budget: $P_{\mathrm{optical}} = \mathrm{NEP_{Ge}} \cdot \sqrt{\frac{f_s}{2}} \cdot N \cdot 10^{0.1(\mathrm{SNR_{dB}} + \mathrm{P_{PIC\,loss}})} \approx 35$ W, in case of a total optical loss $\mathrm{P_{PIC\,loss}} = 20$ dB from laser to integrated Ge detector.

As the throughput for such architecture scales as $N^2 \cdot f_s$, the optical energy efficiency is ~ $10^4/35$ = 285 TMAC/W. Considering the plug energy efficiency of the optical power source (~40% for the most efficient sources), the total energy budget per MAC for such a system would be at least ~10 fJ/MAC (ADC/DAC overheads neglected). Hence, the optical source overhead per MAC scales as $10^{2+0.1\mathrm{P_{PIC\,loss}}}/N$ fJ/MAC.

While various approaches could be used to mitigate optical losses in such a system (see below), further decreasing the energy budget per MAC by increasing N seems infeasible due to the quadratic scaling of the modulator components in the system.

To alleviate the existing power-efficiency overheads in single-pass architectures, excess free-carrier losses and nonlinearities can be mitigated by transitioning to field-effect modulators on thin-film lithium niobate (TFLN) [S3] or, potentially, barium strontium titanate [S4]. Passive optical losses, such as in edge couplers, splitters, and multiplexers, can be reduced to sub-dB levels. This can be achieved through inverse-designed components [S5] and metamaterial-assisted couplers [S6], which have shown said efficiency limits while maintaining compact footprints.

Supplementary Note 2: PA Energy Consumption and Its Scaling

A typical general-purpose PA comprises several power-consuming modules, including light sources, DACs for vector and weight updates, and photodetectors, followed by ADCs. The total energy consumption of a PA-based system depends on both photonic and electronic components (including ADC, DAC, memory, and CPU). In most of the published studies, the energy breakdown is typically provided for the PA part of the system, and depending on the exact

architecture, the scaling laws for the energy efficiency and hence the minimal number of required components for reaching a given performance level might be quite different, as shown below.

The reported photonic platforms aimed at performing general MVM and GMMs can be divided into two major classes, with the first (which we term hypermultiplexing platforms) relying on time-domain-encoded vector inputs that can be batched in parallel using WDM. In these systems, vectors and weights are constantly streamed in time; the multiplication is performed at the modulator [S7] or the detector [S8], while the summation operation is performed using photocurrent integration in the analog domain. Such systems have much lower ADC overhead (due to low-frequency readout) and can exploit low light levels (and hence minimal power-source overhead) through signal accumulation. Such systems generally have a fixed optical depth on a chip, and their overall energy consumption per MAC is usually dominated by the DAC costs, scaling as:

$$\alpha \propto E_{\mathrm{DAC}}/n + \beta E_{\mathrm{ADC}},$$

where n is the batch/WDM scale, i.e., the number of independent vectors that are processed simultaneously with given weights, E_{DAC} is the typical energy required for a single DAC operation, which is of the order of 500–1000 fJ for 8-bit encodings at 20 GHz [S7], and $\beta \ll 1$ when input vector dimension $N \gg 1$, which is usually the case. In this expression, we neglected the energy required to drive the light sources as it is significantly smaller than the energy required for weight encodings [S7], which scales as $E = E_{DAC} N_s N$, where N_s is the number of distinct spatial components used to apply weights to the data input (typically $N_s = n < N$). Such a photonic tensor core operating with $n \sim 100$ frequencies and a single array of $N_s \sim 100$ MZM components for weight application can implement high-performance GMMs with a low budget of $\alpha \sim 10$ fJ/MAC, which is formally an order-of-magnitude improvement over the state-of-the-art NVidia GB200 ($\alpha \sim 200$ fJ/MAC).

The second type of the architecture the we term a single-pass system takes the full advantage of spatial parallelism: in this case the MVM operation is performed during the passage of the vector with N components through N different spatial channels (e.g., waveguides) and MVM is performed in one go with N^2 modulators: while weights are usually applied in the optical domain [S1], the summation part can be also performed in the analog electronic domain after photoconversion. The energy budget of such systems is usually dominated by the laser power and high-frequency ADCs, and the energy per MAC in the case of the constant optical depth systems [S1] generally scales as:

$$\alpha = E_{\mathrm{ADC}}/N + E_{\mathrm{source}}/N + \gamma E_{\mathrm{DAC}},$$

where γ is the ratio of the weights update frequencies and vector update frequency (can be downscaled, for example, in case batches of inputs are sequentially delivered). Importantly, due to high-frequency readout, even in case of very low weight update frequencies the power overhead of such systems is dominated by optical energy overheads (required to reach sufficient SNR at high detection frequency): even for $N = 1000$ components, α is dominated by E_{source}/N, which in

case of 10 GHz readout is ~ 10 fJ/MAC (see Supplementary Note 1 for details). Importantly, such a system will require ~10^6 modulators and, while having a rather high throughput of ~10 POPS/s, will have the same energy efficiency as a hypermultiplexing platform built with just 100 modulators that performs GMM using 100 wavelengths (and having ~100 TMAC/s of throughput).

One of the possible strategies to further mitigate power overheads could be based on mixed architectures, for example, wherein time-encoded and frequency-multiplexed inputs are converted (using high-speed routing and delay lines) into spatially multiplexed pulses delivered into a weights module (with a lower frequency update rate) – and further detected using low-frequency ADC modules.

Supplementary Note 3: On PA-assisted Training

AI training is a much more demanding task than inference due to memory-intensive workloads, the enormous amount of total computation required, a greater number of optimized algorithms, and sensitive numerical stability, amongst other requirements. Practical training requires memory bandwidth of 500 GB/s – 2 TB/s per device (as training is a memory-bound workload) and numerical (mixed) precision of at least FP8/FP16, as higher numerical stability is needed for optimal convergence of training algorithms. The total computation required to train most AI models lies in the range of 10^{21} – 10^{27} FLOPs, depending on model size. To achieve such computation in a reasonable time, cluster sizes typically range from 10^2 to 10^5 GPUs.

As an early example, the Meta Llama 65B model was trained on a cluster of 2048 Nvidia A100 GPUs on 1.4T training tokens, taking a total of 21 days. The A100s support a bandwidth of 1.56 TB/s and precision of up to 64 bits [S10].

Among these requirements, memory-intensive workloads that demand high memory bandwidth (TB/s) within a GPU and high numerical precision represent the most critical roadblocks for future PAs to address.

Some of these fundamental challenges see a path forward, such as numerical precision, memory usage, and data efficiency. Low-precision training is an active area of research, with the potential to achieve end-to-end 8-bit and 4-bit training pipelines. This can be achieved with further advances in quantization-aware training, novel quantization functions, and scaling methods [S11]. Traditional backpropagation requires memory that scales as O(no. of parameters + no. of activations). New training paradigms, such as the forward-forward algorithm, significantly mitigate the memory-intensive nature of training by optimizing local layer-wise objectives rather than a global loss propagated through the entire network, thereby reducing the need to store all model activations [S12]. Similarly, physics-aware training incorporates prior knowledge of governing physical laws into the learning objective, reducing reliance on labelled datasets. By augmenting data-driven losses with physics-based constraints, such methods improve data

efficiency and promote physically consistent predictions, particularly in scientific and engineering domains where data is limited but underlying dynamics are well understood [S13].

Whether highly efficient co-packaged PAs could be practical for substantially reducing energy overheads during training of large AI models needs careful analysis and simulations of hybrid architectures.

Supplementary References

S1. Hua, S., Divita, E., Yu, S. et al. An integrated large-scale photonic accelerator with ultralow latency. *Nature* **640**, 361–367 (2025).

S2. Sorianello et al. Germanium photodetectors for Si photonics. *J. Lightwave Tech*. (2018).

S3. Lin, Z., Shastri, B.J., Yu, S. *et al.* 120 GOPS Photonic tensor core in thin-film lithium niobate for inference and in situ training. *Nat Commun* 15, 9081 (2024).

S4. Abel, S., Eltes, F., Ortmann, J.E. *et al.* Large Pockels effect in micro- and nanostructured barium titanate integrated on silicon. *Nature Mater* 18, 42–47 (2019)

S5. MacLellan, B., Roztocki, P., Belleville, J., Romero Cortés, L., Ruscitti, K. *et al.* Inverse design of photonic systems. *Laser Photonics Rev.* **18**, 2300500 (2024).

S6. Chen, R., Chang, Y. K., Zhuang, Z. P., Liu, Y. & Chen, W. J. *et al.* Metasurface-based fiber-to-chip multiplexing coupler. *Adv. Opt. Mater***. 11**, 2202317 (2023)

S7. Ou, S., Xue, K., Zhou, L., Lee, C. H., Sludds, A. *et al.* Hypermultiplexed integrated photonics–based optical tensor processor. *Sci. Adv.* **11**, eadu0228 (2025).

S8. Chen, Z., Sludds, A., Davis, R. et al. Deep learning with coherent VCSEL neural networks. *Nat. Photonics* **17**, 723–730 (2023).

S9. Peccerillo B., Mannino M., Mondelli A. & Bartolini S. A survey on hardware accelerators: Taxonomy, trends, challenges, and perspectives. *J. Syst. Archit.* 129, 102561 (2022).

S10. Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M.-A., Lacroix T., Rozière B., Goyal N., Hambro E., Azhar F., Rodriguez A., Joulin A., Grave E. & Lample G. LLaMA: Open and efficient foundation language models. *arXiv 2302.13971* (2023).

S11. Marion B., Liu D. & Khoury A. The case for training large models in low precision. *TechRxiv* DOI: 10.36227/techrxiv.175493457.73187131/v1 (2025).

S12. Hinton, G. The forward-forward algorithm: Some preliminary investigations. *arXiv* 2212.13345 (2022).

S13. Wright, L. G. *et al.* Deep physical neural networks trained with backpropagation. *Nature* 601, 549–555 (2022).

Supplementary Table 1: Accelerator speed vs energy efficiency. References are provided per the main text. GPU — graphics processing unit, PTC — photonic tensor core, FS — free space, MAC — multiply and accumulate, INT8 — 8-bit signed integer number representation.

Platform/reference	Type	Year	Speed (TMAC/s)	Energy (fJ/MAC)
Nvidia A100 (INT8)	GPU	2020	312	1603
Nvidia H100 SXM (INT8)	GPU	2023	2000	350
Nvidia GB200 Gr (INT8)	GPU	2025	5000	200
Nvidia MAGNET (ref. 22)	GPU	2019	1	40
S Ahmed (ref. 3)	PTC	2025	33	2424
C Pappas (ref. 19)	PTC	2025	131	546
S Bandyopadhyay (ref. 11)	PTC	2024	0.3	6000
S Ou (ref. 12)	PTC	2025	0.5	84
S Hua (ref. 4)	PTC	2025	4.1	840
X Xu (ref. 15)	PTC	2021	5.5	1580
J Feldman (ref. 14)	PTC	2021	1	5000
Y Xie (ref. 18)	PTC	2025	0.64	1360
Y Huang (ref. 20)	PTC	2023	0.04	474
C Feng (ref. 17)	PTC	2022	3.796	273.97
F Ashtiani (ref. 16)	PTC	2021	0.14	28000
B Bai (ref. 13)	PTC	2023	0.065	1380000
Z Chen (ref. 21)	FS	2023	15	12

Supplementary Table 2. Implemented model size vs. publication year. References are provided per the main text.

Reference	Type	Year	Model size
F Ashtiani (ref. 16)	PTC	2021	66
C Feng (ref. 17)	PTC	2022	288
Z Chen (ref. 21)	FS	2023	79410
Z Xu (ref. 24)	PTC	2024	13000000
C Wang (ref. 23)	PTC	2025	30640000

Supplementary Table 3: Accelerator speed vs energy efficiency for edge-oriented platforms. References are provided per the main text. DE — digital electronics, AE — analog electronics, PTC — photonic tensor core, FS — free space, MAC — multiply and accumulate.

Platform/reference	Type	Year	Speed (TMAC/s)	Energy (fJ/MAC)
T Zhou (ref. 44)	FS	2021	120	1267
M Luo (arXiv:2504.20416)	FS	2025	2350	8.3
M Miscuglio (ref. 45)	FS	2025	200	250
Y Chen (ref. 34)	FS	2023	2300	0.0267
Nakajima (ref. 43)	PTC	2021	21	30000
D Wang (ref. 42)	PTC	2024	105.5	48
Z Xu (ref. 24)	PTC	2024	25000	12.43
Nvidia Jetson Orin	DE	2023	5.32	11280
Google Coral edge TPU	DE	2023	2	1000
AMD Kria K26 (DPU4096)	DE	2021	0.68	11030
Ti AM69A	DE	2023	12	1583.33
Nvidia Jetson Thor T5000	DE	2025	1035	251.2
Axelera Metis	DE	2022	107	133.33
DIMC	DE	2022	10.016	0.901
Encharge AI EN100	AE	2025	100	82.5

Supplementary Table 4. Input size vs. latency for free space platforms. References are provided per the main text. DONN — diffraction optical neural networks, CONN — convolution optical neural networks, FCONN — fully connected optical neural networks.

Reference	**Network type**	**Year**	**Input size**	**Latency, ns**
X Lin (ref. 47)	DONN	2018	90000	$>10^9$
T Wang (ref. 48)	FCONN	2023	1600	$1.25 \cdot 10^7$
T Zhou (ref. 44)	DONN	2021	490000	$6.25 \cdot 10^5$
M Miscuglio (ref. 45)	CONN	2020	43264	10^6
Y Chen (ref. 34)	DONN	2023	65536	72
Z Chen (ref. 21)	FCONN	2023	784	30