nature reviews physics

https://doi.org/10.1038/s42254-023-00645-5

Perspective

Check for updates

The physics of optical computing

Peter L. McMahon

Abstract

There has been a resurgence of interest in optical computing since the early 2010s, both in academia and in industry, with much of the excitement centred around special-purpose optical computers for neural-network processing. Optical computing has been a topic of periodic study since the 1960s, including for neural networks in the 1980s and early 1990s, and a wide variety of optical-computing schemes and architectures have been proposed. In this Perspective article, we provide a systematic explanation of why and how optics might be able to give speed or energy-efficiency benefits over electronics for computing, enumerating 11 features of optics that can be harnessed when designing an optical computer. One often-mentioned motivation for optical computing — that the speed of light is fast — is emphatically not a key differentiating physical property of optics for computing; understanding where an advantage could come from is more subtle. We discuss how gaining an advantage over state-of-the-art electronic processors will likely only be achievable by careful design that harnesses more than 1 of the 11 features, while avoiding a number of pitfalls that we describe.

Sections

School of Applied and Engineering Physics, Cornell University, Ithaca, NY, USA. e-mail: pmcmahon@cornell.edu

Introduction

There has been a resurgence of interest in optical computing since the early 2010s, both in industry and in academia[1–4]. What is the fundamental physical basis on which we can expect an optical computer to outperform an electronic computer, at least for some tasks? In this Perspective article, we enumerate and discuss 11 features of optics and optical computing that can contribute to an advantage for an optical computer. Any optical computer that achieves an advantage in practice will likely need to harness more than one of these features. An explicit list of features can help to make clear what ingredients the architect of an optical computer has to work with. It also allows researchers to systematically identify the fundamental physical principles behind the operation of different proposed optical computers, aids them in analysing what advantage they can hope to achieve and how their designs might be improved by exploiting further features. The design of a successful optical computer must be carefully engineered to avoid bottlenecks or overhead that would outweigh the optical benefits. We discuss some of the pitfalls and approaches one can take to mitigate them.

The high bar set by electronic processors has contributed to periods when there has been pessimism about the prospects for optical computing (for example, see refs. 5,6 from 2010). Given the continued improvements in CMOS technology[7], why is there now renewed excitement about optical computing, including commercial efforts[8,9]? One of the major criticisms of optical computing has been that optical transistors are not competitive with their electronic counterparts. The current wave of interest in optical computing is primarily focused on optical-computer architectures that are not based on replicating digital logic with optical transistors. Instead of trying to construct general-purpose, digital computers, the community is largely targeting building special-purpose, analog computers. Both these shifts — to special-purpose and to analog processing — are important. Trying to build performant general-purpose processors with optics remains out of reach, essentially because general-purpose processors are expected to have no errors (accountants want sums in their spreadsheets to be exactly correct, for example), and it is only known how to achieve error-free machines with digital logic; to build digital logic requires an optical transistor satisfying the criteria given in ref. 6 or something similar. However, one can alternatively build optical processors that are specialized to particular applications for which completely error-free operation is not necessary.

There are several application areas being targeted by special-purpose optical computers presently, including neural networks[1]; scientific computing[10]; combinatorial optimization[4] and cryptography[9,11,12]. All four application areas have as a key algorithmic primitive the process of matrix–vector multiplication, which is the target of much of the current research in optical computing. Fourier transforms and convolutions have applicability across neural networks, scientific computing and cryptography, contributing to their prominence in current research. Optical correlators have been released as commercial products during several periods over the past few decades[13], so this is not a new direction even commercially, but one that has been revitalized. There is also a substantial thrust in performing computations for neural networks that are not explicitly engineered to be matrix–vector multiplications or convolutions[1,14–18]. A commonality among all four application areas is that the subroutines performed optically are still useful even if they suffer from some error (noise). This factor is crucial because it is difficult to achieve an effective precision greater than 10 bits in any analog computer, including analog optical computers, so applications of analog optical computers should be robust to this level of noise. Neural networks are a particularly good match because, at least during inference (as opposed to training), neural networks do not suffer a substantial decrease in accuracy even if they are restricted to integer arithmetic with fewer than 8 bits of precision[1,19]. A concern for any analog neural-network processor, including analog optical processors, is the potential for accumulation of errors in executing deep neural networks. This has recently been theoretically analysed, with a conclusion that deleterious effects of noise accumulation can be mitigated, even in the case of correlated noise[20]. Uncorrelated noise that merely leads to an effective low-bit precision has been shown in simulations of deep optical neural networks (having 60 optically executed layers) to yield accuracies that are the same as or better than that of digital-electronic processors executing the same neural network with 8-bit integer arithmetic[21], that is, the simulations predicted that the accumulation of error in an optical implementation of the neural network would not have a noticeable impact on accuracy compared with a standard digital-electronic implementation. For all applications of analog optical processors, neural networks, intuition and simulations about resilience to noise ultimately need to be validated by optical experiments.

With this context, we can now give a fuller answer to why there is renewed excitement in optical computing. The first reason is the rise of neural networks: over the past decade, neural networks have become a dominant approach in machine learning and have become extremely compute-resource-intensive. This has led to strong interest in alternative hardware approaches specialized to neural networks, and the intrinsic resilience of neural networks to noise makes them well suited to analog optical implementations. Second, CMOS improvements would not be enough to satisfy application demand: although there has been remarkable progress in CMOS hardware[7], it is also simultaneously true that both for neural networks and for some other applications (such as combinatorial optimization), the anticipated future improvements in CMOS hardware[22] are less than users would like and will limit application capabilities[23]. For instance, the number of parameters in neural networks — one measure of their size and computational demand — has been growing much faster than hardware improvements[24], primarily because of the finding that increased scale often leads to increased capability or accuracy[25,26]. Third, there have been large improvements in photonics hardware: driven largely by the consumer-electronics and the optical-communications industries, there have been enormous advances in the scale, speed and energy efficiency of photonic devices over the past 30 years since the last big surge of interest in optical neural networks. As examples, Samsung now offers a camera with 200 million pixels[27], and 400-gigabit-per-second optical transceivers using on the order of 10 W of power are commercially available. This period has also seen the development and commercialization of photonic integrated circuits[28], giving a miniaturized alternative to bulk optics; there have also been substantial developments in optical materials and devices[29–35].

A complementary trend in the electronics community (both in CMOS and beyond-CMOS technologies), which has provided further support for the development of optical computers for neural networks, has been the development of special-purpose electronic chips for neural-network processing[36]. In many cases, these chips also perform analog rather than digital matrix–vector multiplications; this fact has led to the development of methods for training neural networks to work well on analog hardware, many of which are also applicable to analog optical neural networks. Both analog and digital-electronic

neural-network chips often have dataflow architectures, especially systolic-array architectures. They also often implement the concept of compute-in-memory, meaning that the physical element storing an element of the weight matrix of a neural network, for example, is also the physical element in which the multiplication by that weight takes place[37]; often, the stored values can only be updated slowly, but this is acceptable for neural-network inference or other scenarios in which the weights will be re-used many times. Systolic-array and especially compute-in-memory architectures can have a close mapping to optical processors in which information encoded in optical signals flows through processing elements, be they arrays of spatial light modulator pixels[38], meshes of Mach–Zehnder interferometers[39], crossbars of phase-change-memory cells[40] or networks of microring resonators[41]. This parallel between the architectures of analog electronic neural-network processors and analog optical neural-network processors has allowed optical-computer architects to borrow insights from the electronic-processor community. Architectural similarities also make it easier to predict how the performance of future electronic and photonic implementations are likely to compare. Not every optical computer for neural networks is based on similar architectures to electronic neural-network processors — and there are good reasons to deviate[17,42] — but in the cases in which the architectures and algorithms are comparable, performance analysis is simpler because one does not have to disentangle the effects of different algorithms and different architectures and can focus on the underlying physical differences: how many parallel elements are there, how fast can data be sent through them and so on. There are likewise architectural and algorithmic parallels between many special-purpose electronic processors for combinatorial optimization and optical approaches for the same application area[4].

In this Perspective article, we limit ourselves to discussing classical optical computing and do not review the benefits of optics for building quantum computers[43]. We will also not attempt to compare optical classical computers with optical quantum computers, other than to say that both are competing against classical digital-electronic computers but with rather different applications targeted for potential advantage[44]. We briefly discuss why electronic processors are hard to beat, before explaining what physics differences between electronics and optics can contribute to an advantage for optical computers. We then discuss strategies for optical processors to achieve advantage, before describing remaining challenges in the Outlook section.

What do optical computers need to beat?

Before we discuss how an optical computer could beat an electronic computer, let us first briefly describe what they are up against and why this makes electronic processors such stiff competition. There is both a hardware and an algorithm or software component to this. On the hardware side, electronic processors based on CMOS transistors have enormous parallelism, with up to ~10^{11} transistors per chip, operating at a clock rate of between ~1 GHz and ~10 GHz, and a switching energy of <10 aJ (that is, <10^{-17} J)[7]. These features allow modern processors to have enormous computing throughput — for example, the Nvidia H100 processor[45] can perform 4×10^{15} 8-bit scalar multiplications per second, which corresponds to performing approximately 4×10^{6} multiplications in parallel per clock cycle; the chip draws <1,000 W of power. On the software side, in parallel with >50 years of effort that has gone into improving transistor-based hardware, there has been >50 years of effort in designing algorithms, which in some cases has been responsible for almost as much benefit as improvements in hardware[22]. In many cases, the algorithms have been implicitly or explicitly designed to be optimized for the kinds of hardware that were or are available at the time[42], raising the barrier to entry for new hardware paradigms.

The 11 features

Paraphrasing journalist H.L. Mencken, there is an explanation for potential advantage of optical computing that is neat, plausible and wrong: the fact that light travels fast. We list below 11 features of either optics itself or of a way computing can be done with optics, which are ingredients for the construction of optical computers; these features allow for explanations of how optics can deliver an advantage that are subtler, but correct. We also address how the speed of light is related to optical computing, even though it is not the cause of optical advantage.

Bandwidth

Photonics has an ~100,000× larger bandwidth B than electronics. The bandwidth of photonics is ~500 THz, whereas for electronic circuits it is typically ~5 GHz (Fig. 1a). Small analog electronic circuits can have bandwidth >5 GHz (refs. 46,47) and small digital-electronic circuits can be clocked at rates >5 GHz, but both analog and digital electronics for computing systems tend to be limited to speeds ≪ 5 GHz by wire delays[48,49] and, since the mid-2000s, also by power dissipation[23]. The large bandwidth of photonics leads to two potential benefits.

Massive frequency-multiplexing parallelism. For example, there can be >10^7 comb lines in a frequency comb[50] and >10^9 frequency modes in a long fibre-ring cavity; data represented in each comb line (frequency mode) can be acted on in parallel (Fig. 1b) — not just individually (that is, element-wise) but also with operations that, for example, add or multiply data in different frequency modes[17]. The parallelism of optical-frequency modes is commonly exploited in optical communications, in which wavelength-division multiplexing enables communication over a single-mode fibre at rates >10^{13} bits per second[51]. This technology can also be used for computing; for example, reservoir computing on coherent linear photonic processors has been achieved with a bandwidth of B ~ 5 THz (ref. 15).

Fast dynamics of optical systems. The dynamics of optical systems can be very fast, which can translate to very high operation speeds, which in turn can lead to higher computing throughput and lower latency: the limit in the delay for an operation, $\tau_{\text{delay}} \gtrsim 1/B$, can be ~100,000× smaller for optics than electronics if the full bandwidth of optics is used. (We expand on what we mean by throughput and latency in computing in the section on how optical computers might beat electronic computers.)

However, this perspective on potential optical advantage from bandwidth has some subtlety. For one, the bandwidth limit on τ_{delay} is only a limit and the delay can be substantially longer if the device has a propagation length such that the time taken for light to travel through the device is long compared with $1/B$ (that is, a speed-of-light limit begins to dominate). Note that when the delay from propagation dominates the total delay, it is still possible to benefit from the fast bandwidth-limited speed in throughput by pipelining[52] — for instance, by sending multiple optical pulses into the system spaced apart by more than the temporal pulse width ~ $1/B$ but by less than the propagation delay. However, as electronic computers can — and generally do — also take advantage of pipelining, care again needs to be taken in making performance comparisons.

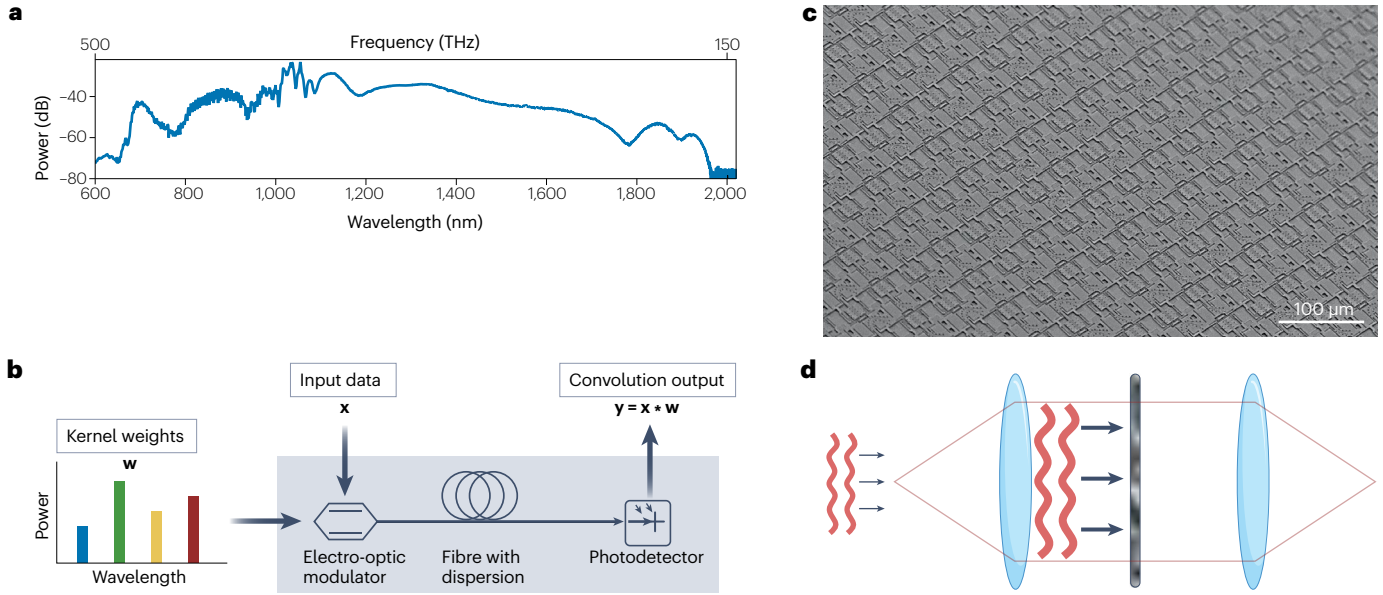

Fig. 1 | The three features most likely to have a key role in any future optical processor that does deliver an overall advantage in latency, throughput or energy efficiency. a,**b**, Bandwidth. An optical signal with bandwidth >300 THz (part **a**) and an example of the use of frequency multiplexing in optical computing (part **b**): kernel weights for a convolution are input as intensity modulations of spectral lines in a frequency comb; the use of multiple comb lines allows multiple computations to be performed in parallel. **c**, Spatial parallelism. Part of a state-of-the-art silicon-photonic device with 16,384 pixels on a 10 × 11 mm^2 chip, illustrating the degree of spatial parallelism possible in modern photonic devices. **d**, Nearly dissipationless dynamics. An example of computing with linear optics: light propagating through a lens undergoes a Fourier transform, and in a two-lens 4*f* system with a scattering medium in between, a convolution is performed on the input light. In the absence of optical loss (as would arise from absorption in the lenses, for example), the computation of the convolution happens without any energy loss. However, if one considers how to use this building block in an end-to-end computing system, there is typically an energy cost associated with converting an electrical signal into an optical input, and there is also typically an energy cost associated with converting the optical output back into an electrical signal. Part **a** adapted with permission from ref. 146, Optica Publishing Group. Part **b** adapted with permission from ref. 97, Springer Nature Ltd. Part **c** adapted with permission from ref. 147 under a Creative Commons licence CC BY 4.0. Part **d** adapted with permission from ref. 1, Springer Nature Ltd.

Another subtlety is that the delay for an individual modern electronic transistor under typical load is ~1 ps[53] so if one compared photonics with electronics at the level of an individual switch, the bandwidth benefit of optics would be much smaller than ~100,000× (perhaps 'only' ~1,000×). At the level of an entire chip, electronic processors are clocked ~10–100× more slowly than the circuit delays[54] would suggest are possible, largely owing to limits on power dissipation[23]. By contrast, photonic processors can have low dissipation (discussed subsequently). Thus, at a system level, it is a combination of both intrinsic bandwidth and low dissipation that gives rise to a ~100,000× potential system-wide bandwidth advantage for optics.

Reference[55] has demonstrated optical switching of ~46 fs pulses — highlighting the fast speeds possible with THz-bandwidth optical pulses and the quasi-instantaneous nature of nonlinear-optical operations.

Spatial parallelism

Photonic systems can exploit a large number (>10^6) of parallel spatial modes[56]. Consumer electronics using >10^8 spatial modes in an ~2.5-cm^2 area have been realized[27], illustrating that massive parallelism can be achieved in practice. Sophisticated integrated-photonics devices controlling many modes have also been created in academia (Fig. 1c).

For photonic systems in which light is confined in a single 2D plane, such as in 2D photonic integrated circuits, the density of photonic components can be as high as ~10^6 cm^{-2} (ref. 57), and we can roughly think of each component as enabling one or more computing operations (such as a multiplication) to be performed in parallel. There are multiple reasons to write one or more operations and not just exactly one operation. For example, one is that a single component in space can act on many frequency modes in parallel, as mentioned earlier, or on multiple polarization modes. Another is that depending on one's definition of an operation, and one's definition of a single component, a component may naturally perform multiple operations in a single pass of light through it, such as a single 50:50 coupler arguably performing two multiplications and two additions.

Although this component density is in absolute terms a high number, we should compare it against the spatial parallelism available in CMOS electronics, in which the achieved density of transistors is ~10^{10} cm^{-2} (ref. 45). As another point of comparison, to give an example of a candidate future electronics technology, an analog matrix–vector-multiplier core based on a crossbar array of phase-change memory, built by IBM[58], featured 65,536 phase-change-memory cells within a chip area of ~0.6 mm^2. This is a density of ~10^7 cells per cm^2, and each cell can be interpreted as performing one scalar, analog multiplication per clock cycle.

In this setting of 2D photonic integrated circuits, optics is at a disadvantage compared with electronics in the pure density of fabricable components, because the transistor density in electronics

is ~10^4× larger than the component density in on-chip photonics. This comparison is arguably the most relevant, as transistor-based electronic processors are, in most cases, the systems to beat. However, other comparisons can be made. Even 2D photonics can have a spatial-parallelism advantage over 2D microwave electronics: for example, photonic-crystal cavities (resonators) can have areas ~1 μm^2 (refs. 59,60), whereas electronic microwave resonators are typically orders of magnitude larger[61].

However, if the third spatial dimension is used[1,62], optics may gain a several-orders-of-magnitude advantage in spatial parallelism because electronics is in practice limited to very modest 3D integration. A typical modern electronic chip is thin — on the order of 1 mm — and comprises only tens of layers[63], whereas optical processors that are centimetres or even metres thick, using propagation through bulk crystals[62,64] or multimode optical fibre[16], for example, have been constructed. However, in the specific case of NAND memory, electronic integrated circuits have been scaled to 128 layers[65] — which suggests that for memory rather than computing, photonics has less room for advantage over electronics by extending in the third dimension.

Let us use an example to make a rough estimate of the kind of advantage that is in principle possible for 3D optical computing. Consider a 2D photonic device with dimensions $L \times L$ and a 3D photonic device with dimensions $L \times L \times L$. Assume we address each device with light of wavelength $\lambda \approx 500$ nm and that the device length is $L \approx 5$ cm. The number of resolvable spots in the former case is on the order of $(L/\lambda)^2 = 10^{10}$, whereas the number of resolvable voxels in the latter case is on the order of $(L/\lambda)^3 = 10^{15}$ — an advantage of $(L/\lambda) = 10^5$ times when going from 2D to 3D. We can also compare these numbers with the counts of transistors in electronic processors: at the state-of-the-art fabrication density of ~10^{10} transistors per cm^2, a 5 cm × 5 cm-chip would have 2.5×10^{11} transistors. This figure is an order of magnitude greater than the number of resolvable spots in the same-area photonic device, but several orders of magnitude smaller than the number of voxels in the same-length 3D device. Of course, an addressable voxel of material is not the same thing as a transistor; one ultimately needs to carefully analyse the computation and memory that is achieved using a particular device in a particular way, but these crude estimates hopefully convey two key messages: that by going from 2D to 3D devices, there can be an orders-of-magnitude increase in the achievable complexity of the device stemming from the fact that L/λ can be a large number and that although 2D photonic devices offer lower spatial parallelism than transistor-based electronic chips, moving to 3D devices may enable an orders-of-magnitude benefit in spatial parallelism for optics over electronics.

There is an important additional perspective on spatial parallelism: it is not only the density or number of components that can be fabricated that is important but also how many of the components one can in practice use in parallel. In other words, increased component density does not necessarily translate to proportionately greater computing performance. Modern CMOS electronic processors are typically only able to switch a small percentage (in one example, 3%[66]) of their transistors in a single clock cycle, largely owing to limitations in cooling[52]. When taking into account how many components can actually be operated in parallel with the constraints of power dissipation (discussed in the next section), 2D photonic integrated circuits may be at less of a disadvantage in spatial parallelism compared with electronic integrated circuits than the fabricated component densities alone would suggest.

As an example of spatial parallelism in optical computing, free-space optical processors have been prototyped using commercial spatial light modulators, which have ~10^6–10^7 controllable pixels — making them useful tools in building highly parallel systems[67]. Computation of ~5×10^5 scalar multiplications in parallel per pass of light through an optical setup with ~5×10^5 pixels has been achieved[38,68]. For applications in which the programmability of spatial light modulators is not required (such as in neural-network inference), fabricated metasurfaces offer a route to even larger parallelism: on the basis of the linear-with-area scaling of the space–bandwidth product of imaging systems[69], we expect it to be possible to create metasurface-based matrix multiplications or convolutions with >10^9 preprogrammed pixels (parameters) using ~10 × 10 cm^2 (ref. 70).

Nearly dissipationless dynamics

Photons can propagate through free-space optical setups with nearly no energy loss and perform computation by their mere propagation. (They can even propagate with nearly no energy loss in some on-chip setups: for example, thin-film lithium niobate chips can have waveguide propagation losses of 0.06 dB cm^{-1} (ref. 71)). How much computation is performed? We consider the cases of linear-optical and nonlinear-optical systems.

Linear optics. An example of computation by propagating light is that a single lens effectively performs a 2D Fourier transform on light that impinges on it[72] — optical correlators[13] and convolutional layers in optical neural networks[1] (Fig. 1d) both take advantage of this phenomenon. More generally, propagation of light through a linear-optical system can be modelled by a matrix–vector multiplication, so matrix–vector multiplication can be performed by merely shining light encoding a vector (of dimension N) in its spatial pattern onto an optical system[1].

As a rather extreme example, shining light through white paint can be used to perform the multiplication of a vector by a random matrix with dimension >$10^6 \times 10^6$ (ref. 73). In that example, the matrix is fixed and random, but various linear-optical systems in which the matrix can be programmed have also been demonstrated[1,57], although in these cases the matrix size has generally been limited by the number of programmable elements that can be engineered. An example programmable element is a pixel of a spatial light modulator, which can be used to represent a single programmable element of a matrix; spatial light modulators with ~10^7 pixels are commercially available. In principle, the dissipationless nature of optical propagation can lead to matrix–vector multiplications being performed that beat the Landauer limit[74] for multiplications performed on digital-electronic processors — intuitively, because in a coherent setup, the optical interference that occurs is a reversible process[75].

For the sake of concreteness, we have discussed examples of vectors encoded in space, but this is not the only possibility: the propagation of light in just a single spatial mode can also result in nearly dissipationless computation of inputs encoded in other ways, such as in frequency or time[2].

Nonlinear optics. Nearly dissipationless dynamics that can be harnessed for computation can also be seen in light propagating through nonlinear-optical systems. For example, propagation of light through an optical medium with a non-zero second-order nonlinear-optical susceptibility, $\chi^{(2)}$, can in general result in sum-frequency-generation and difference-frequency-generation processes, in which the optical amplitude of the output scales as the product of the amplitudes of light at two frequencies at the input, for instance, $E_{out}(\omega_1 + \omega_2) \propto E_{in}(\omega_1)E_{in}(\omega_2)$[76]. We can interpret such a nonlinear-optical process as performing a scalar multiplication

of the two numbers $E_{in}(\omega_1)$ and $E_{in}(\omega_2)$[17]. Nonlinear-optical dynamics enable the implementation of mathematical functions that are nonlinear – which is essential in deep neural networks[77] and in computing more generally[78]. For example, in a $\chi^{(2)}$ process, if the frequencies of the input light are equal ($\omega_1 = \omega_2$), then one may obtain output light at twice the frequency with amplitude $E_{out}(2\omega_1) \propto (E_{in}(\omega_1))^2$, so the function realized is $f(x) = x^2$, which is nonlinear.

Furthermore, just as the propagation of multiple spatial beams through a linear-optical system can be seen as performing a matrix–vector product, propagation of multiple spatial beams through a nonlinear-optical system can realize a higher-dimensional generalization of matrix–vector multiplication, namely, tensor contraction involving tensors of order $n + 1$, in which n is the order of the nonlinearity-optical susceptibility, $\chi^{(n)}$. This is an impressive feature for computing[16,17]: with the lowest-order nonlinearity, $n = 2$, the computation performed – by the mere propagation of the light through the system – is a tensor contraction that comprises ~N^3 multiplication operations, in which N is again the number of spatial modes. Higher orders of optical nonlinearity can result in even larger amounts of computation being performed by a single pass of light through the system, as even higher-order tensors are involved.

Benefits. There are benefits to the fact that computations can be performed nearly dissipationlessly in optics. The first is that one can potentially harness dissipationless dynamics to perform computation using less energy than would have been needed in a different platform that did have substantial dissipation (such as electronics).

A second benefit is higher performance. Dissipation does not only cause a computation to cost more energy, but can also limit the clock speed and parallelism of a processor, ultimately limiting its total computing throughput (operations per second) and latency. Modern CMOS electronic processors are limited – both in clock speed and in 3D density of transistors – by the ability to extract dissipated heat from them[23]. By markedly reducing dissipation per computing operation, one potentially allows for a marked increase in both the clock speed and spatial parallelism (number of operations performed simultaneously per unit volume).

In the context of 3D chips, photonics has another potential benefit over electronics with regard to dissipation: although the loss of electrical energy in a chip is generally by the generation of heat at the point where the energy is lost – in resistive heating of a wire, for example – the situation in photonics can be quite different because the loss of optical energy is often not due to absorption and accompanying generation of heat, but rather by scattering. This is true for waveguides in silicon-photonics integrated circuits, for example, and suggests that if one constructs a 3D silicon-photonic chip, the losses of waveguides within the chip will primarily not cause heating, but instead will result in photons being scattered within the chip until they emerge at the surfaces. In summary, nearly dissipationless dynamics in optics makes it possible to create 3D photonic chips that do not suffer from the extreme heat-extraction challenges of 3D electronic chips, and even the small photonic dissipation that does occur does not cause heating within the bulk of the chip if it is due to scattering, so we may not even need to worry about the residual photon loss causing heat-management difficulties provided that components that absorb photons are avoided.

There is, however, a snag to these benefits, namely, input/output costs: how does the input data for the computation get loaded and the result get read out? If the input comes from an electronic memory and the result needs to be stored in an electronic memory, then even though the computation itself can happen nearly 'for free', one needs to convert electronic data to the optical domain for the data input, and then convert the optical answer back to the electronic domain. This memory access and transduction, which typically also involves digital-to-analog and analog-to-digital conversion, will cost substantial energy (and be limited in speed when compared with optical bandwidths of terahertz).

Fortunately, this energy cost only scales as the size of the input vector, N, whereas the amount of computation being performed may scale as N^2 (linear propagation) or N^3 (or even higher powers; nonlinear propagation), and so for sufficiently large N, the energy cost of the input and output will be small compared with the cost that the computation would have required in an electronic processor. Similarly, the time required for input and output for N-dimensional vectors can, for sufficiently large N, be very small compared with the time the N^2-complexity or N^3-complexity computation would have taken on an electronic processor. The loading of coefficients, such as the matrix elements in the case of linear propagation, in general, also has a cost in both energy and time, but this can be amortized over many runs, such as in the case of batched inference with neural networks[39].

Low-loss transmission

The energy cost to transmit information 'long' distances with light is much lower than that with electrical signals[79], mostly because signal attenuation (energy loss) per unit length is much higher in electrical wires than in optical fibres or waveguides (Fig. 2a). There are several subtleties in evaluating the energy cost of optical and electrical communication, discussed in detail in refs. 48,79,80, which necessitate the use of the word 'mostly' here. For one, optical communications between electronic devices require transduction of signals from electrical to optical, and back to electrical, and the transduction devices have energy costs[79]. For another, electrical signal transmission along a wire requires energy that increases with length because the resistance of the wire increases with length – but this is not the end of the story: for thin wires, such as those used in CMOS electronic processors, the wire delay grows quadratically with length and to mitigate this, repeaters are used to regain a linear scaling of delay with length, and the repeaters also have an energy cost (associated with the switching of their driver transistors)[48,80].

For on-chip photonic processors, commercial foundries such as AIM Photonics can produce silicon-nitride waveguides with losses ~0.06 dB cm^{-1} for wavelengths ~1,600–1,640 nm and <0.25 dB cm^{-1} across the telecommunications C band (~1,530–1,565 nm)[81].

An important caveat for both free-space and on-chip optical processors is that although propagation losses between components can be very low, typically there are losses from reflections or scattering as light propagates into or out of a component (such as Fresnel reflections owing to mismatch in refractive index). As a result, optical processors still need careful design to avoid excessive overall optical loss.

The low-loss transmission of optics is already being taken advantage of in electronic computing: optical links in data centres[82], and even directly between chips[83], use light to communicate information over length scales from centimetres to many metres. It is anticipated that even some communications within a single chip might eventually use optics[79,82].

A major reason that light is not already used for communications within single electronic-processor chips, especially over very short distances, is that the optoelectronic components to transduce signals between the optical and electrical domains cost both space and energy, and it is only worth paying these costs when the distance the signal needs to travel is long enough[79]. An optical computer, however, could

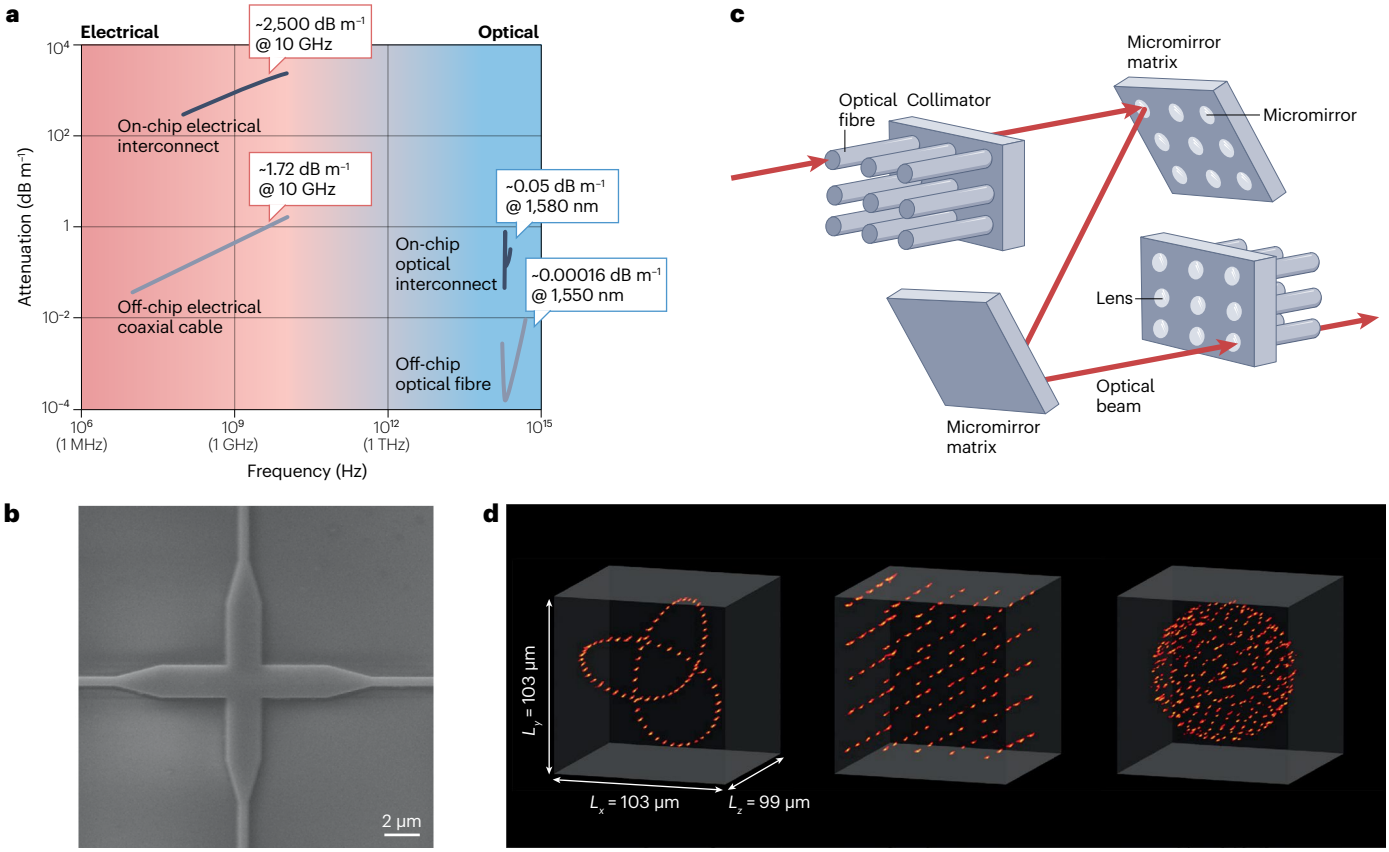

Fig. 2 | Signal transmission in optical systems. a, Low-loss transmission in optical systems. For both on-chip and off-chip transmission, the signal attenuation (in dB per metre) is orders of magnitude lower (better) with optical instead of electrical signals. For example, electrical signals at 10 GHz have ~10^4× higher attenuation than equivalent on-chip or off-chip transmission with optical signals. Inspired by ref. 148, Fig. 4.3. Data sources: on-chip electrical interconnect: ref. 149; off-chip electrical coaxial cable: ref. 150; on-chip optical interconnect: ref. 151; off-chip optical fibre: ref. 152, Fig. 22.2 and ref. 153. This figure is intended to give a heuristic comparison; it does not comprehensively cover all transmission technologies, but is based on just a few illustrative examples that convey the relevant orders of magnitude. For more examples and details, see: ref. 154 (electrical interconnects and cables); ref. 149 (on-chip electrical interconnects with different dimensions); ref. 155 (electrical interconnects on printed circuit boards) and ref. 156 (integrated-photonics waveguides with lithium niobate). **b**, Optical beams and 'wires' can cross. It is in free space that optical paths can cross: in integrated photonics, waveguides can pass through one another with minimal impact on the signal propagation. The waveguide crossing in this image had a crosstalk of less than −50 dB. **c**,**d**, Optical beams can be steered programmably. Optical beams inside a micro-electro-mechanical systems optical switch can be rerouted on timescales on the order of milliseconds using arrays of micro-electro-mechanical systems-actuated micromirrors (part **c**). Optical-tweezer beams can be reconfigured to trap atoms in arbitrary geometries in 3D (part **d**); the results shown here are from an experiment in which a liquid-crystal-based spatial light modulator was used to programme the beams; such modulators can also be updated on a timescale on the order of milliseconds. Part **b** adapted with permission from ref. 157 under a Creative Commons licence CC BY 4.0. Part **c** adapted with permission from ref. 158, IEEE. Part **d** adapted with permission from ref. 159, Springer Nature Ltd.

in principle take advantage of optics for low-energy cost, nearly dissipationless information transmission at all length scales, and without paying space or energy costs for transduction — because the signals would already be optical. Note however that an optical processor will inevitably need to use some energy for transduction, for example, to load the initial input data for the computation and/or to read out the final answer, which will typically need to be in the electrical domain. But the transductions — and their costs — that would have occurred within a computation can be avoided.

Optical beams and 'wires' can cross; electrical wires cannot

In many cases, there is negligible optical nonlinearity — not only in free-space settings but also in materials when the optical power is low and the propagation length is short; informally: we do not have lightsabers in ordinary optical situations[84]. In these cases, optical beams can pass through one another without suffering from crosstalk. Likewise, optical on-chip wires (waveguides; Fig. 2b) can cross with very low crosstalk — not just in principle but also in practice in the presence of fabrication imperfections. By contrast, electrical wires need their own region of isolated physical space and, in addition to not being able to pass through one another, also often suffer from crosstalk even if they are merely close to one another[85].

This difference provides the possibility for photonic processors to be more compact than electronic processors when interconnect is an important contributor to processor size, although the use of optical beams for communicating information is not without its own crosstalk

challenges owing to diffraction, scattering and unwanted reflections[86]. (One might also wonder about the size of optical beams compared with electrical wires, as optical beams or waveguides are limited to sizes on the order of a wavelength, whereas electrical wires can be made only nanometres wide. However, interconnects in electronic processors have trace widths and spacings on the order of 1 μm[87], which is a design choice in part motivated by the fact that the resistance of a wire decreases as its cross-sectional area increases[48].)

One can interpret the ability for optical beams to cross as a key enabler of many free-space, spatially multiplexed optical implementations of convolution and matrix–vector multiplication[1]. For example, in implementations[88] of matrix–vector multipliers that use arrays of lenses for fan-out (Fig. 3b), the rays between the input vector and the fanned-out copies cross. The crossing supports the implementation, in principle, of large convolutions and dense matrix–vector multiplications in small volumes. Optical switches (Fig. 2c) provide another example in which crossing of beams enables a more compact design.

Optical beams can be steered programmably at high speed; electrical wires are either fixed or reconfigurable only slowly

Free-space optical beams can readily be redirected (for example, using an acousto-optic deflector, with a delay on the order of microseconds) (Fig. 2d), enabling the creation of reconfigurable optical interconnects[89,90] (Fig. 2c). By contrast, electrical wires on chips are fixed at the time of fabrication, and wires joining nodes in an interconnect between processors, boards or racks can only be moved slowly (typically on the order of seconds). Electronic processors typically mitigate the disadvantage of having a fixed network by using multihop communications — relying on there being a path between a sender and a receiver involving some intermediate nodes — and switching, which achieves fast rerouting of signals within a fixed network topology. These strategies come with the cost of increased latency and potential bandwidth bottlenecks.

Fan-in (summation) and fan-out (copying) work differently in optics

Copying data to be processed in parallel (fan-out) and summing the outputs from a number of parallel-processing units (fan-in) are important primitives in parallel processing. Both can be implemented in optics in a different way to electronics and have different tradeoffs[89,91]. Optics has a potential advantage from supporting large (>1,000) fan-in and fan-out without the *RC* and *LC* delays of fan-in and fan-out with electrical wires, for which fan-in and fan-out are typically kept lower than 10 in digital processors, necessitating multiple buffering stages (and hence further delay) whenever larger fan-in/fan-out is needed[92,93]. Note that as ref. 89 points out, when evaluating an optical scheme, one needs to take care to evaluate the *RC* and *LC* delays of photodetectors that are involved.

In free space, fan-in of signals encoded in spatial modes can be performed by directing beams to a common point in space (via the use of a lens, for example; Fig. 3a), at which there could be, for example, a photodetector (if the next processing step required conversion from optical to electrical signals), a holographic element (to combine the beams travelling in different directions into a beam that travels in one direction, albeit at the cost of loss of optical power)[89] or an intensifier (which can amplify the summed beams and re-emit a single optical signal)[88].

Fan-out of a signal in a single spatial mode to multiple spatial modes can also be performed conceptually easily in free space, where it happens essentially without any special engineering effort (Fig. 3b): imagine an optical display (such as a light-emitting diode display on a cell phone) that emits in multiple directions — multiple people looking at the display from different vantage points can all see the same image, and we can interpret what happened is that multiple copies of the data on the display were made and transmitted to different receivers. Another example of optical fan-out in everyday life is in a kaleidoscope. Arrays of lenslets (microlenses) can be used to collimate the image copies[88,94]. Free-space fan-out can also be implemented and understood in the Fourier domain[95].

Both fan-in and fan-out for spatial modes can also readily be implemented in integrated-photonics platforms[96]. However, in an on-chip setting, light propagation is typically practically restricted to be in a single plane, whereas in free space it is natural for signals to propagate in all three dimensions, enabling a much higher degree of fan-in and fan-out. For this reason, it is easier to imagine gaining an advantage over on-chip electronic processors (which are also quasi-planar) from the use of optical fan-in or fan-out in free-space settings.

So far, we have discussed fan-in and fan-out in the context of spatial modes. For optical computers using frequency or temporal modes, fan-in and fan-out may be realized using other means. For example, fan-out of data input as electronic signals can be performed in the frequency domain by modulating an optical-frequency comb[97], and weighted fan-in can be performed using wavelength-division multiplexing, including in on-chip platforms[2].

To reason about why or when optical fan-in or fan-out may have an advantage over electrical fan-in or fan-out, it is useful to consider the bandwidth and low-loss transmission possible in optics and that optical beams can cross. However, the fan-in/fan-out possibilities of optics are distinct from the potential benefits of bandwidth, low-loss transmission and beam-crossing in optics, and it is fruitful to think of fan-in and fan-out in optics as special features that can be used in an optical-computing architecture, even though they may also use other features of optics to operate well.

Indeed, teasing out the source of a potential advantage can be quite subtle. For example, fan-in arguably has an important role in enabling vector–vector or matrix–vector multiplication engines that use extremely small amounts of optical energy per multiplication[38] — in which the amount of optical energy needed to achieve a particular signal-to-noise ratio for a vector–vector dot product is fixed regardless of the vector size — but similar efficiency can be achieved with optoelectronic fan-in, in which summation is performed in the electrical domain[75,98]. Purely analog electronic approaches to compute vector–vector dot products can also show favourable energy consumption compared with digital-electronic approaches[99], so for any given computing scheme using optical fan-in, one can ask: which part of the potential benefit comes from performing the summation in an analog rather than digital fashion, and which part comes from using optics instead of electronics?

One-way propagation

One can readily construct optical systems in which the propagation is naturally one-way (if one, for example, forms an optical cavity in part of the system, then the situation becomes more complicated). By contrast, electrical signals can propagate backwards (Fig. 3c). In electronic processors, backwards propagation (from inputs to other inputs, or from the output to the inputs) can cause unwanted dynamics as well as unnecessary power consumption. This difference leads to an advantage of optics over electronics for some analog architectures.

Although backwards propagation is a general feature of electrical circuits — without isolating elements such as buffers or diodes

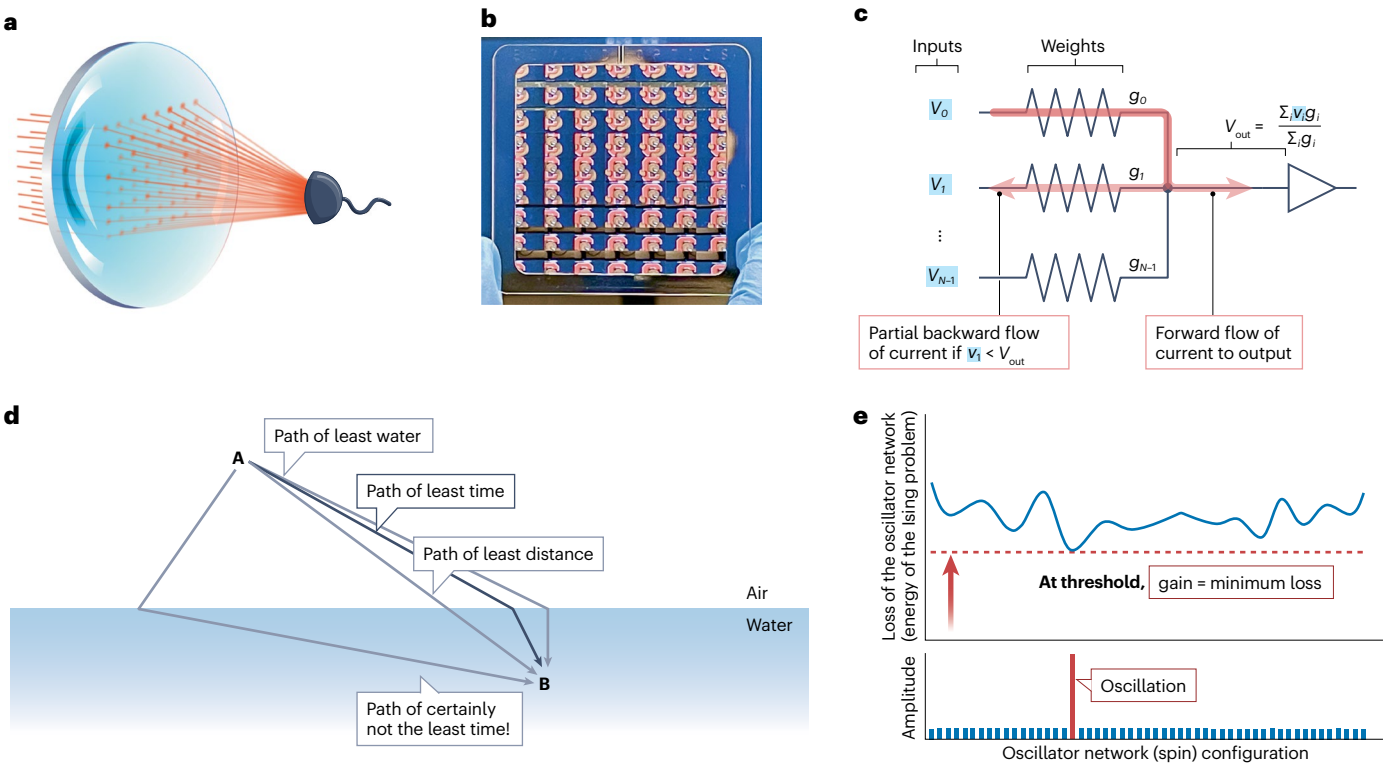

Fig. 3 | Additional ways that optical systems are different from electrical systems. a,**b**, Optical fan-in can be performed in free space using a lens (part **a**); here, a lens causes many beams to converge on a single-pixel detector. Optical fan-out can be performed in free space using an array of lenses, in which each lens 'captures' a copy of the incoming image (part **b**). **c**, An electrical fan-in (weighted sum of voltage inputs v_i by conductance weights g_i) exhibiting undesired backward flow of current. The current contributions from the input v_0 to the output (desired) and, if $v_1 < v_{out}$, from the input v_0 to the input v_1 (undesired) are shown in pink. Only the current contributions from v_0 to the output and to v_1 are illustrated here, but in general current will flow backwards from the common node v_{out} to v_i if $v_i < v_{out}$. By contrast, one-way, forward-only propagation of light in a fan-in is shown in part **a**. **d**,**e**, Optimization principles. The principle of least time in optics (part **d**). Light travels between starting point A and ending point B by taking the path of least time. A computational interpretation is that the light solves an optimization problem (of finding the path of least time), given the constraints of where the path starts and ends. A network of oscillators (part **e**) – which in optics could, for example, be optical parametric oscillators or laser oscillators – will in principle oscillate in the collective mode/configuration corresponding to the lowest loss if the gain is set to be equal to the minimum loss. Part **a** adapted with permission from ref. 38 under a Creative Commons licence CC BY 4.0. Part **b** courtesy of Mandar Sohoni and Tianyu Wang. Part **c** adapted with permission from ref. 100, IEEE. Part **d** adapted with permission from ref. 105, Princeton Univ. Press. Part **e** adapted with permission from ref. 107, Springer Nature Ltd.

in a circuit, any time there is a voltage difference between two connected circuit nodes there will be a current flow between them, even if those two nodes are inputs – concerns about backwards propagation have arisen mostly in the context of analog crossbar-array processors, related to their fan-in stage[100] and also the sneak-path issue[101]. Analog optical matrix–vector-product engines[1] generally feature one-way propagation, avoiding some of the issues that arise in analog electronic matrix–vector-product engines (that is, crossbar arrays), and there is a broader notion of optics providing natural isolation[102] that can be useful in computing.

A caveat is that although perfectly one-way propagation is possible if light does not pass through any interfaces, any useful optical processor involves at least some interfaces (light going from air into a glass lens, for example), and as a consequence have some unavoidable reflections. The reflections can be made small by appropriate choices of geometry and materials but will never be completely eliminated. In many cases, there may be an engineering tradeoff between, for example, the compactness of the optical processor and the magnitude of the reflections (in other words, the one-way-ness) in the system.

Different realizations of adiabatic, least-action and least-power-dissipation principles

There are general physics principles – such as adiabaticity, the principle of least action and the principle of least energy dissipation – that can lead to a physical system heuristically solving optimization problems[103]; variations of these principles can be leveraged to construct optimization machines (such as Ising machines[4]). Given how central optimization is in machine learning, and especially in neural networks, computers designed to perform optimization are often also well suited to perform machine learning – so an advantage on optimization can quite plausibly be translated into an advantage in machine learning too. Similarly, one can recast the problem of solving partial differential equations as a variational optimization problem[104], providing another potential application of physics optimization principles to a broader class of computations.

For example, Fermat's principle of least time for optics states that light follows the path that minimizes its time to travel between two points (Fig. 3d). Feynman gave an explanation of this principle with a path-integral formulation in which the light can take all possible paths but only the paths that constructively interfere contribute substantially, and paths with substantially different propagation times than Fermat's solution destructively interfere[105]. This perspective is possibly helpful for thinking about how to design optimization machines that use Fermat's principle. By contrast, Fermat's principle does not have a direct analog in electrical circuits — so a computer performing optimization using Fermat's principle is more natural to try to create with optics.

Onsager's principle of least energy dissipation can apply in both optics and in electronics, but the behaviour and resulting computing performance may be different because of differences in the underlying physics. For example, lasers and parametric oscillators in optics have a threshold when gain is equal to loss, and the fact that they first oscillate in the mode with lowest loss can be used to design optical Ising machines[106,107] (Fig. 3e). Electrical circuits, including oscillators, also have dynamics that heuristically minimize the energy dissipated[103], but they are not identical to lasers or optical parametric oscillators and in general have different behaviours.

It is an open question whether, or in which situations, optics systems using Onsager's principle have an advantage over electronics realizations, but the possibility is one that a designer of an optical computer may wish to explore. The question has multiple facets: if the equations governing the optics and electronics dynamics were identical, one might still achieve an advantage of optics over electronics for some of the other reasons described in this article, such as bandwidth. However, one can also ask whether the differences between the underlying equations lead to different behaviours beyond a faster timescale resulting from higher bandwidth, or a larger system size resulting from larger spatial parallelism — in other words, differences beyond the other optics versus electronics distinctions drawn so far.

The quantum nature of light is accessible at room temperature

It is possible to store and process information encoded with single optical-frequency photons, and it is possible to detect individual optical photons with low noise, all at room temperature. This is in contrast to the situation at microwave frequencies, in which thermal noise at room temperature rapidly swamps any information stored in single photons, and low-noise single-photon detection is not available (Fig. 4a). The quantum nature of microwave photons is accessible at temperatures ~10 mK, but such cold temperatures are generally only achievable using a dilution refrigerator, which is bulky and expensive (in money and energy).

For classical information processing, the fact that small numbers of photons can be manipulated and measured naturally leads to a potentially lower energy cost than if more photons were needed for reliable operation[38,75]. It is also possible to produce and measure squeezed states of light at room temperature[108]; the reduced noise in squeezed states could prove useful in classical information processing, for example, for achieving higher numerical precision with a fixed energy budget (average number of photons).

The lack of a strong single-photon nonlinearity in optics, which is an advantage for communicating information without crosstalk but can be a disadvantage for processing information with small numbers of photons, can be circumvented using single-photon detection (Fig. 4b). The nonlinearity of the detection process itself is a feature one can use[1,75], but it is also possible to use photodetection to probabilistically induce nonlinear operations across multiple optical modes[109]. Reference[109] develops and motivates probabilistic nonlinear operations for use in quantum computing, but these operations could potentially also be used for classical computing .

In this Perspective article, we do not consider quantum information processing[110]; here, when we talk of operating in the quantum regime, we mean in the sense that light comprises photons and we are operating at such low powers that the quantum noise and discrete nature of the light are relevant to modelling the operation of the

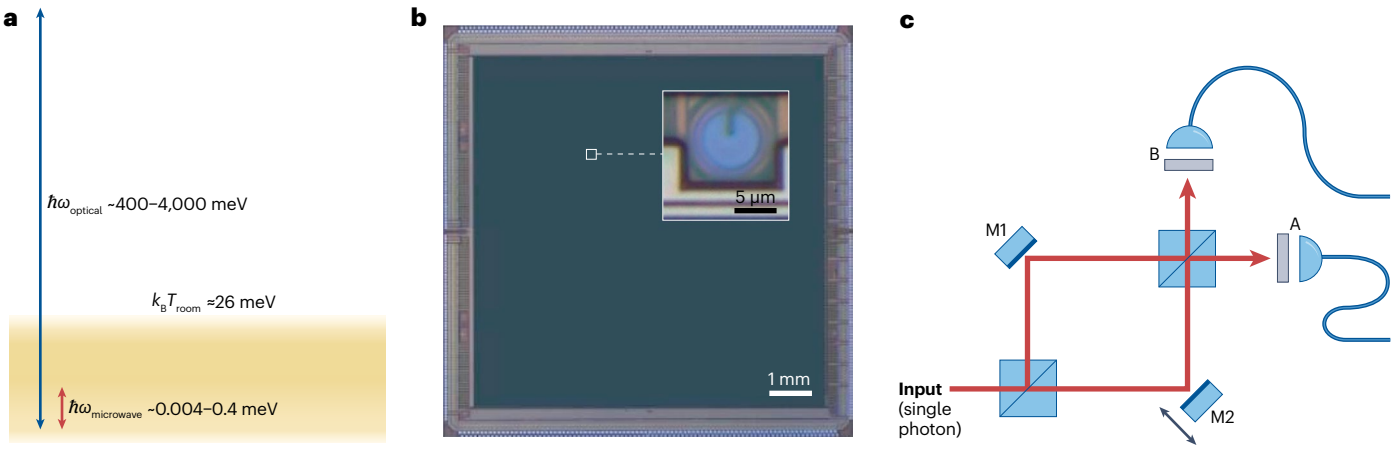

Fig. 4 | The quantum nature of light. a, The energy of optical photons is much higher than that of the thermal energy scale k_BT at room temperature ($T_{room} \approx 300$ K), whereas microwave photons have much lower energy than k_BT_{room}. Consequently, thermal noise 'drowns out' quantum effects of microwave signals at room temperature, but quantum effects in optical signals can be observed. **b**, An array of 250,000 single-photon detectors, which is sensitive to light at visible wavelengths and operates at room temperature. **c**, Wave physics. Interference can be observed in a Mach–Zehnder interferometer with only a single photon input at a time. This schematic is from an undergraduate-laboratory experiment using just a few commercial optical components, highlighting the relative ease of observing wave phenonmena at the single-photon level with optics. (The counts at photodetector A oscillate as a function of the position of mirror M2, which controls a phase difference between the upper and lower arms of the interferometer.) Part **b** adapted with permission from ref. 160, IEEE. Part **c** adapted with permission from ref. 112, AIP.

computer. The topic of using quantum phenomena such as entanglement to build quantum computers is exciting but beyond the scope of this paper; ref. 111 provides a helpful description delineating the first and second quantum revolutions, and it is only the former that we consider here.

Wave physics

It is easy to observe the wave nature of individual photons — observing interference of single photons in a Mach–Zehnder interferometer is an undergraduate-laboratory experiment[112] (Fig. 4c), and photon coherence is well preserved in on-chip photonic processors[113] — but it is difficult to observe the wave nature of individual electrons. Even in advanced on-chip electron-transport experiments, the electron coherence length is less than ~250 μm, with values between 1 μm and 20 μm[114] more typical, and only at cryogenic temperatures.

The wave nature of electrons being difficult to observe and exploit is due to cryogenic temperatures being required — on-chip electron coherence lengths are also much more dependent on the properties of the material host than on-chip photon coherence lengths. For this reason, we are treating the accessibility of wave physics for photons as a separate advantage to the accessibility of their quantum nature, even though the wave–particle duality for both photons and electrons is part of quantum physics.

A counterpoint is that even though the wave nature of individual electrons is impractical to observe, wave phenomena of microwave signals in electronics can readily be observed and exploited for computation[115]. However, these are not wave phenomena of single electrons, but rather of signals that comprise many microwave photons. A key engineering consequence of this distinction is that electronic microwave signals have long wavelengths (for example, gigahertz signals have centimetre-scale wavelengths), which markedly limits the possible spatial parallelism relative to the parallelism possible with optical-frequency photonic signals — leading to a potential advantage of optics over electronics (and in particular, microwaves). Note that a completely different kind of microwave signal can also be created and used for computation: an acoustic wave at microwave frequencies[116]. These waves can have short wavelengths despite their low frequencies, but at the cost of propagating at vastly slower speeds than photonic signals — the speed of sound rather than the speed of light — which is a disadvantage for computing with them.

But not that the speed of light is fast

The speed of light is often brought up as a contributing factor for how optical computing will obtain a large speed advantage over electronic computers, but this is misleading because both optical and electrical signals can travel at roughly the same speed: in vacuum, light (and microwaves) travels at speed c; in silicon-photonic waveguides, light travels at speed ~0.4c[117]; in wires on printed circuit boards, signals can travel at speed ~0.43c[118]; and in CMOS electronic circuits, signals can travel at speed ~0.2c[79] or ≈0.5c in CMOS wires with careful design[80].

There is a mere 5× difference between the speed of light in vacuum and the speed of signal propagation in wires in CMOS electronic processors, so the speed of light is not a key distinction of optics. The notion of 'computing at the speed of light'[1] is more useful to think of as a goal for an optical computer, rather than a cause of advantage. The speed of light provides a physical limit on how fast a computer can operate[119] and one framing of the goal of the optical-computer engineer is to design a computer that leverages the benefits of optics (as discussed earlier) to reach this limit for a particular computing task, in as small a volume as possible, so that the total time for a computation is as small as possible.

This framing implicitly makes the goal about the latency of the computer (how long does it take for the answer to be output from the time the input is provided?) — which can be important, especially in real-time-computing scenarios — but often we are instead interested in improving its throughput or energy efficiency. Optimizing for throughput may involve trying to maximize the number of computing operations performed in parallel, and optimizing for energy efficiency may involve minimizing the dissipation in the system, neither of which have much to do with ensuring that the latency of the computer saturates the bound set by the speed of light. 'Computing at the speed of light' is not only a goal rather than a cause but it is just one of several possible goals for an optical computer.

Some of the items listed earlier are interrelated, and some of them even have a common physical root but are listed separately because the root leads to multiple features of light or has multiple consequences for computing. For example, the large bandwidth of optics relies on the large carrier frequency ω of optical signals. The wavelength of light λ is directly connected with its frequency ω: λ is proportional to $1/\omega$, so the large values of ω for light make it possible to achieve large spatial parallelism and to observe and exploit wave physics in small volumes. The fact that optical photons have a large energy $\hbar\omega$ relative to thermal energy k_BT at room temperature $T \approx 300$ K (k_B is the Boltzmann's constant) is directly responsible for the quantum nature of light being accessible at room temperature. Low-dissipation dynamics and transmission of information with optics are also connected with the short wavelength λ for optical photons, which allows tight waveguided confinement with nearly lossless dielectrics rather than with metals. So all six of these features are connected by the fact that ω is large, as multiple aspects of optical physics are influenced by the value taken by ω.

Not all of these features are equally important for obtaining advantage in optical computing but they are also not presented in order of importance, partially because determining such an order would require knowing what ingredients future optical computers will ultimately most heavily rely on. Nevertheless, in the next section, we discuss how these features may be used and opine on which ones are most likely to be critical.

How might optical computers beat electronic computers?

In this section, we describe some strategies for the design of optical computers that may enable them to have an advantage over electronic computers.

There are three main metrics of computing performance for which we might aim to achieve an advantage: latency, throughput and energy efficiency. Which of the three (or which combination) should be targeted in designing an optical computer depends on the goals of the user, but there are arguments for how optics could enable advantage in all three of these metrics.

Note that there are several other metrics of computers that are important, such as size, robustness, cost, security (susceptibility to hacking) and accuracy. We do not have any reason to believe that an optical computer could deliver superior accuracy, for example, than all possible electronic computers, so accuracy is not a metric we expect an optical advantage for, but instead we typically aim to achieve an advantage in latency, throughput and/or energy efficiency for a specified accuracy. Similarly, the other metrics provide other constraints

that an optical computer must satisfy to be competitive for some particular use case.

We now briefly describe these metrics using a particular computing example: machine-learning inference, more specifically, face recognition in an image. Latency (also called delay) refers to the time it takes for the computer to make a prediction of the name of the person in an image from the moment the computer is given the input image. Throughput refers to how many inferences can be performed per second; for face recognition in images, a throughput metric is images processed per second. Note that in general (1/Latency) ≠ Throughput; by pipelining[52], throughput can be much higher than the inverse of latency. As an intuitive example of this, consider a factory producing cars using an assembly line (pipeline): from start to finish, it might take the factory 1 day to manufacture a car (latency), but the total number of cars manufactured per day could be hundreds (throughput). Energy efficiency refers to how much energy is used by the computer to complete a single inference computation with a specified accuracy; for face recognition in images, an energy-efficiency metric is joules per image processed.

There may be tradeoffs when optimizing for these three metrics, so it is important to decide before starting the design of a computer what one's goals are. For example, although minimizing latency is sometimes the main goal (for instance, in high-frequency trading[120]), often improving the throughput of a processor or its energy efficiency is the more important goal — and in many cases the goal will involve all three metrics, such as maximizing throughput and energy efficiency, subject to the constraint that the latency meets a particular target (for example, in neural-network inference[121], where in many applications — such as language translation — we may require the latency to be <1 s).

Despite the fact that there are typically tradeoffs in the optimization of computer performance metrics (between latency and throughput, for example), the following strategies should help in designing a computer that optimizes any combination of latency, throughput and energy efficiency.

Avoid or mitigate input and output bottlenecks and overheads

Optical computers generally do not operate entirely with optics: typically some inputs to the computer originate in electronics, and/or the output from the computer is ultimately electronic. For example, if an optical processor is used for determining whether there is a pedestrian walking in front of a self-driving car, the output needs to be electronic so that it can be input to the control systems in the car, which can use the information to actuate the brakes. If the processor uses a neural network, the trained parameters for the neural network may well be stored in electronic memory and need to be input to the processor in some way. Unfortunately, the interfaces between optics and electronics can cause major bottlenecks in speed and be a major source of energy usage by a processor. For an optical processor to offer an advantage over electronic processors — in any of latency, throughput or energy efficiency — the processor architecture needs to be designed to minimize the negative impact of transduction between optical and electrical signals and the conversion between analog and digital signals.

To illustrate some of the challenges that can arise from optics–electronics interfaces, imagine an optical processor that intrinsically has a processing bandwidth of 100 THz. If data can only be input to the processor at a rate of 10 GHz, limited by, for example, the bandwidth of electro-optic modulators and digital-to-analog converters, then without careful design, the intrinsic bandwidth benefit of the optical system — which could have led to improved latency and/or improved throughput — may go to waste. Similarly, although an optical processor can be designed to perform computation on optical signals nearly dissipationlessly, there is an energy cost to optical–electrical transduction and analog–digital conversion for getting electronic data into and out of the optical processor, and these costs may be so large that they not only dominate the total energy cost of the optical processor but also make the energy cost so high that the processor is less energy-efficient than an all-electronic processor.

A crucial mitigation strategy is that inputted data should be re-used as much as possible — once both the time and energy penalties for sending electronic data into an optical processor have been paid, one would like to extract as much benefit as possible from those data. This applies both to data converted into optical signals and to data that may remain as electrical signals but that nevertheless has time and energy costs to be input to the processor. Re-use of optical signals can be enabled by various forms of optical memory[122], as well as by copying via fan-out. As a consequence, an optical-computer designer is usually motivated to make the fan-out factor be as large as possible. In an optical matrix–vector multiplier, fanning out 10^3 or more copies of the input vector is desirable and likely necessary to achieve a substantial advantage over electronics.

As an example of the re-use of electrical control signals, optical processors performing neural-network inference (as opposed to training) can load the neural-network weights into phase shifters that consume either little or no static power[1,39] and then use those weights many times by performing many inference computations with them (for example, by batching individual inferences[75]). This allows both the time and energy costs of loading the weights to be amortized. Another example of data re-use in photonic neural-network processors is in convolutional neural networks: the same convolutional kernel can be applied to many different subsets of the input data, so the kernel weights can — at least conceptually — be loaded once and used many times[1,40,97].

A general design principle is that — all else held equal — it is better to perform more computations per bit of input data. This principle is essentially the concept of maximizing arithmetic intensity in conventional computer architecture[52]. Data re-use is one way to achieve this, but an important complementary conceptual approach is to choose computational tasks such that the optical processor for that task performs computations whose complexity scales rapidly with the input data size. For example, a computation on input data of size N that requires only $O(N)$ operations is less attractive than one that needs $O(N^2)$ operations; a computation requiring $O(N^3)$ operations is even better. The cost in time and energy of inputting data of size N is generally $O(N)$, so if the computation performed by the optical system has complexity $O(N^2)$ (and we assume that, through a combination of the 11 features discussed earlier, the cost of this computation in optics is far lower than it is in electronics), then there exists some threshold size such that for any N larger than the threshold, the costs of loading the data can be compensated for by the benefits of doing the $O(N^2)$ computations optically — leading the optical computer to outperform electronic computers even when the data-transfer costs are considered.

A key practical fact is that for current speed and energy numbers for CMOS electronics, it seems likely that optical processors will need to support very large values of N (say, $N > 10^4$) to reach the crossover point where they start delivering a throughput or energy-efficiency advantage for computations on the basis of matrix–vector multiplication (which is an $O(N^2)$ computation, for square matrices)[21]. This fact

motivates both scaling optical matrix–vector-multiplication processors to large sizes and designing optical processors with computations that have complexity greater than $O(N^2)$. From this perspective, combinatorial optimization such as Ising solving[4] is an attractive problem for optical computing because the computing effort is generally expected to scale exponentially, that is, as $O(2^N)$, with respect to the number N of variables being optimized, and also with the amount of data required to specify the optimization problem — for example, an N-spin Ising problem is specified by $O(N^2)$ numbers.

When an optical processor loads data from electronic memory, there is not only a cost for the memory access — which an electronic processor would also have had to pay — but also there is a cost for transducing the data from an electrical to an optical signal, and potentially also a digital-to-analog conversion involved, which also has a cost. Because the cost of loading data is generally larger for optical processors than for electronic processors, there is a strong motivation to choose algorithms for optical processors that have higher intrinsic data re-use or higher algorithmic complexity. This kind of hardware–software co-design can lead to considerable improvements compared with fixing the algorithm on the basis of what works well on current electronic processors and trying to forcibly design an optical processor to work in the same way.

Although minimizing and compensating for the costs of loading input data are crucial, it is also important to avoid having the output of data be too costly in time or energy. It is similarly beneficial to minimize how much data needs to be output, by doing as much of the computation and data reduction within the optical processor as possible. This design principle motivates choosing algorithms that require a large amount of computation relative to the size of the output. As an example, this is typically true in machine-learning inference — where for the overall computation the answer may be just a few tens of bits, outputting the predicted class of the input data.

Do not try to directly take on digital-electronic processors at their own game

Arguably, the biggest challenge in building optical processors that surpass electronic processors in throughput or energy efficiency is overcoming the limiting performance of electronics-to-optics and optics-to-electronics conversion technology. If one starts with data in electronics — as is most typically the case — and wants the computed answers to end up in electronics — as is also most often the case — then one has little choice but to apply the strategies above and hope to be able to amortize the input/output costs. However, given how large state-of-the-art CMOS electronic processors are and that they have a home-ground advantage in working on data that are already in electronics, it seems likely that modern optical processors would not first gain an advantage as drop-in replacement accelerators in conventional electronic-processing workflows. Instead, one can target applications in which the inputs and/or outputs are naturally optical — and in this way eliminate the conversion costs.

Machine-learning application in which the input is conventionally an image from a camera is an example[1,88,123]: one can replace the camera and subsequent electronic neural network with an optical neural network that directly processes the scene in front of it, in applications such as self-driving cars[124], microscopy[88] or spectroscopy. It is not necessary to replace all the electronic image-processing computation with optics if the output is ultimately going to be electronic anyway — one can adopt the strategy of using optics to pre-process the optical image data[70,125], intelligently encoding it so that the output conversion from optics to electronics has much lower bandwidth than naively digitizing the images to begin with, which could lead to benefits in latency, throughput and energy efficiency[88].

Although image processing enables the elimination of the input conversion stage because the input can be directly optical, applications in which both the input and the output are optical may be even more promising for immediate attack. Optical communications have inputs and outputs that are both optical, but current approaches involve a number of stages at which optical signals are converted to electrical signals for electronic processing and then converted back to the optical domain. This makes optical communications signal processing a natural target for all-optical signal processing, which could reduce latency, increase throughput and improve energy efficiency[20,126–128].

Many neural-network models have become large enough that they can no longer practically be run on a single electronic processor, which has motivated the design of optical interconnects specifically for neural-network processing[129]. This trend provides another motivation for neural-network processing as an application for optical processors: if the electronic-processor competition needs to pay the relatively high energy costs of conversion between optics and electronics too, then these conversion costs are at least not an exclusive disadvantage of using optical processors. One can think of a single processor in an optically interconnected data centre for performing neural-network processing as a system whose inputs and outputs are both optical — so from this perspective, it is a promising candidate to try replace with an optical processor.

Combine multiple optical features to try gain an advantage

This point might sound trite, but it is important — any optical processor that has an advantage over the best equivalent electronic processors will most likely need to take advantage of not just one of the features of optics but also will need to carefully combine several of them. For example, just taking advantage of the large bandwidth of optics in a single spatial mode — even if we ignore for now input/output bottlenecks — is probably not sufficient to enable a throughput benefit as electronic processors compensate for lower bandwidth with enormous spatial parallelism (having on the order of 10^{11} transistors in modern chips). Similarly, relying only on spatial parallelism will likely also be insufficient: although the spatial parallelism of optics is considerable, especially in 3D systems, the spatial parallelism of transistors is typically even more impressive. (Optical multiplication of vectors by random matrices is an exception in which the spatial parallelism is so large that even very low bandwidth does not prevent the system from having higher throughput than electronic processors[73]. Even in this case though, more than one property of optics is being used: not only spatial parallelism but also nearly dissipationless dynamics).

However, if one can combine the bandwidth and spatial-parallelism features of optics in a single system, then there is potential to surpass electronics. For example, imagine being able to process data in 10^7 spatial modes in parallel at a clock rate of 10 THz, or processing data in parallel in 10^7 spatial modes, each with 10^7 frequency modes — in other words, 10^{14} parallel spatio-frequency modes. The numbers 10^7 and 10^7 are chosen somewhat arbitrarily but as believably practical, as, for example, we already have technology — spatial light modulators — for manipulating 10^7 spatial modes. We could have even higher numbers of spatial and frequency modes though — this is an example, not a bound. Although it is far from a solved problem how to fully take advantage of the combination of bandwidth and spatial parallelism afforded by optics, when combined with the fact that operations can be performed

nearly dissipationlessly in optics, there is great potential for optics to outperform electronics.

Accurately predicting the future of technology is difficult, but it seems reasonable to hypothesize that of the 11 features explored in this Perspective article, bandwidth, spatial parallelism and nearly dissipationless dynamics are most likely to have a key role in any future optical processor that does deliver an overall advantage in latency, throughput or energy efficiency. However, many of the other features may very well end up playing important roles too, so should not be ignored — but they will probably need to be combined with one of the 'big three' for a processor using them to achieve an overall advantage over electronics.

Many of the demonstrations of optical processors to date have shown a proof of principle of the use of some features of optics for computing in a way that could lead to an advantage, but with a system that does not suitably leverage some of the other available features, ultimately leading to a prototype that is inferior to current electronic processors. An example of this from my own group is ref. 38, which reports using spatial parallelism to realize >500,000 scalar multiplications per pass of light through a free-space optical processor, but the prototype is extremely limited in bandwidth owing to the speed limits of the input and output stages, leading to performance that is ultimately many orders of magnitude worse than an electronic processor. In that project, we were not expecting to beat an electronic processor but rather were aiming to demonstrate how few photons are needed for matrix–vector multiplication in optical neural networks; nevertheless, to advance this proof-of-principle system to be competitive with electronics would require markedly increasing the system bandwidth. Besides spatial parallelism, the optical processor presented in ref. 38 also used some other features of optics, such as nearly dissipationless dynamics — without which the ultra-low optical energy usage demonstrated would not have been possible — and optical fan-in.

Outlook

My opinion is that the most likely route to building an optical processor that delivers a large advantage over electronic processors in throughput or energy efficiency (or both) in the near term is by constructing a free-space optical matrix–vector multiplier that takes advantage of large spatial parallelism and nearly dissipationless dynamics[1]. With a vector dimension of $N \approx 10^4$ and a matrix size of $N \times N$, it seems promising that one can achieve an advantage provided that the system can be operated at a rate of one matrix–vector multiplication per nanosecond and the surrounding electronics for input and output operate with state-of-the-art energy efficiency[21,38].

Such a system will require careful optical and electronic engineering to realize — it is a major engineering undertaking whose difficulty should not be underplayed — but is all based on existing technology components that can in principle be appropriately scaled. I find this candidate architecture the most promising in the near term largely because it has been well studied and many of the necessary building blocks are fairly advanced. An optical matrix–vector multiplier whose inputs are optical, such as when it is used as a preprocessor for visual scenes[88], would have a lower bar to deliver an advantage over electronic solutions, so I expect that if an optical matrix–vector multiplier does outperform an electronic processor it will probably first be for an application involving optical inputs. However, I certainly do not want to give the impression that I think a free-space spatially multiplexed architecture is the only one worth pursuing. There are many other architectures[1,2] — including those based on photonic integrated circuits rather than free-space systems and those involving frequency multiplexing rather than, or in addition to, spatial multiplexing — that are appealing and very much worth pursuing.

When evaluating an optical-computing scheme, it can be helpful to determine what the cost of simulating the scheme with a digital-electronic processor would be. For example, wave physics can be simulated by digital-electronic processors, so when seeking an advantage for optics from wave phenomena, one needs to consider the cost of equivalent digital-electronic approaches, and depending on the wave phenomena being exploited, the digital approaches may be competitive or outright superior. As another example, least-power-dissipation principles can be used to realize Ising optimizers from networks of coupled optical oscillators[4], but simulating the equations of motion of the network on a digital-electronic computer can yield the same behaviour as a physical, optical implementation, so the intrinsic least-power-dissipation phenomenon does not automatically give rise to a computing benefit. Instead, one also needs to leverage other benefits of optics, such as parallelism and low dissipation.

We conclude by summarizing some of the major outstanding challenges that, if addressed, would move us substantially closer to realizing practically useful optical computers:

- **Optical-processor architecture design**. There is a major challenge to design optical-processor architectures that most effectively use the features of optics to gain an advantage. It is not obvious that the existing optical-processor architectures (using free space or integrated photonics) — some of which are decades old[13] — are optimal, and there is an opportunity to invent refined or completely new designs to meet this challenge.
- **Applications**. We need to find good applications to target with optical processors. As one of the major roadblocks to achieving advantage with optical computing are issues associated with input/output, we want to find valuable applications in which we can avoid or mitigate input/output bottlenecks and costs. For example, it has proven very difficult to build an optical matrix–vector multiplier at a scale (N) at which the input/output costs can be sufficiently amortized, even though an optical matrix–vector multiplier can perform $O(N^2)$ operations with input/output costs of just $O(N)$. Given that even matrix–vector multiplication, with its $O(N^2)$ complexity, does not have a high enough ratio of computation to input data, it would be helpful to find useful subroutines, algorithms or applications that have higher complexity than $O(N^2)$ for input and output data sizes ~ N. An additional direction is to find applications that could benefit from other aspects of optical computing besides potential performance advantages. For example, direct optical processing of visual scenes could give a privacy advantage: an electronic processor of images captured by a camera that stores the images in memory could be hacked, but an optical processor that directly processes what it 'sees' and never converts the full incoming images to electronic format could be far harder to maliciously copy images from.
- **Nonlinearity**. Nonlinearity is crucial in many computations, and a low-energy, fast, small-footprint, reliably manufacturable nonlinearity would be a useful building block. The nonlinearity need not necessarily be all-optical—optoelectronic nonlinearity can also be useful[130], although generally one can hope to benefit from higher bandwidths and possibly lower energy consumption in all-optical nonlinearities[55]. A fast, few-photon nonlinearity capable of attojoule switching has been demonstrated[131]; one important

direction is in scalably manufacturing the nonlinearities that have already been established.

- **Cascadability**. In many computations — for example, in deep neural networks — the input data is fed not through one function but a sequence of functions. An optical implementation of the computation then often involves passing an optical signal either through the same optical setup multiple times or through multiple different optical setups (or both). Doing so requires being able to cascade optical processes in time or space. We mention three challenges that can arise in cascading optical processing stages: the first of which is nearly universal in optical processors and the latter two of which are specific to optical-computing schemes using particular implementations of optical nonlinearity. These challenges are: attenuation of the optical signal owing to optical loss, effective attentuation of the optical signal owing to weakness in optical nonlinearity and nonlinear-optical processes generating output light that is at wavelengths incompatible with being input to the next optical stage (for example, directly cascading many second-harmonic-generation processes is infeasible, because the frequency of the optical signal is doubled at each stage, so after just a few stages one reaches wavelengths that are beyond the optical spectrum and are impractical to use). The attenuation owing to weak nonlinearity is an attenuation that is fundamental and unrelated to optical loss, that is, it would occur even if the optical system were lossless. The attenuation arises because the part of a signal coming out of a nonlinear stage that was not affected by the nonlinearity is discarded — either explicitly, or implicitly by not taking meaningful part in later stages of the computation — but because optical nonlinearity is generally weak[76], less than 100% of the light input to a nonlinear stage will generally be acted on nonlinearly. Designing suitably cascadable systems can be approached in multiple ways: for example, at the level of processor architecture, one may opt to insert gain into the system to compensate for the signal attenuation. Doing so leads to further architectural and system-design decisions about the type of gain (purely optical or optoelectronic, in which case the gain is essentially provided electronically by transistors, a common architectural choice in optical-neural-network prototypes[2] that often also serves the dual purpose of providing nonlinearity), and its required speed, preservation of information encoded in the optical spectrum and so on, as well as new engineering challenges in realizing suitable gain components. One may also approach cascadability challenges at the component or physical-implementation level, seeking to realize lower-loss optical systems, or materials with higher nonlinear coefficients.
- **3D design and manufacturing**. Spatial parallelism can be massively enhanced by using a third dimension, and if the dissipation is kept low, this provides a path to advantage over electronics. Separately, enabling long-range coupling between modes by using a third dimension (and advantages relating to how transmission works in optics) can also bring benefits[132,133]. The key question here is how to engineer and fabricate programmable, large-scale, possibly dense, 3D processors[62,132,134,135].
- **Energy costs for electronic and optoelectronic components**. The energy cost of optical processors is typically dominated by the energy costs of the electronic parts of the computer (for example, in an analysis of optical neural networks running large transformer models, the optical energy used accounts for <1% of the total energy cost[21]; see also refs. 75,136). Many optical-computing schemes could benefit from — and to deliver advantage, may even require — the availability of large arrays of high-speed, low-power and low-cost detectors, analog-to-digital converters, modulators and digital-to-analog converters. Increasing the energy efficiency of these components is an important challenge.
- **Scale**. Most optical-computing schemes rely on parallelism — be it from frequency or time multiplexing, or spatial multiplexing or a combination — for part of how they will achieve an advantage over electronics. However, throughput and energy-efficiency advantages typically only materialize when the system size (that is, the number of parallel operations) is very large[21]. (The situation for latency, as opposed to throughput or energy-efficiency advantages, is more subtle in that it is more application-dependent: if an application requires a certain amount of highly parallelizable computation (such as matrix–vector multiplication) to be performed in as little time as possible, so long as an optical processor is large enough to perform all that computation in parallel, it is big enough and won't necessarily benefit from larger scale (from the perspective of latency). A latency advantage could then arise from how the system is designed to minimize the time it takes to get the data into and out of the constituent parallel-processing units. But conversely, an optical processor could also deliver a latency advantage that is directly attributable to its scale: if it has parallelism far beyond that of an electronic processor it may achieve a throughput advantage that then will typically give a latency advantage as a side benefit for large tasks in which an electronic processor would need to perform the computation in multiple stages in series on account of the task being larger than the parallel-processing capacity of the electronic processor.) For example, we would like optical matrix–vector multipliers to be large enough to amortize the energy costs of loading the input vector and reading out the output vector. We would also like them to be large enough to be able to compete in throughput with electronic processors, which can perform $>10^6$ 8-bit-precision scalar multiplications per nanosecond[45] — so if vectors are input at a rate of 1 GHz, we would like the optical processor to also be able to perform $>10^6$ scalar multiplications in parallel. However, in optical matrix–vector multipliers made from arrays of Mach–Zehnder interferometers[1], even a state-of-the-art commercial prototype with a 64×64 array[137] does >100× fewer parallel operations than seems necessary to compete in throughput with state-of-the-art electronics solutions. A major challenge is how to scale arrays of size 64×64 to something much larger, like $1{,}000 \times 1{,}000$, which would put them roughly on par with the degree of parallelism in a single state-of-the-art electronic chip[45], or $10^4 \times 10^4$, which would then be in the regime in which a substantial throughput advantage could be achieved provided the system was clocked at a comparable rate to electronics (that is, at ~1 GHz). How can Mach–Zehnder-interferometer arrays be scaled from sizes ~ 64×64 to sizes ~ $10^4 \times 10^4$? This question is a major challenge for the community working on this approach. The challenge of scaling to achieve a far greater degree of parallelism than current prototypes is certainly not unique to optical matrix–vector multipliers or Mach–Zehnder-interferometer arrays — most optical-computing schemes face a major scaling challenge for them to be able to deliver a practical advantage. In some cases, we do not even have a solid practical roadmap for how to scale yet: for example, what is a feasible way to scale a scheme that combines spatial and frequency multiplexing (such as that in ref. 40, using 16 spatial and 4 frequency degrees of freedom) to a point where it can

achieve advantage? There is the potential for very large numbers of both spatial and frequency modes to be harnessed to perform parallel computations (for example, $>10^{14}$ spatio-frequency modes being operated on in parallel), but how can we reach this scale for a concrete scheme that performs useful computation?

- **Robustness, reliability and fabrication variation.** Although many optical components, such as those appearing in consumer-electronics devices such as cellphones and in optical-fibre-communications systems, are generally very reliable, there are many optical technologies that are being considered for use in optical computers that present challenges in robustness (for example, how well they can perform in the presence of environmental perturbations such as temperature changes or mechanical vibrations), reliability (for example, how likely they are to keep functioning correctly under normal operation conditions) and fabrication variation (for example, how much fabricated devices will differ in specifications from their designed values). For example, many optical phase-change-memory technologies have stringent limits on how many times they can be switched, and it is desirable for these limits to be raised[138,139]. As another example, in integrated photonics, Mach–Zehnder interferometers typically suffer from the constituent splitters having small deviations from the ideal splitting ratio owing to variations in fabrication; one research direction is to improve the fabrication processes, and another is to construct designs that can compensate for these fabrication errors[140]. Generally, for each photonic technology platform that might be used in an optical computer, there are open problems in how to stabilize it — passively or actively.
- **Storage**. To avoid the costs of converting between electronics and optics, and to avoid the cost of electronic memory accesses (which is a dominant cost even in electronic computing[23]), we would often like to be able to store data for use in optical processing. For example, in matrix–vector multipliers, we typically want to be able to store matrices with as low energy cost as possible for maintaining the storage, but in a way that the matrix can be updated on demand many times, at reasonably high accuracy (say, 8 bits), and also with relatively low energy cost[57,138]. In some applications or architectures, it is advantageous to be able to store optical signals (corresponding to intermediate calculation results, for example) so that conversion from optics to electronics and then back to optics can be avoided. There is active study and much room for improvement in both these use cases of storage.
- **Pushing towards quantum limits**. One path towards minimizing optical energy consumption is to operate optical computers in a regime in which the quantum nature of light cannot be ignored — for example, by using ultra-low optical powers in which signals comprise small numbers of photons and are measured by single-photon detectors. Note that optical computers will inevitably involve some electronics, if only for control or readout, and it is often the electronics energy costs that dominate[136], so it is only in some cases that there is strong benefit to minimizing the optical power used. Nevertheless, for these situations, there is much work to be done in both designing architectures and realizing practical devices that benefit from operating in the quantum regime[141–145].

Constructing an optical computer that beats an electronic computer in any metric is challenging, given how advanced electronic processors are. However, the physics of optical computing gives promise that if optical computers are carefully engineered, for certain classes of tasks — especially those involving data that are already in an optical format or that have a very high ratio of computation to data — they may deliver orders-of-magnitude benefits in latency, throughput or energy efficiency.

Published online: 9 October 2023

References

1. Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
2. Shastri, B. J. et al. Photonics for artificial intelligence and neuromorphic computing. *Nat. Photon.* **15**, 102–114 (2021).
3. Greengard, S. Photonic processors light the way. *Commun. ACM* **64**, 16–18 (2021).
4. Mohseni, N., McMahon, P. L. & Byrnes, T. Ising machines as hardware solvers of combinatorial optimization problems. *Nat. Rev. Phys.* **4**, 363–379 (2022).
5. Tucker, R. S. The role of optics in computing. *Nat. Photon.* **4**, 405 (2010).
6. Miller, D. A. Are optical transistors the logical next step? *Nat. Photon.* **4**, 3–5 (2010).
7. Datta, S., Chakraborty, W. & Radosavljevic, M. Toward attojoule switching energy in logic transistors. *Science* **378**, 733–740 (2022).
8. Artificial intelligence and the rise of optical computing. *The Economist* (20 December 2022).
9. Cartlidge, E. Photonic computing for sale. *Opt. Photon. News* **34**, 26–33 (2023).
10. Feinberg, B., Vengalam, U. K. R., Whitehair, N., Wang, S. & Ipek, E. Enabling scientific computing on memristive accelerators. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture* (*ISCA*), 367–382 (IEEE, 2018).
11. Dubrovsky, M., Ball, M., Kiffer, L. & Penkovsky, B. Towards optical proof of work. In *Cryptoeconomic Systems '20 Conference* (2020).
12. Pai, S. et al. Experimental evaluation of digitally verifiable photonic computing for blockchain and cryptocurrency. *Optica* **10**, 552–560 (2023).
13. Ambs, P. Optical computing: a 60-year adventure. *Adv. Opt. Technol.* **2010**, 372652 (2010).
14. Van der Sande, G., Brunner, D. & Soriano, M. C. Advances in photonic reservoir computing. *Nanophotonics* **6**, 561–576 (2017).
15. Nakajima, M., Tanaka, K. & Hashimoto, T. Scalable reservoir computing on coherent linear photonic processor. *Commun. Phys.* **4**, 1–12 (2021).
16. Teğin, U., Yıldırım, M., Oğuz, İ., Moser, C. & Psaltis, D. Scalable optical learning operator. *Nat. Comput. Sci.* **1**, 542–549 (2021).
17. Wright, L. G. et al. Deep physical neural networks trained with backpropagation. *Nature* **601**, 549–555 (2022).
18. Zhou, T., Scalzo, F. & Jalali, B. Nonlinear Schrödinger kernel for hardware acceleration of machine learning. *J. Lightwave Technol.* **40**, 1308–1319 (2022).
19. Semenova, N., Larger, L. & Brunner, D. Understanding and mitigating noise in trained deep neural networks. *Neural Netw.* **146**, 151–160 (2022).
20. Huang, C. et al. Prospects and applications of photonic neural networks. *Adv. Phys. X* **7**, 1981155 (2022).
21. Anderson, M. G., Ma, S.-Y., Wang, T., Wright, L. G. & McMahon, P. L. Optical transformers. Preprint at https://arxiv.org/abs/2302.10360 (2023).
22. Leiserson, C. E. et al. There's plenty of room at the top: what will drive computer performance after Moore's law? *Science* **368**, eaam9744 (2020).
23. Horowitz, M. Computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers* (*ISSCC*), 10–14 (IEEE, 2014).
24. Xu, X. et al. Scaling for edge inference of deep neural networks. *Nat. Electron.* **1**, 216–222 (2018).
25. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://arxiv.org/abs/2001.08361 (2020).
26. Zhai, X., Kolesnikov, A., Houlsby, N. & Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12104–12113 (IEEE, 2022).
27. Samsung unveils isocell image sensor with industry's smallest 0.56 µm pixel. https://news.samsung.com/global/samsung-unveils-isocell-image-sensor-with-industrys-smallest-0-56%CE%BCm-pixel (2022).
28. Fahrenkopf, N. M. et al. The AIM photonics MPW: a highly accessible cutting edge technology for rapid prototyping of photonic integrated circuits. *IEEE J. Sel. Top. Quantum Electron.* **25**, 1–6 (2019).
29. Chen, X., Li, C. & Tsang, H. K. Device engineering for silicon photonics. *NPG Asia Mater.* **3**, 34–40 (2011).
30. Borghi, M., Castellan, C., Signorini, S., Trenti, A. & Pavesi, L. Nonlinear silicon photonics. *J. Opt.* **19**, 093002 (2017).
31. Blumenthal, D. J., Heideman, R., Geuzebroek, D., Leinse, A. & Roeloffzen, C. Silicon nitride in silicon photonics. *Proc. IEEE* **106**, 2209–2231 (2018).
32. Gaeta, A. L., Lipson, M. & Kippenberg, T. J. Photonic-chip-based frequency combs. *Nat. Photon.* **13**, 158–169 (2019).
33. Chen, R. et al. Opportunities and challenges for large-scale phase-change material integrated electro-photonics. *ACS Photon.* **9**, 3181–3195 (2022).
34. Panuski, C. L. et al. A full degree-of-freedom spatiotemporal light modulator. *Nat. Photon.* **16**, 834–842 (2022).

35. Boes, A. et al. Lithium niobate photonics: unlocking the electromagnetic spectrum. *Science* **379**, eabj4396 (2023).
36. Chen, Y., Xie, Y., Song, L., Chen, F. & Tang, T. A survey of accelerator architectures for deep neural networks. *Engineering* **6**, 264–274 (2020).
37. Yu, S., Jiang, H., Huang, S., Peng, X. & Lu, A. Compute-in-memory chips for deep learning: recent trends and prospects. *IEEE Circuits Syst. Mag.* **21**, 31–56 (2021).
38. Wang, T. et al. An optical neural network using less than 1 photon per multiplication. *Nat. Commun.* **13**, 123 (2022).
39. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11**, 441–446 (2017).
40. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
41. Ohno, S., Tang, R., Toprasertpong, K., Takagi, S. & Takenaka, M. Si microring resonator crossbar array for on-chip inference and training of the optical neural network. *ACS Photon.* **9**, 2614–2622 (2022).
42. Hooker, S. The hardware lottery. *Commun. ACM* **64**, 58–65 (2021).
43. Rudolph, T. Why I am optimistic about the silicon-photonic route to quantum computing. *APL Photon.* **2**, 030901 (2017).
44. Hoefler, T., Häner, T. & Troyer, M. Disentangling hype from practicality: on realistically achieving quantum advantage. *Commun. ACM* **66**, 82–87 (2023).
45. NVIDIA Hopper architecture in-depth. https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/ (2022).
46. Deal, W., Leong, K., Yoshida, W., Zamora, A. & Mei, X. InP HEMT integrated circuits operating above 1,000 GHz. In *2016 IEEE International Electron Devices Meeting (IEDM)*, 29-1 (IEEE, 2016).
47. Thome, F. & Leuther, A. First demonstration of distributed amplifier MMICs with more than 300-GHz bandwidth. *IEEE J. Solid-State Circuits* **56**, 2647–2655 (2021).
48. Ho, R., Mai, K. W. & Horowitz, M. A. The future of wires. *Proc. IEEE* **89**, 490–504 (2001).
49. Rabaey, J. M., Chandrakasan, A. & Nikolic, B. *Digital Integrated Circuits: A Design Perspective* 2nd edn (Pearson, 2002).
50. Diddams, S. A., Vahala, K. & Udem, T. Optical frequency combs: coherently uniting the electromagnetic spectrum. *Science* **369**, eaay3676 (2020).
51. Marin-Palomo, P. et al. Microresonator-based solitons for massively parallel coherent optical communications. *Nature* **546**, 274–279 (2017).
52. Hennessy, J. L. & Patterson, D. A. Computer architecture: a quantitative approach. in *The Morgan Kaufmann Series in Computer Architecture and Design* 6th edn (Morgan Kaufmann, 2017).
53. Sicard, E. & Trojman, L. Introducing 5-nm FinFET technology in Microwind. https://hal.archives-ouvertes.fr/hal-03254444 (2021).
54. Xie, Q. et al. Performance comparisons between 7-nm FinFET and conventional bulk CMOS standard cell libraries. *IEEE Trans. Circuits Syst. II Express Briefs* **62**, 761–765 (2015).
55. Guo, Q. et al. Femtojoule femtosecond all-optical switching in lithium niobate nanophotonics. *Nat. Photon.* **16**, 625–631 (2022).
56. Kahn, J. M. & Miller, D. A. Communications expands its space. *Nat. Photon.* **11**, 5–8 (2017).
57. Bogaerts, W. et al. Programmable photonic circuits. *Nature* **586**, 207–216 (2020).
58. Khaddam-Aljameh, R. et al. HERMES-core — a 1.59-tops/mm^2 PCM on 14-nm CMOS in-memory compute core using 300-ps/LSB linearized CCO-based ADCs. *IEEE J. Solid-State Circuits* **57**, 1027–1038 (2022).
59. Vahala, K. J. Optical microcavities. *Nature* **424**, 839–846 (2003).
60. Majumdar, A. et al. Design and analysis of photonic crystal coupled cavity arrays for quantum simulation. *Phys. Rev. B* **86**, 195312 (2012).
61. Da Dalt, N., Knopf, C., Burian, M., Hartig, T. & Eul, H. A 10b 10GHz digitally controlled LC oscillator in 65nm CMOS. In *2006 IEEE International Solid State Circuits Conference-Digest of Technical Papers*, 669–678 (IEEE, 2006).
62. Psaltis, D., Brady, D., Gu, X.-G. & Lin, S. Holography in artificial neural networks. *Nature* **343**, 325–330 (1990).
63. Sell, B. et al. Intel 4 CMOS technology featuring advanced FinFET transistors optimized for high density and high-performance computing. In 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), 282–283 (IEEE, 2022).
64. Li, H.-Y. S., Qiao, Y. & Psaltis, D. Optical network for real-time face recognition. *Appl. Opt.* **32**, 5026–5035 (1993).
65. Goda, A. 3-D NAND technology achievements and future scaling perspectives. *IEEE Trans. Electron Dev.* **67**, 1373–1381 (2020).
66. Dally, W. J. The future of high-performance computing: are neuromorphic systems the answer? https://www.youtube.com/watch?v=lH3wKXZK9Zc (2022).
67. Neff, J. A., Athale, R. A. & Lee, S. H. Two-dimensional spatial light modulators: a tutorial. *Proc. IEEE* **78**, 826–855 (1990).
68. Zhou, T. et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nat. Photon.* **15**, 367–373 (2021).
69. Mait, J. N., Euliss, G. W. & Athale, R. A. Computational imaging. *Adv. Opt. Photon.* **10**, 409–483 (2018).
70. Colburn, S., Chu, Y., Shilzerman, E. & Majumdar, A. Optical frontend for a convolutional neural network. *Appl. Opt.* **58**, 3179–3186 (2019).
71. Desiatov, B., Shams-Ansari, A., Zhang, M., Wang, C. & Lončar, M. Ultra-low-loss integrated visible photonics using thin-film lithium niobate. *Optica* **6**, 380–384 (2019).
72. Goodman, J. *Introduction to Fourier Optics* (Roberts and Company Publishers, 2004).
73. Saade, A. et al. Random projections through multiple optical scattering: approximating kernels at the speed of light. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6215–6219 (IEEE, 2016).
74. Lent, C. S., Orlov, A. O., Porod, W. & Snider, G. L. (eds) *Energy Limits in Computation* 1st edn (Springer International Publishing, 2018).
75. Hamerly, R., Bernstein, L., Sludds, A., Soljačić, M. & Englund, D. Large-scale optical neural networks based on photoelectric multiplication. *Phys. Rev. X* **9**, 021032 (2019).
76. Boyd, R. W. *Nonlinear Optics* (Academic Press, 2020).
77. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
78. Kia, B., Lindner, J. F. & Ditto, W. L. Nonlinear dynamics as an engine of computation. *Philos. Trans. Royal Soc. A Math. Phys. Eng. Sci.* **375**, 20160222 (2017).
79. Miller, D. A. Attojoule optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.* **35**, 346–396 (2017).
80. Jose, A. P., Patounakis, G. & Shepard, K. L. Pulsed current-mode signaling for nearly speed-of-light intrachip communication. *IEEE J. Solid-State Circuits* **41**, 772–780 (2006).
81. Tyndall, N. F. et al. A low-loss, broadband, nitride-only photonic integrated circuit platform. In *Quantum 2.0*, QTu4B-5 (Optica Publishing Group, 2022).
82. Cheng, Q., Glick, M. & Bergman, K. Optical interconnection networks for high-performance systems. In *Optical Fiber Telecommunications* VII, 785–825 (Elsevier, 2020).
83. Sun, C. et al. Single-chip microprocessor that communicates directly using light. *Nature* **528**, 534–538 (2015).
84. Fillion-Gourdeau, F. & Gagnon, J.-S. On the physical (im)possibility of lightsabers. *Eur. J. Phys.* **40**, 055201 (2019).
85. Duan, C., LaMeres, B. J. & Khatri, S. P. *On and Off-Chip Crosstalk Avoidance in VLSI Design* (Springer, 2010).
86. Lee, J. N. *Design Issues in Optical Processing* (Cambridge Univ. Press, 1995).
87. Nassif, N. et al. Sapphire rapids: the next-generation intel xeon scalable processor. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, Vol. 65, 44–46 (IEEE, 2022).
88. Wang, T. et al. Image sensing with multilayer nonlinear optical neural networks. *Nat. Photon.* **17**, 408–415 (2023).
89. Goodman, J. W. Fan-in and fan-out with optical interconnections. *Optica Acta Int. J. Opt.* **32**, 1489–1496 (1985).
90. McArdle, N., Naruse, M., Toyoda, H., Kobayashi, Y. & Ishikawa, M. Reconfigurable optical interconnections for parallel computing. *Proc. IEEE* **88**, 829–837 (2000).
91. Wang, T. & Arrathoon, R. Limits of optical and electrical fan-out versus power and fan-out versus bandwidth. In *Hybrid Image and Signal Processing II* Vol. 1297, 133–149 (SPIE, 1990).
92. Ji, L. & Heuring, V. P. Impact of gate fan-in and fan-out limits on optoelectronic digital circuits. *Appl. Opt.* **36**, 3927–3940 (1997).
93. Chen, J., Clark, L. & Cao, Y. Maximum fan-in/out: ultra-low voltage circuit design in the presence of variations. *IEEE Circ. Dev. Mag.* **21**, 12–20 (2006).
94. de Groot, P. J. & Noll, R. J. Adaptive neural network in a hybrid optical/electronic architecture using lateral inhibition. *Appl. Opt.* **28**, 3852–3859 (1989).
95. Bernstein, L. et al. Single-shot optical neural network. *Sci. Adv.* **9**, eadg7904 (2023).
96. Yao, R. et al. Compact and low-insertion-loss 1×N power splitter in silicon photonics. *J. Lightwave Technol.* **39**, 6253–6259 (2021).
97. Xu, X. et al. 11 Tops photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
98. Sludds, A. et al. Delocalized photonic deep learning on the internet's edge. *Science* **378**, 270–276 (2022).
99. Murmann, B., Bankman, D., Chai, E., Miyashita, D. & Yang, L. Mixed-signal circuits for embedded machine-learning applications. In *2015 49th Asilomar Conference on Signals, Systems and Computers*, 1341–1345 (IEEE, 2015).
100. DeBenedictis, E. P. Computational complexity and new computing approaches. *Computer* **49**, 76–79 (2016).
101. Shi, L., Zheng, G., Tian, B., Dkhil, B. & Duan, C. Research progress on solutions to the sneak path issue in memristor crossbar arrays. *Nanoscale Adv.* **2**, 1811–1827 (2020).
102. Aluf, O. *Optoisolation Circuits: Nonlinear Applications in Engineering* (World Scientific, 2012).
103. Vadlamani, S. K., Xiao, T. P. & Yablonovitch, E. Physics successfully implements Lagrange multiplier optimization. *Proc. Natl Acad. Sci. USA* **117**, 26639–26650 (2020).
104. E, W. & Yu, B. The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. *Commun. Math. Stat.* **6**, 1–12 (2018).
105. Feynman, R. P. *QED: The Strange Theory of Light and Matter* (Princeton Univ. Press, 2006).
106. Wen, K. *Injection-locked Laser Network for Solving NP-Complete Problems*. PhD thesis, Stanford Univ. (2012).
107. Marandi, A., Wang, Z., Takata, K., Byer, R. L. & Yamamoto, Y. Network of time-multiplexed optical parametric oscillators as a coherent Ising machine. *Nat. Photon.* **8**, 937–942 (2014).
108. Andersen, U. L., Gehring, T., Marquardt, C. & Leuchs, G. 30 years of squeezed light generation. *Phys. Scr.* **91**, 053001 (2016).
109. Knill, E., Laflamme, R. & Milburn, G. J. A scheme for efficient quantum computation with linear optics. *Nature* **409**, 46–52 (2001).
110. Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge Univ. Press, 2010).
111. Dowling, J. P. & Milburn, G. J. Quantum technology: the second quantum revolution. *Philos. Trans. Royal Soc. Lond. A Math. Phys. Eng. Sci.* **361**, 1655–1674 (2003).

112. Pearson, B. J. & Jackson, D. P. A hands-on introduction to single photons and quantum mechanics for undergraduates. *Am. J. Phys.* **78**, 471–484 (2010).
113. Carolan, J. et al. Universal linear optics. *Science* **349**, 711–716 (2015).
114. Duprez, H. et al. Macroscopic electron quantum coherence in a solid-state circuit. *Phys. Rev. X* **9**, 021030 (2019).
115. Lee, T. H. *The Design of CMOS Radio-frequency Integrated Circuits* (Cambridge Univ. Press, 2003).
116. Safavi-Naeini, A. H., Van Thourhout, D., Baets, R. & Van Laer, R. Controlling phonons and photons at the wavelength scale: integrated photonics meets integrated phononics: publisher's note. *Optica* **6**, 410 (2019).
117. Dwivedi, S. et al. Experimental extraction of effective refractive index and thermo-optic coefficients of silicon-on-insulator waveguides using interferometers. *J. Lightwave Technol.* **33**, 4471–4477 (2015).
118. Rabaey, J. M., Chandrakasan, A. & Nikolic, B. *Digital Integrated Circuits: A Design Perspective* 2nd edn, Ch. 4 (Pearson, 2002).
119. Bremermann, H. J. Quantum noise and information. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, 15–20 (1967).
120. Kao, Y.-C., Chen, H.-A. & Ma, H.-P. An FPGA-based high-frequency trading system for 10 gigabit ethernet with a latency of 433 ns. In *2022 International Symposium on VLSI Design, Automation and Test* (*VLSI-DAT*) 1–4 (IEEE, 2022).
121. Pope, R. et al. Efficiently scaling transformer inference. In *Proc. of Machine Learning and Systems* (MLSys, 2023).
122. Alexoudi, T., Kanellos, G. T. & Pleros, N. Optical RAM and integrated optical memories: a survey. *Light Sci. Appl.* **9**, 1–16 (2020).
123. Ashtiani, F., Geers, A. J. & Aflatouni, F. An on-chip photonic deep neural network for image classification. *Nature* **606**, 501–506 (2022).
124. Rodrigues, S. P. et al. Weighing in on photonic-based machine learning for automotive mobility. *Nat. Photon.* **15**, 66–67 (2021).
125. Chang, J., Sitzmann, V., Dun, X., Heidrich, W. & Wetzstein, G. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Sci. Rep.* **8**, 1–10 (2018).
126. Minzioni, P. et al. Roadmap on all-optical processing. *J. Opt.* **21**, 063001 (2019).
127. Huang, C. et al. A silicon photonic–electronic neural network for fibre nonlinearity compensation. *Nat. Electron.* **4**, 837–844 (2021).
128. Chen, Y. et al. Photonic unsupervised learning processor for secure and high-throughput optical fiber communication. Preprint at https://arxiv.org/abs/2203.03807 (2022).
129. Ghobadi, M. Emerging optical interconnects for AI systems. In *2022 Optical Fiber Communications Conference and Exhibition* (*OFC*) 1–3 (IEEE, 2022).
130. Williamson, I. A. et al. Reprogrammable electro-optic nonlinear activation functions for optical neural networks. *IEEE J. Sel. Top. Quantum Electron.* **26**, 1–12 (2019).
131. Zasedatelev, A. V. et al. Single-photon nonlinearity at room temperature. *Nature* **597**, 493–497 (2021).
132. Dinc, N. U., Psaltis, D. & Brunner, D. Optical neural networks: the 3D connection. *Photoniques* **104**, 34–38 (2020).
133. Boahen, K. Dendrocentric learning for synthetic intelligence. *Nature* **612**, 43–50 (2022).
134. Morris, R., Kodi, A. K. & Louri, A. Dynamic reconfiguration of 3d photonic networks-on-chip for maximizing performance and improving fault tolerance. In *2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*, 282–293 (IEEE, 2012).
135. Moughames, J. et al. Three-dimensional waveguide interconnects for scalable integration of photonic neural networks. *Optica* **7**, 640–646 (2020).
136. Tait, A. N. Quantifying power in silicon photonic neural networks. *Phys. Rev. Appl.* **17**, 054029 (2022).
137. Ramey, C. Silicon photonics for artificial intelligence acceleration: Hotchips 32. In *IEEE Hot Chips 32 Symposium*, 1–26 (IEEE, 2020).
138. Zhang, Y. et al. Myths and truths about optical phase change materials: a perspective. *Appl. Phys. Lett.* **118**, 210501 (2021).
139. Martin-Monier, L. et al. Endurance of chalcogenide optical phase change materials: a review. *Opt. Mater. Express* **12**, 2145–2167 (2022).
140. Hamerly, R., Bandyopadhyay, S. & Englund, D. Asymptotically fault-tolerant programmable photonics. *Nat. Commun.* **13**, 6831 (2022).
141. Mabuchi, H. Nonlinear interferometry approach to photonic sequential logic. *Appl. Phys. Lett.* **99**, 153103 (2011).
142. Kerckhoff, J., Armen, M. A. & Mabuchi, H. Remnants of semiclassical bistability in the few-photon regime of cavity QED. *Opt. Expr.* **19**, 24468–24482 (2011).
143. Tezak, N. & Mabuchi, H. A coherent perceptron for all-optical learning. *EPJ Quant. Technol.* **2**, 1–22 (2015).
144. Shainline, J. M., Buckley, S. M., Mirin, R. P. & Nam, S. W. Superconducting optoelectronic circuits for neuromorphic computing. *Phys. Rev. Appl.* **7**, 034013 (2017).
145. Ma, S.-Y., Wang, T., Laydevant, J., Wright, L. G. & McMahon, P. L. Quantum-noise-limited optical neural networks operating at a few quanta per activation. Preprint at https://arxiv.org/abs/2307.15712 (2023).
146. Johnson, A. R. et al. Octave-spanning coherent supercontinuum generation in a silicon nitride waveguide. *Opt. Lett.* **40**, 5117–5120 (2015).
147. Zhang, X., Kwon, K., Henriksson, J., Luo, J. & Wu, M. C. A large-scale microelectromechanical-systems-based silicon photonics lidar. *Nature* **603**, 253–258 (2022).
148. Stallings, W. *Data and Computer Communications* 8th edn (Pearson, 2007).
149. Kleveland, B. et al. High-frequency characterization of on-chip digital interconnects. *IEEE J. Solid-State Circuits* **37**, 716–725 (2002).
150. Qaxial. *RG142B/U Flexible PTFE High Power Coaxial Cable Datasheet* (2022).
151. Bauters, J. F. et al. Planar waveguides with less than 0.1 dB/m propagation loss fabricated with wafer bonding. *Opt. Expr.* **19**, 24090–24101 (2011).
152. Schubert, E. F. *Light-Emitting Diodes* 2nd edn (Cambridge Univ. Press, 2012).
153. Corning. *SMF-28 ULL Optical Fiber Portfolio Product Information* (2021).
154. Miller, D. A. B. & Ozaktas, H. M. Limit to the bit-rate capacity of electrical interconnects from the aspect ratio of the system architecture. *J. Parallel Distrib. Comput.* **41**, 42–52 (1997).
155. Huang, D., Sze, T., Landin, A., Lytel, R. & Davidson, H. L. Optical interconnects: out of the box forever? *IEEE J. Sel. Top. Quantum Electron.* **9**, 614–623 (2003).
156. Shams-Ansari, A. et al. Reduced material loss in thin-film lithium niobate waveguides. *APL Photon.* **7**, 081301 (2022).
157. Johnson, M., Thompson, M. G. & Sahin, D. Low-loss, low-crosstalk waveguide crossing for scalable integrated silicon photonics applications. *Opt. Expr.* **28**, 12498–12507 (2020).
158. Stepanovsky, M. A comparative review of MEMS-based optical cross-connects for all-optical networks from the past to the present day. *IEEE Commun. Surv. Tutor.* **21**, 2928–2946 (2019).
159. Barredo, D., Lienhard, V., De Leseleuc, S., Lahaye, T. & Browaeys, A. Synthetic three-dimensional atomic structures assembled atom by atom. *Nature* **561**, 79–82 (2018).
160. Wayne, M. et al. A 500 × 500 dual-gate SPAD imager with 100% temporal aperture and 1 ns minimum gate length for film and phasor imaging applications. *IEEE Trans. Electron Devices* **69**, 2865–2872 (2022).

Acknowledgements

The author gratefully acknowledges many helpful conversations with co-workers including D. Brunner, R. Hamerly, H. Mabuchi, A. Majumdar, A. Marandi, E. Ng, T. Onodera, T. Wang, L. Wright and Y. Yamamoto; these conversations over several years have shaped his understanding of optical computing. The author also gratefully acknowledges S. Agarwal for explanations about analog-electronic crossbars and B. Govind for discussions about electrical interconnects. The author thanks M. Anderson, T. Wang and F. Wu for providing detailed feedback on a draft of this manuscript. This work has been financially supported in part by the National Science Foundation (Award CCF-1918549), NTT Research and a David and Lucile Packard Foundation Fellowship.

Competing interests

The author is listed as an inventor on several U.S. provisional patent applications relating to optical computing (63/149,974; 63/178,318; 63/392,042).

Additional information

Peer review information *Nature Reviews Physics* thanks the anonymous reviewers for their contribution to the peer review of this work.

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 2023