UC Irvine
UC Irvine Previously Published Works

Title
Photonic accelerators for AI across cloud and edge platforms

Permalink
https://escholarship.org/uc/item/04w354jq

Journal
Nature Photonics, 20(10)

ISSN
1749-4885

Authors
Nechipurenko, Dmitry
Gautam, Abhishek
Shcherbakov, Maxim R

Publication Date
2026-10-01

DOI
10.1038/s41566-026-01987-z

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

nature photonics

Perspective

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

Photonic accelerators for AI across cloud and edge platforms

Received: 14 October 2025

Accepted: 13 July 2026

Published online: 15 September 2026

Check for updates

Dmitry Nechipurenko, Abhishek Gautam & Maxim R. Shcherbakov

Light-driven computing systems are closing the gap with state-of-the-art digital accelerators in terms of energy efficiency, compute throughput, density and latency. Here we map where photonic accelerators can matter in modern artificial intelligence infrastructure, and what must change for their impact at scale. We distinguish cloud-oriented, general-purpose photonic accelerators from application-specific, edge-oriented systems. In the cloud, the central challenge is scalability under tight energy budgets: heterogeneous systems with large vector inputs suffer from losses and electro–optic interfaces that dominate power and impede throughput, preventing the commercial deployment of photonic tensor cores. At the edge, opportunities arise wherever ultralow latency or high spatial parallelism are required and tight energy constraints dominate—for instance, in on-fibre optical processing and vision—yet progress is limited by the underwhelming scalability of all-optical nonlinearities, the lack of efficient reconfigurability, and a sizable optics footprint. We conclude that photonics is a viable near-term strategy within the existing digital ecosystem, and with focused advances along these axes, it can enable more sustainable artificial intelligence and open the door to broader adoption over the next decade.

The ongoing artificial intelligence (AI) revolution would be impossible without high-performance computing hardware. The development of massively parallel architectures enabled training of large deep neural networks (DNNs), sometimes comprising over a trillion trainable parameters[1]. However, modern AI hardware relies on power-hungry electronics, and sustaining the growth of both AI model size and demand poses a recognized challenge due to the rapidly accelerating power consumption for training and inference. Photonics is among the possible routes to mitigate the increasing power costs. From fibre-optic communication to light detection and ranging and camera-based navigation, light-driven information systems have become critical components of modern AI infrastructure, continually expanding their impact. Therefore, further delegation of information processing to the optical domain could become the natural step to alleviate existing AI bottlenecks.

Computing with photonic accelerators (PAs) can complement digital and analogue electronics. Computation via light propagation is often cited as advantageous due to the fundamental features of light–matter interactions: nearly dissipationless dynamics, ultrahigh bandwidth and natural spatial parallelism[2]. With recent demonstrations of increasingly general-purpose PAs capable of executing transformer-like workloads[3] and heterogeneously integrated PAs with nanosecond latency[4], photonics is in strong contention for the future of AI hardware. Nevertheless, converting fundamental features into computational optical advantage remains a daunting challenge. Despite a decade of explosive growth, photonic computing still faces a bar set by the maturity of digital electronics even as it takes important steps towards accelerating and decarbonizing AI[5–8].

In this Perspective we provide an overview of current progress, challenges and possible approaches to address the major roadblocks to photonic AI at scale. In our analysis we view future developments in PAs in the context of their application goals and hence distinguish between cloud- and edge-oriented platforms, as outlined in Fig. 1.

For cloud-oriented PAs, which aim to facilitate large-scale, general-purpose linear algebra operations, it is critical to bridge the current gap in both compute throughput and energy efficiency while retaining

Department of Electrical Engineering and Computer Science, University of California, Irvine, CA, USA. e-mail: maxim.shcherbakov@uci.edu

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

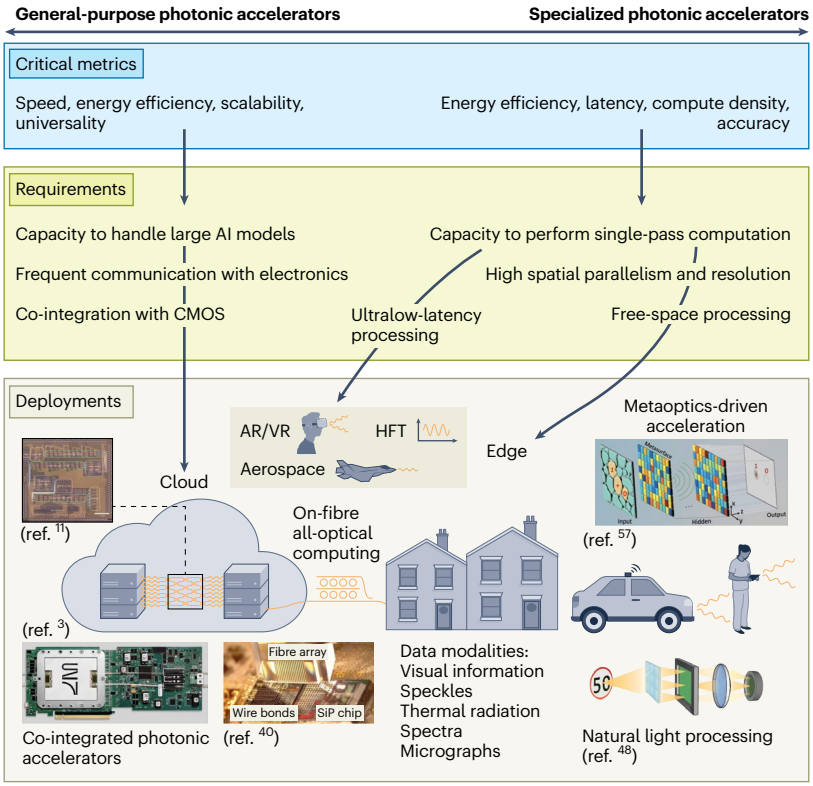


Fig. 1 | Quo vadis, photonic AI? Mapping the critical metrics, requirements and deployments of light-accelerated computing. From general-purpose accelerators in the cloud to application-specific processors, light-based computing faces stringent requirements to become deployable. Addressing underlying issues effectively may lead to adoption across a variety of scenarios. AR/VR, augmented reality/virtual reality; HFT, high-frequency trading; SiP, silicon photonics. Deployment images reproduced with permission from: top-left circuit, ref. 11, Springer Nature Ltd; bottom-left printed circuit board, ref. 3, Springer Nature Ltd; bottom-right printed circuit board, ref. 40, Springer Nature Ltd; metaoptics-driven acceleration graphic, ref. 57 under a Creative Commons license CC BY 4.0. Natural light processing diagram adapted with permission from ref. 48, Springer Nature Ltd.

system universality. The major roadblocks to achieving these goals with current devices are scalability challenges and power-inefficient interfaces between photonic integrated circuits (PICs) and digital processing units. Addressing these challenges requires multidisciplinary efforts from photonics, electronics and AI communities, which we briefly outline.

For application-specific photonic solutions specialized to handle smaller-scale AI models at the edge, the critical metrics depend on the specific application type, with potential niches requiring various combinations of energy efficiency, latency, compute density, and accuracy[9]. Leveraging optical computing in natural-light processing for smart cameras and sensors, as well as for on-fibre data analysis, will require overcoming multiple challenges. The major roadblocks in this application domain include poorly scalable all-optical nonlinearities, bulky optics and difficulties implementing energy-efficient reconfigurable devices. We believe that most of the challenges discussed in the Perspective can be addressed by the photonics community in the near future, opening a new era for light-enabled, sustainable AI.

General-purpose photonic AI

A brief performance survey of PAs

Photonic accelerators offer a potentially fast and energy-efficient approach to matrix–vector multiplication (MVM), enabling a resurgence of interest in neuromorphic computing using PICs. The impetus behind this interest lies in the low-latency nature of signal propagation through a PIC: a simple back-of-the-envelope calculation shows that for an $L = 1$-mm-long chip operating on non-resonant elements, the optical latency of an inference task is $\tau \approx Ln_{Si}/c \approx 0.01$ ns (where n_{Si} is the refractive index of silicon and c is the speed of light in a vacuum), a figure unrivalled by modern electronics.

Building on this simple idea, a vast zoo of PIC platforms for photonic computing has been established[10–20]. Rather than focusing on the various ways in which MVM and nonlinear operations have been realized—detailed surveys of which can be found in recent reviews[7,8]—we analyse the overall performance of recent systems using several important metrics and identify the performance gaps. We divide the reported PAs into two major groups: fully dynamically reconfigurable but partially integrated solutions that can handle a wide variety of computations[11–20], but face substantial data-movement overheads when scaled to large models; and fully reconfigurable and fully integrated platforms that are designed to handle larger-scale models, allowing sequential execution of large models on one or more photonic tensor cores (PTCs)[3,4].

Figure 2a summarizes the reported performances of 13 published general-purpose PAs (12 PIC-based systems and one free-space system[21]) in terms of computational throughput (in multiply-and-accumulate (MAC) operations per second) and energy efficiency (in joules per MAC) against the last three generations of graphics processing units (GPUs) and one experimental, power-efficient system[22]. Due to the large footprint of photonic components and the much higher dimensionality of the envisioned PTCs, a single PA module is expected to comprise a handful of cores rather than the thousands found in modern GPUs. In our analysis, we consider the whole-system performance metrics for PAs and GPUs. The data suggest that modern general-purpose PAs are, overall, behind digital electronics in terms of throughput and energy efficiency.

While most reported PAs have been benchmarked on relatively simple AI models, the size of implemented AI models has grown over the years, as shown in Fig. 2b. Single-core light-based accelerators have reached the capacity to execute models with millions[23,24] of parameters, commensurate with tinyML and small embedding models.

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

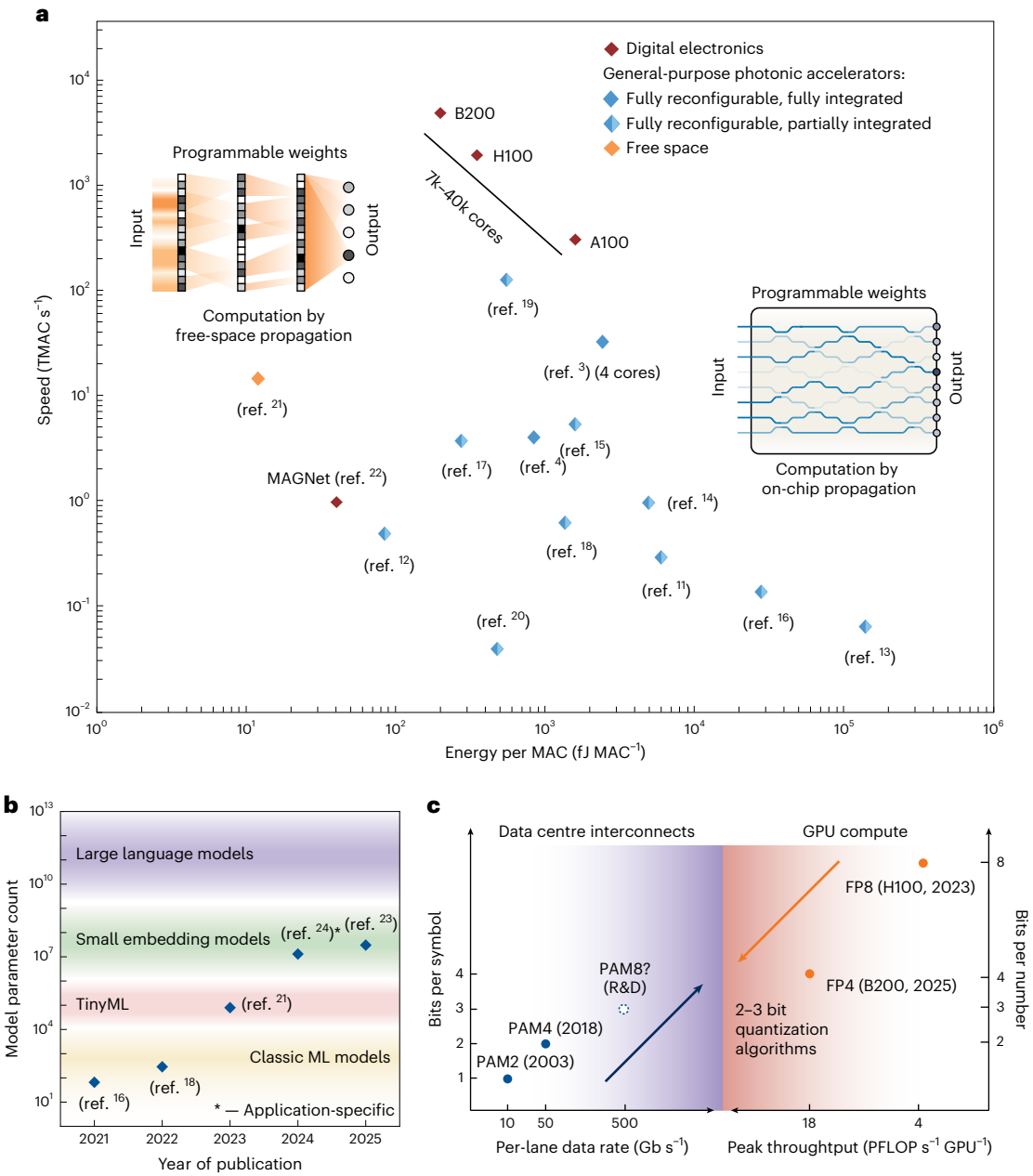


Fig. 2 | Chasing electronics at the speed of light. Overview of the critical PA metrics and their evolution. **a**, Compute speed and energy efficiency in modern GPUs (red; INT8 precision), PICs (blue) and free-space systems (orange). **b**, Ranges corresponding to parameter counts for various types of ML model are shown with horizontal bars, and the parameter counts for models executed on PAs (diamonds) are depicted as a function of time; for each year, the largest PA-implemented model was selected. References [23,24] represent partially reconfigurable platforms, with most of the reported weights not tunable due to their diffractive nature. **c**, Current trends in optical interconnects involve increasing the bandwidth by increasing the modulation bit depth, while boosting GPU performance has been achieved by decreasing the bit depth used for number representation. Corresponding trends are shown with blue and red arrows, respectively. PAM, pulse amplitude modulation; FP, floating-point precision.

Closing the performance gap

To keep pace with electronics over the next decade, the performance of PAs will need to improve substantially. Scaling existing systems will require a multifaceted approach. We understand the scalability of a PA in terms of the following:

- the capacity to increase the number of inputs and operations in individual PTCs (scaling up);
- the capacity to use multiple PTCs simultaneously (scaling out); and
- the ability to increase the number of DNN layers executed in the optical domain (scaling through).

Scaling up. The limited number of inputs and weights processed concurrently in each PTC is considered a major roadblock to viable PAs[8]. Larger systems exhibit higher throughput and energy efficiency[2,3,10], benefiting from spatial multiplexing or exploiting other degrees of freedom (for example, wavelength-division multiplexing (WDM)). Simulations predict that building a PTC capable of performing single-pass MVMs with vectors of $N \approx 1{,}000$ elements could surpass electronics in energy efficiency while executing large language model-scale workloads[25].

Is $N > 1{,}000$ feasible in PIC PAs? The challenge of achieving high N in single-pass architectures lies in accommodating a large number of

inputs and weights within realistic constraints on chip area, optical loss, signal-to-noise ratio (SNR), power dissipation and packaging bandwidth. Coherent Mach–Zehnder interferometer meshes, while realizing arbitrary unitary operations, do not allow $N > 10$–20, and this limitation has caused the field to shift to geometries with fixed optical depth[3,4], where scalability is governed by optical power distribution and detector noise rather than cascading loss (see Supplementary Note 1 for details). In fixed-depth architectures, to move from the state-of-the-art optical MAC ($N \approx 100$) to the desired scale of $N > 1{,}000$, a dense array of over $N^2 \approx 10^6$ modulators would be required. Although the spatial footprint of typical Mach–Zehnder modulators may be potentially compatible with packaging millions of modulators in a 300-mm wafer, a number of additional issues, such as component variability and electronic co-packaging complexities, call for alternative approaches.

Hypermultiplexing and overcoming the spatial complexity in PICs. To address the scaling challenge of single-pass systems, architectures with alternative scaling laws are being explored, including several platforms that take advantage of space–time–wavelength hypermultiplexing[12,19,26]. For this type of architecture, the number of modulator components on a chip can scale as $O(N_s)$, where N_s is the number of spatial, concurrently processed vector channels, using WDM, while the total length of each vector can be $N \gg 1{,}000$ due to constant streaming of both vector and matrix components and signal integration (summation) performed in the analogue electronic domain. These platforms perform matrix–matrix[12] or matrix–tensor[19] multiplication, and the energy efficiency scaling law predicts an efficiency of 5–10 fJ MAC^{-1} for $N_s = 100$ (see Supplementary Note 2 for details). A recent implementation with $N_s = 7$ has enabled computations at 84 fJ MAC^{-1} (ref. 12). Expanding the current systems to $N_s \approx 100$ is a critical next step and appears less challenging than the expansion of single-pass architectures.

Free-space processors. The natural spatial parallelism of free-space systems offers advantages for both general-purpose and application-specific computing with light. Such systems have already demonstrated record-high energy efficiency for general-purpose analogue computations[21] (Fig. 2a) and have recently achieved single-shot matrix–matrix multiplication using wavelength–spatial–temporal hypermultiplexing[27]. A free-space platform has also been utilized as a critical part in an all-analogue optical computer for AI inference and beyond[28].

Lithography-free architectures. Recently, several promising architectures have been reported in which the propagating medium is a programmable thin-film waveguide[29,30]. Although the energy footprint of these PAs remains to be optimized, centimetre-scale input facets can host thousands of parallel diffraction-limited inputs. Challenges remain in all-to-all input connectivity, as well as material programmability.

Bottom line. To summarize the prospects of scaling up, our analysis suggests that monolithic single-pass PIC architectures are unlikely to reach $N \gg 1{,}000$. By contrast, hypermultiplexed, free-space and lithography-free architectures can achieve effective large-N operation at the system level. This may, however, come with reduced universality in lithography-free systems and increased reliance on electronic accumulation in hypermultiplexed systems.

Scaling out. 'Scaling out' implies an increase in the number of PTCs used in parallel within a single system to boost its overall throughput[3]. As it stands, this strategy is stonewalled by frequent communication between PTCs and their digital control interface units. Intense data movement between the modules, including memory access, analogue-to-digital (ADC) and digital-to-analogue (DAC) conversion—all performed in the electronic domain—results in substantial energy and bandwidth overheads, leading to an underperformance of fully integrated platforms (Fig. 2a).

Mitigation of ADC/DAC overheads. A notable decrease in ADC-based energy consumption can be achieved by performing the summation step of MVM using differential photodetector current integration while utilizing techniques such as time-division multiplexing[12,31]. Importantly, this approach offers synergy with overall system throughput boosting by using high input and weight modulation frequencies, as allowed, for example, by thin-film lithium niobate (TFLN)-based components[32]. DAC-based overheads at high bandwidths in future platforms can be mitigated by more efficient integrated electro–optic digital-to-analogue links that leverage TFLN-based nanophotonics[33]. For hypermultiplexing architectures, the DAC conversion overheads can be mitigated through parallel processing of multiple vectors using WDM[12], whereas for single-pass architectures, these overheads can be decreased by using the same weights multiple times[2] (sequentially batched vectors), lowering the overall weight update frequency. Supplementary Note 2 provides details on energy breakdown and corresponding scaling laws for both architecture types.

Mixed-signal electronics. The substantial energy budget of the digital electronic modules in current PAs could also be reduced through analogue processing of photocurrents[28,34] or more radical solutions relying on analogue electronic and/or optical memory cells[35].

PA-specific algorithms and system co-design. Modern AI algorithms, such as fine-tuning and quantization-aware training[36], allow the retention of model accuracy while reducing the bit depth of the data, thereby mitigating energy overheads for conversion, memory usage, and data movement in the power-hungry electronic domain. New algorithms and hardware solutions for hypermultiplexing platforms could further optimize overall energy budgets of the future fully integrated solutions.

Scaling through. Scaling through implies an increase in the number of PA layers that implement cascaded linear and nonlinear operations all-optically. Scaling through is fundamentally constrained by the availability of suitable nonlinear activation functions, because without inter-layer nonlinearity, multiple optical layers collapse into a single linear operation. The pursuit of nonlinearity in the optical domain, in the context of general PAs, has been carried out mostly in the form of electro–optical nonlinearities[11]. However, in many general PA architectures, the summation step of MVMs is performed via electronic accumulation, followed by nonlinear activation. This approach enables scaling to very large models without requiring cascaded optical nonlinearities. Nevertheless, scalable optical nonlinear activation remains an important enabler for small models and for application-specific acceleration, as discussed in the following. A recent review outlines potential pathways toward scaling through in future hybrid or fully optical systems[37].

Application-specific photonic acceleration

In this section, we explore several promising avenues for future specialized PAs and examine the major challenges these platforms face.

On-fibre all-optical computing

Most PA components—lasers, modulators, photodetectors, ADC/DAC and amplifiers—are readily available and well established in modern fibre communications. Specialized PAs performing what can be described as on-fibre all-optical computing can operate on incoming and outgoing fibre-optic signals at transponders to mitigate both latency and energy overheads in current wide-area networks[38]. Such optical co-processors could handle both user-facing (for example, video encoding/decoding using diffraction-based processing[39] or optical convolutional neural network-based algorithms[14,15]) and network-driven (for example, packet classification) tasks.

Although nonlinear effects in fibre-optic communications are usually avoided or compensated by various means (including ultrafast silicon photonic DNNs[40]), they can also serve as the physical

substrate for analogue on-fibre reservoir computing[41]. The ultrahigh fibre data throughput can be used for large-scale computations, such as running large AI models on the fly at the edge via optical streaming of weight data[31], and for ultralow-latency applications such as high-frequency trading.

On-fibre computing looks promising due to multiple driving factors. First, current trends in both fibre-optic technology and AI hardware and algorithm development are convergent and are expected to maintain the same bit depth per symbol (Fig. 2c), potentially enabling direct analogue processing of data in the optical domain. Second, some fibre-based data protocols support the raw data transfer suitable for optical analogue processing after DAC conversion, which can be performed in the optical domain. All-optical DACs can facilitate direct optical input from photonic waveguides or optical fibres, which currently host protocols with one or two bits per symbol, thereby reducing electronic overheads and enabling 'compression through transmission' solutions. However, commercial deployment will also require overcoming integration complexity, ensuring compatibility with existing network hardware, and implementing robust error correction to ensure reliability at scale.

Light-based computing at the edge

Edge computing enables near-real-time data processing on resource-constrained devices located near the data source, thereby minimizing communication delays and bandwidth usage[9]. Although ubiquitous in modern infrastructure, it suffers from inherent challenges, including limited computational resources, storage and memory constraints, and energymanagement challenges. All-optical data preprocessing in photonic devices presents an exciting opportunity to mitigate energy and latency overheads at the edge.

To illustrate both the advances and limitations of edge photonic computing, we next compare existing implementations of edge PAs across different architectures—including chip-based[24,42,43] and free-space systems[34,44,45]. Figure 3a shows the performance of state-of-the-art application-specific photonic systems and modern edge-oriented digital electronic processors. By leveraging passive computations with limited reconfigurability, several photonic platforms achieve speed and power efficiency exceeding those of digital electronic systems, demonstrating the great potential of such systems for inference tasks. However, it should be noted that both the throughput and energy efficiency of the top performers are estimated using the so-called number of optical operations in diffraction-based systems, which should be interpreted and compared with other hardware with caution[23].

PIC-based edge PAs. PICs offer robust, co-integrated designs that have been proposed to address edge problems, such as image and sound classification and time-series prediction[11,24,42,43]. The major challenges in PIC-based edge acceleration echo those listed in the section General-purpose photonic AI, resulting in smaller-scale tasks and reliance on resource-intensive DAC/ADC. Nevertheless, promising architectures have recently emerged in which all-optical processing of analogue inputs on a chip helps mitigate the initial ADC stages for ultralow-latency inference[46].

Free-space edge PAs. The edge environment is characterized by analogue signals from the surroundings detected by sensors. Light carries a large proportion of edge-processed information, typically captured with cameras. Free-space PAs can process this information, bypassing ADCs and optoelectronic conversion, allowing for passive computation, extremely low latency, high energy efficiency, and the ability to process large inputs due to high spatial parallelism[2]. Figure 3b compares the three primary edge PA architectures—diffractive optical neural networks (DONNs)[34,44,47], convolutional optical neural networks (CONNs)[45] and fully connected optical neural networks (FCONNs)[21,48]—in terms of reported latency and number of processed inputs, with state-of-the-art low-resolution cameras (LRCs) and high-resolution smart cameras (HRSCs). Although the most universal of the three, FCONNs currently support a modest number of concurrent inputs and optical neurons. In contrast, both CONNs and DONNs have demonstrated the ability to process a large number of inputs simultaneously, comparable to the capacity of embedded LRCs. Despite impressive performance metrics, free-space platforms have several caveats, including difficulties performing nonlinear operations in the optical domain, lack of energy-efficient reconfigurability, as discussed previously[49,50], and the open challenge of system-level miniaturization.

Multilayer ONNs with scalable optical nonlinearities. Current free-space PAs rarely exceed two or three layers, whereas even simple edge AI models executed on mobile devices typically comprise tens of layers. Implementing multiple layers in a PA without optoelectronic conversion is challenging due to several limitations, including the difficulty of achieving efficient, scalable all-optical nonlinear activations. Although nonlinear activations with complete optoelectronic conversion followed by regeneration of the optical signal have been realized[48,51], this approach is associated with increased latency and energy losses. Nonlinear activations with either no or partial optoelectronic conversion have been implemented in many ways, including those relying on partial conversion-driven modulation of the primary optical signal in resonators[11] and those based on recently explored structural nonlinearity[52,53]. Nonlinear optical effects, such as second-harmonic generation, have recently been utilized in various reservoir computing platforms based on multimode fibres[41] and lithium niobate polycrystals[54]. However, they have not been reported for more universal systems, such as optical DONNs, CONNs and FCONNs.

To alleviate the lack of scalable, low-threshold nonlinearity, several approaches are envisioned. Leveraging state-of-the-art nanophotonics capable of light concentration and confinement through micro and nano-resonators and nonlinear metasurfaces can boost optical nonlinearities by orders of magnitude[55]. Other approaches, such as material-based nonlinear activation relying on saturable absorption, phase-change media or nonlinear propagation and accumulation in disordered or multimode systems, have potential for scaling through in edge PAs[37]. Integrated mode-locked lasers[56] could, in the future, be used as ultracompact pulse sources that deliver high-intensity bursts of light sufficient to drive nonlinear optical responses with low average driving power. However, challenges such as activation threshold, device variability, limited reconfigurability, trade-offs between speed and losses, and stability under repeated cascading are yet to be addressed.

Metaoptics-driven compactification. Future implementations of edge PAs may leverage metaoptics[57]. Besides subwavelength thickness and near-arbitrary diffraction control, leading to ultrashort working distances, metasurfaces provide unique opportunities to increase the number of independently processed degrees of freedom of the incoming light, including wavelength, polarization and angle of incidence[58]. In photonic reservoir computing, near-field coupling between meta-atoms can significantly increase the effective computational density of the system, potentially reaching billions of optical neurons per cubic millimetre of nanophotonic media (for example, in multilayer metasurfaces). Recently, polychromatic metasurface-based convolutional image processing was performed using a knowledge distillation approach, which allows notable downsizing of the number of DNN layers[59]. Another prospective approach based on random projection with 41 million meta-atoms on a single metasurface achieved high-accuracy image classification at up to megapixel resolution, according to ref. 60.

Towards reconfigurable free-space PAs. Tunable photonic elements are essential in free-space PAs to enable reconfigurability without forfeiting the intrinsic latency advantage of propagation-based computation. Among the available approaches, low-power electro–optic

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

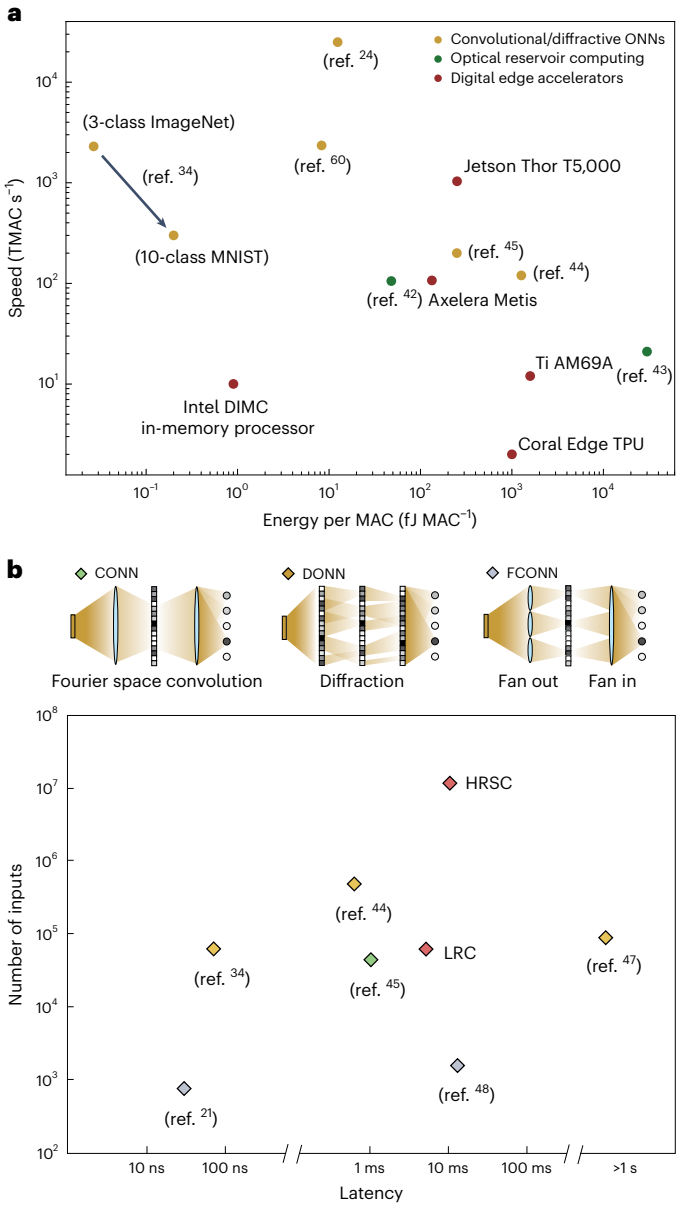


Fig. 3 | Survey of application-specific PAs. a, Performance of application-specific PAs (orange dots for convolutional/diffractive ONNs; green dots for optical reservoir computers) and edge-oriented digital accelerators (red dots). Although competitive performance is evident for some PA architectures, metrics remain heavily model-size-dependent, as shown in the class performance variability in ref. 35. **b**, Latency versus the number of concurrent inputs processed in reported free-space optical neural networks, compared with state-of-the-art smart cameras. Diffractive (DONN), convolutional (CONN) and fully connected (FCONN) optical neural networks demonstrate large-scale spatial parallelism and, in some cases, latencies rivalling embedded vision hardware. However, while smart cameras achieve compactness and high integration, free-space systems currently face challenges in scalability, reconfigurability and nonlinear processing, which must be addressed before they can transition beyond proof-of-concept demonstrations. HRSC, high-resolution smart camera; LRC, low-resolution camera.

tunability is preferred over thermo–optic control, with viable implementations based on liquid crystals, Pockels-effect materials, carrier-depletion modulation, phase-change media and emerging optomechanical and electro-absorptive mechanisms[61,62]. Such reconfigurability enables adaptive weighting and edge-deployable operation in free-space PAs, while remaining compatible with the nonlinear optical primitives discussed above.

Moving forward. An unexplored opportunity for PAs at the edge is to harness the untapped potential of light-based sensors beyond cameras. Co-designing PAs that interface with optical microphones, interferometric, speckle-based or spectroscopic biomedical probes, environmental probes, inertial and motion sensors, and computational imaging platforms with lensless and diffraction-driven modalities[49,50], could open a plethora of all-optical pipelines that collapse latency and minimize energy overheads, enabling a new class of intelligent, self-processing sensors that blur the boundary between measurement and inference. Besides sensor data processing, PAs have recently been utilized for image generation, providing unique opportunities for next-generation AR/VR systems and beyond[63].

Conclusion

Light-based computing is a viable strategy for mitigating the rapidly growing power consumption and compute-throughput bottlenecks of AI hardware. Despite recent breakthroughs and growing interest from the research community and industry, this emerging field still faces multiple challenges, as we highlight in this Perspective.

In cloud platforms, PAs are beginning to demonstrate competitive efficiency but require significant scaling up to compete with digital electronics. Closing this gap demands advances on several fronts: expanding hypermultiplexed platforms, exploring novel free-space and lithography-free architectures, and mixed-signal electronic–photonic co-design that minimizes conversion overheads.

At the edge, photonics aligns naturally with application-driven needs, providing low latency, high spatial parallelism, and direct compatibility with optical signals. Emerging opportunities span free-space and on-fibre optical preprocessing, leveraging the existing telecommunications infrastructure, metasurface platforms and compact nonlinear elements that enable cascaded neural network architectures. Progress here depends on overcoming current barriers in reconfigurability, nonlinear scalability and device footprint.

Future co-integrated PAs can facilitate both inference and training tasks performed in the cloud domain (see Supplementary Note 3 for details), while application-specific photonic platforms can take full advantage of the low-latency inference enabled by computation via propagation. Most reported systems are trained in silico using their digital twins, but the community is exploring the in situ training of physical neural networks[64], including PAs, which appears attractive, as evidenced by recent demonstrations[11,64].

In summary, optical computing has a compelling outlook as a complementary AI technology. The energy efficiency and throughput potential of PAs remains far from exhausted, and hardware advances are likely to be swift. As the photonics–AI frontier matures, deployable optical computing is emerging as a credible contender for specialized acceleration, with the potential for broader adoption as key challenges are addressed.

References

1. Patil, R. & Gudivada, V. A review of current trends, techniques and challenges in large language models (LLMs). *Appl. Sci.* **14**, 2074 (2024).
2. McMahon, P. L. The physics of optical computing. *Nat. Rev. Phys.* **5**, 717–734 (2023).
3. Ahmed, S. R. et al. Universal photonic artificial intelligence acceleration. *Nature* **640**, 368–374 (2025).
4. Hua, S. et al. An integrated large-scale photonic accelerator with ultralow latency. *Nature* **640**, 361–367 (2025).
5. Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
6. Shastri, B. J. et al. Photonics for artificial intelligence and neuromorphic computing. *Nat. Photon.* **15**, 102–114 (2021).
7. Farmakidis, N., Dong, B. & Bhaskaran, H. Integrated photonic neuromorphic computing: opportunities and challenges. *Nat. Rev. Electr. Eng.* **1**, 358–373 (2024).

8. Fu, T. et al. Optical neural networks: progress and challenges. *Light Sci. Appl.* **13**, 263 (2024).
9. Wang, X. et al. Empowering edge intelligence: a comprehensive survey on on-device AI models. *ACM Comput. Surv.* **57**, 228 (2025).
10. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11**, 441–446 (2017).
11. Bandyopadhyay, S. et al. Single-chip photonic deep neural network with forward-only training. *Nat. Photon.* **18**, 1335–1343 (2024).
12. Ou, S. et al. Hypermultiplexed integrated photonics-based optical tensor processor. *Sci. Adv.* **11**, eadu0228 (2025).
13. Bai, B. et al. Microcomb-based integrated photonic processing unit. *Nat. Commun.* **14**, 66 (2023).
14. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
15. Xu, X. et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
16. Ashtiani, F., Geers, A. J. & Aflatouni, F. An on-chip photonic deep neural network for image classification. *Nature* **606**, 501–506 (2022).
17. Feng, C. et al. A compact butterfly-style silicon photonic-electronic neural chip for hardware-efficient deep learning. *ACS Photonics* **9**, 3906–3916 (2022).
18. Xie, Y. et al. Complex-valued matrix-vector multiplication using a scalable coherent photonic processor. *Sci. Adv.* **11**, eads7475 (2025).
19. Pappas, C. et al. A 262 TOPS hyperdimensional photonic AI accelerator powered by a Si_3N_4 microcomb laser. *APL Photonics* **10**, 110805 (2025).
20. Huang, Y. et al. Easily scalable photonic tensor core based on tunable units with single internal phase shifters. *Laser Photonics Rev.* **17**, 2300001 (2023).
21. Chen, Z. et al. Deep learning with coherent VCSEL neural networks. *Nat. Photon.* **17**, 723–730 (2023).
22. Venkatesan, R. et al. MAGNet: a modular accelerator generator for neural networks. In *Proc. 2019 IEEE/ACM International Conference on Computer-Aided Design* (*ICCAD*) https://doi.org/10.1109/ICCAD45719.2019.8942127 (IEEE, 2019).
23. Wang, C., Cheng, Y., Xu, Z., Dai, Q. & Fang, L. Diffractive tensorized unit for million-TOPS general-purpose computing. *Nat. Photon.* **19**, 1078–1087 (2025).
24. Xu, Z. et al. Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence. *Science* **384**, 202–209 (2024).
25. Anderson, M., Ma, S. Y., Wang, T., Wright, L. & McMahon, P. L. Optical transformers. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=Xxw0edFFQC (2024).
26. Kovaios, S. et al. On-chip 1 TOPS hyperdimensional photonic tensor core using a WDM silicon photonic coherent crossbar. *J. Lightwave Technol.* **43**, 8799–8805 (2025).
27. Luan, C., Davis III, R., Chen, Z., Englund, D. & Hamerly, R. Single-shot matrix-matrix photonic processor based on spatial-spectral hypermultiplexed parallel diffraction. *Nat. Commun.* **17**, 484 (2026).
28. Kalinin, K. P. et al. Analog optical computer for AI inference and combinatorial optimization. *Nature* **645**, 354–361 (2025).
29. Wu, T., Li, Y., Ge, L. & Feng, L. Field-programmable photonic nonlinearity. *Nat. Photon.* **19**, 725–732 (2025).
30. Onodera, T. et al. Arbitrary control over multimode wave propagation for machine learning. *Nat. Phys.* **22**, 164–171 (2026).
31. Sludds, A. et al. Delocalized photonic deep learning on the internet's edge. *Science* **378**, 270–276 (2022).
32. Lin, Z. et al. 120 GOPS photonic tensor core in thin-film lithium niobate for inference and in situ training. *Nat. Commun.* **15**, 9081 (2024).
33. Song, Y. et al. Integrated electro-optic digital-to-analogue link for efficient computing and arbitrary waveform generation. *Nat. Photon.* **19**, 1107–1115 (2025).
34. Chen, Y. et al. All-analog photoelectronic chip for high-speed vision tasks. *Nature* **623**, 48–57 (2023).
35. Kari, S. R. et al. Optical and electrical memories for analog optical computing. *IEEE J. Sel. Top. Quantum Electron.* **29**, 1–12 (2023).
36. Kirtas, M. et al. Mixed-precision quantization-aware training for photonic neural networks. *Neural Comput. Appl.* **35**, 21361–21379 (2023).
37. Shi, W., Huang, Z., Fu, T. & Chen, H. Review of nonlinear activation functions in optical neural networks. *Adv. Photonics* **7**, 064004 (2025).
38. Yang, M., Zhong, Z. & Ghobadi, M. On-fiber photonic computing. In *Proc. 22nd ACM Workshop on Hot Topics in Networks* (ed. Arun, V.) 263–271 (ACM, 2023).
39. Chen, Y. et al. Photonic unsupervised learning variational autoencoder for high-throughput and low-latency image transmission. *Sci. Adv.* **9**, eadf8437 (2023).
40. Huang, C. et al. A silicon photonic–electronic neural network for fibre nonlinearity compensation. *Nat. Electron.* **4**, 837–844 (2021).
41. Oguz, I. et al. Programming nonlinear propagation for efficient optical learning machines. *Adv. Photonics* **6**, 016002 (2024).
42. Wang, D., Nie, Y., Hu, G., Tsang, H. & Huang, C. Ultrafast silicon photonic reservoir computing engine delivering over 200 TOPS. *Nat. Commun.* **15**, 10841 (2024).
43. Nakajima, M., Tanaka, K. & Hashimoto, T. Scalable reservoir computing on coherent linear photonic processor. *Commun. Phys.* **4**, 20 (2021).
44. Zhou, T. et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nat. Photon.* **15**, 367–373 (2021).
45. Miscuglio, M. et al. Massively parallel amplitude-only Fourier neural network. *Optica* **7**, 1812–1819 (2020).
46. Zhang, S. et al. Photonic edge intelligence chip for multi-modal sensing, inference and learning. *Nat. Commun.* **16**, 10136 (2025).
47. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
48. Wang, T. et al. Image sensing with multilayer nonlinear optical neural networks. *Nat. Photon.* **17**, 408–415 (2023).
49. Hu, J. et al. Diffractive optical computing in free space. *Nat. Commun.* **15**, 1525 (2024).
50. Choi, M. & Majumdar, A. Free-space optical encoder for computer vision. *npj Nanophoton.* **2**, 36 (2025).
51. Tait, A. N. et al. Silicon photonic modulator neuron. *Phys. Rev. Appl.* **11**, 064043 (2019).
52. Yildirim, M., Dinc, N. U., Oguz, I., Psaltis, D. & Moser, C. Nonlinear processing with linear optics. *Nat. Photon.* **18**, 1076–1082 (2024).
53. Xia, F. et al. Nonlinear optical encoding enabled by recurrent linear scattering. *Nat. Photon.* **18**, 1067–1075 (2024).
54. Wang, H. et al. Large-scale photonic computing with nonlinear disordered media. *Nat. Comput. Sci.* **4**, 429–439 (2024).
55. Vabishchevich, P. P. & Kivshar, Y. S. Nonlinear photonics with metasurfaces. *Photon. Res.* **11**, B50–B64 (2023).
56. Yang, J. et al. Titanium:sapphire-on-insulator integrated lasers and amplifiers. *Nature* **630**, 853–859 (2024).
57. Qian, C. et al. Performing optical logic operations by a diffractive neural network. *Light Sci. Appl.* **9**, 59 (2020).
58. Brongersma, M. L. et al. The second optical metasurface revolution: moving from science to technology. *Nat. Rev. Electr. Eng.* **2**, 125–143 (2025).

59. Choi, M. et al. Transferable polychromatic optical encoder for neural networks. *Nat. Commun.* **16**, 5623 (2025).
60. Luo, M. et al. Highly scalable machine vision enabled with meta-optics-based ultra-wide neural network. *eLight* **6**, 10 (2026).
61. Ding, F., Meng, C. & Bozhevolnyi, S. I. Electrically tunable optical metasurfaces. *Photonics Ins.* **3**, R07 (2024).
62. Abou-Hamdan, L. et al. Programmable metasurfaces for future photonic artificial intelligence. *Nat. Rev. Phys.* **7**, 331–347 (2025).
63. Chen, S., Li, Y., Wang, Y., Chen, H. & Ozcan, A. Optical generative models. *Nature* **644**, 903–911 (2025).
64. Momeni, A. et al. Training of physical neural networks. *Nature* **645**, 53–61 (2025).

Acknowledgements

This material is based on work supported by the National Science Foundation under grant no. ECCS-2339271.

Author contributions

A.G. and M.R.S. conceived the Perspective. All authors contributed to the literature review, compilation and interpretation of published data, preparation of figures, and writing and revision of the manuscript.

Competing interests

M.R.S. is a co-inventor on US patent application US18/865,192, published as US20250314915A1, 'Programmable photonic waveguides', filed by NTT Research, Inc. and the Center for Technology Licensing at Cornell University. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41566-026-01987-z.

Correspondence should be addressed to Maxim R. Shcherbakov.

Peer review information *Nature Photonics* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature Limited 2026