Integrated Photonic Neuromorphic Computing: Opportunities and Challenges

Nikolaos Farmakidis[1], *Bowei Dong*[1] *and Harish Bhaskaran**[1]

[1]Department of Materials, University of Oxford, Parks Road, Oxford OX1 3PH, UK.
*Corresponding authors: E-mail: *harish.bhaskaran@materials.ox.ac.uk*

Keywords: Photonic integrated circuits, neuromorphic computing, hardware accelerators

Abstract

Employing photons in lieu of electrons to process information has been an exciting technological prospect for decades. Optical computing is gaining renewed enthusiasm, owing to the accumulated maturity of photonic integrated circuits and the pressing need for faster processing to cope with data generated by artificial intelligence. In neuromorphic photonics, the bosonic nature of light is exploited for high-speed, densely multiplexed linear operations while the superior computing modalities of biological neurons are imitated to accelerate computations. Here we provide an overview of recent advances in integrated synaptic optical devices and on-chip photonic neural networks focusing on the point in the architecture where the optical to electrical conversion takes place. We identify challenges associated with electro-optical conversions, implementations of optical nonlinearity, amplification and processing in the time domain and we identify promising paths forward.

Introduction

Computing has reached impressive performance milestones. Achieving these has until recently, seen a synergy between the software and hardware communities, with each building on advances in both. However, this synergy has now been broken. Data generated by artificial intelligence (AI) is heavily reliant on deep neural networks (DNNs) for processing [1, 2], a shift which with few exceptions [3] has not been effectively mapped in the hardware [4] This apparent disjoint is evident both in the processor speed and the energy requirements per operation, an issue which is compounded for edge computing where the power- and bandwidth-hungry GPUs are not sustainably viable [5]. Although the deficiencies of traditional computing architectures in comparison to biological computing schemes have been identified since the 1980’s by Carver Mead et al., [6] the coincidental slowing of Moore’s law [7, 8] and Koomey’s law [9, 10] coupled with the exponential acceleration of data generated by AI implementations are now imposing an immediate and critical need for radical innovations in hardware [4, 5, 11].

Traditional hardware limitations have led to a growing interest in neuro-inspired implementations, attracting both academic and commercial interest [12]. Recent developments in electronic neuromorphic machines have demonstrated energy, latency, and processing speed advantages of neuro-imitative paradigms. These endeavours have replicated aspects of brain's architecture, specifically the vast interconnectivity, organizational hierarchy, and time-dependent synaptic processing. Implementations of these, including IBM’s TrueNorth, Intel’s Loihi, the SpiNNaker project and others, are showcasing the superiority of neuromorphic approaches [12-14]. Photonic neuromorphic computing, although still at its infancy, is following suit promising two to three orders of magnitude improvement in energy consumption and compute density [15, 16]. Operations in the optical domain can be high-speed, near-lossless and massively parallelised. Importantly, in the past decade the development of integrated light sources [17], high-speed optical modulators [18, 19], detectors [20], amplifiers [21], photonic memories and computing elements [22-25], (de)multiplexers [26] as well as the transition to foundry-processed PICs have collectively established the maturity needed to “go optical” [27].

Neuromorphic Photonics

Photonics is particularly well suited to neuromorphic hardware where performance is reliant on interconnectivity. Simply represented, the brain is comprised of 10^{11} cascaded neurons and 10^{15} synapses [28], forming a connected communication network which interacts via spikes. Each neuron carries out a weighted linear accumulation of neural inputs with a non-linear activation function subsequently applied (fig. 1a). Mapping these characteristics into optical hardware, can have analogous gains to the transition from electronic to optical communications where operating in the optical domain can offer:

i. Vast interconnectivity which can be effectively addressed through multiplexing
ii. Energy efficiency which is not compromised by capacitive charging/discharging
iii. Temporal processing which is not hindered by capacitive delays
iv. Sub-ns latency between long-range communication links

While linear multiply-accumulate (MAC) operations are readily available in photonics, fast and efficient all-optical nonlinearities remain challenging. Neuromorphic photonic chips nominally resort to optoelectronic conversions which impose an overhead in terms of latency, energy per operation and chip area. In this review, we categorise photonic neuromorphic implementations based on the depth at which the transition from the optical to the electronic domain takes place. Different approaches broadly include:

(i) ***incoherent electro-optical processors***, ***(ii) coherent electro-optical processors*** and ***(iii) all-optical neural networks including photonic spiking neural networks***

The key difference between incoherent and coherent optical accelerators being that in the former, the summation operation is materialising in the electronic domain whereas in the latter the optical signals are coherently added in the optical domain and only then are they converted into electrical signals at the photodetector. Wavelength multiplexing is rather straightforward for incoherent approaches because optical intensities are simply added in power, but cumbersome for coherent approaches due to the difficulty in simultaneously and accurately controlling the optical phases of different wavelengths due to dispersion. Yet, single-source coherent optical accelerators show a clear advantage in terms of cascadability. Because the full set of optical controls (amplitude, phase, , mode) is preserved upon addition. Incoherent optical accelerators, however, operate and preserve only the intensity information after weighting and addition. This makes wavelength multiplexing readily exploitable for incoherent approaches while the cascadability is

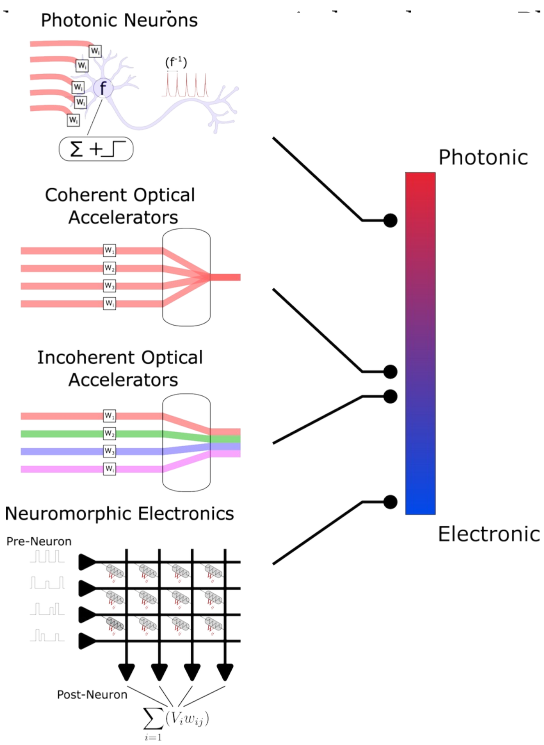


Figure 1: Neuromorphic computing schemes based on the depth of electro-optical conversions. **A)** Electronic neuromorphic crossbar array **B)** Incoherent optical accelerators, where incoherent signals are combined in a bus channel and summed at the photodetector, Since the photodetector is performing the summation, the electro-optical conversion occurs at an early stage. **C)** Coherent optical accelerators, wherein optical signals are added in the optical domain and subsequently converted to photoelectrons. **D)** The long-term prospect of photonic neurons combining weighting, summation and optical nonlinearity.

Incoherent Optical Accelerators

Inspired by electronic crossbar array architectures, incoherent optical accelerators directly map a biological many-neuron-to-many-neuron weighted connection, mathematically represented by a matrix, onto the hardware. Photonic weight banks, crossbar arrays and multimode interferometer architectures are only some of the approaches demonstrated. In most cases these operate on the amplitude of light with wavelength [36, 37] or mode [38] multiplexed optical signals carrying out the parallel matrix-vector-multiplications (MVMs). Figure 2a illustrates a basic dot product engine where signals separated in wavelength are weighted and multiplexed into a common output bus. The optical power at the output (P_o) gives the dot product of the input vector (*I*) and the weight vector (*w*) as $P_o = \sum (I_i \cdot w_i)$. The value of the dot product however is then only accessible when all wavelengths are indiscriminately converted to a photocurrent by the detector, which simultaneously performs the incoherent summation and optoelectronic conversion [39] [36]. Cascading serially what is essentially a weighted multiplexer/router, the detector and a nonlinear function in the electronic domain forms a primitive neuron as shown in figure 2a.

Weight banks

Initial demonstrations of incoherent optical accelerators were implemented on microring resonator (MRR) weight banks functioning as tuneable filters [37, 40, 41]. Individually addressable MRR resonators with integrated phase shifters, each control the amount of light coupled at a given wavelength, simultaneously performing the weighting and MUX-DEMUX operation (figure 2e) [42, 43]. In this so-called broadcast and weight approach, although information is encoded in the amplitude, there are no significant optical losses in coupling optical power to the bus waveguide. Bai et al. recently demonstrated on-chip integration of MRR weight banks with a microcomb source (figure 2g) to achieve 9-bit precision in weight setting and a compute density of more than a trillion operations pes second per square millimetre (TOPS/mm^2) [44]. A challenge of high-Q MRR-based systems however is the requirement for tracking and controlling their resonance as they are sensitive to their environment and fabrication [45]. This places a fixed, continuous energy toll to correct for fabrication and a fluctuating requirement to account for transient drift due to thermal coupling. Several approaches aiming to resolve these issues have been proposed based on ion implantation, localised material deposition/removal, transparent phase-change materials and polymers. [45-47].

Photonic crossbar arrays

A significantly less fabrication- and environment- sensitive approach involves the use of broadband couplers and tuneable optical attenuators to map the weight matrix (W) directly onto a photonic hardware topographically similar to an electronic crossbar array. Based on initial demonstrations of photonic in-memory computing by Rios et al. using phase change materials [48], Feldmann et al. subsequently demonstrated MVMs [36] at 0.256 TOPS (figure 2b). The crossbar array was interfaced with a frequency comb to perform parallel convolution processing, multiplexed in wavelength. Although broadband couplers enable multiplexed parallel processing, the power scaling of the combiners limits the size of the array. Resonant structures can be used to trade-off some WDM capacity with higher scalability (fig. 2D) [49-51]. The crossbar array was later improved to reduce channel crosstalk, to include WDM components on-chip and achieve balanced photodetection and electro-optical addressability of individual phase-change cells [52, 53]. Recently, Dong et al., demonstrated an additional degree of parallelisation with time domain multiplexing (figure 2c). Here, 50 RF – although more are also possible – signals were passed simultaneously through the structure and demultiplexed in Fourier space at the output. Due to the relatively wavelength-insensitive nature of the crossbar design, the degree of parallelisation in wavelength is practically only limited by the light source and not the crossbar itself. However, the scalability has been limited largely by the coupling loss at each column [54]. Initial solutions to expand the achievable size of the crossbar by tiling, expanding the crossbar array in three dimensions and employing factorisation techniques are actively explored [55, 56] while recent works have focused on in-situ training and updating of the weights [57, 58]. Importantly, by encoding weights into the amplitude of optical signals using nonresonant optical devices, the crossbar array architecture is insensitive to phase which not only enhances the system robustness but also permits the incorporation of high-speed foundry compatible components such as electro-absorption modulators (EAMs) [59].

Multimode interferometer networks

An alternative approach, based on broadband multimode interferometers (MMI) was proposed by Qu et al. [60] based on scattering and was demonstrated by Meng et al. on an integrated platform (figure 2f) [15]. Two 4 input, 4 output MMIs were employed to multiplex and distribute four wavelengths while thermo-optic phase shifters, one for each of the 4 channels, were used to reconfigure the transfer matrix of the system. While this approach benefits from a linear scaling relationship between the number of elements and the dimension of the matrix, the theoretical feasibility of scaling up this approach towards larger neural network remains to be investigated.

Conclusions on incoherent optical accelerators

Considering the simplest case of N-to-1 mapping, which in a biological context is analogous to the summation of weighted signals from N dendrites in a single neuron, all incoherent approaches encode these N individual weights into N orthogonal optical properties. This orthogonality can originate from various dimensions, such as the wavelength dimension through wavelength division multiplexing, the spatial dimension via mode division multiplexing, or the temporal dimension through time division multiplexed optical signals [61]. The necessitation of orthogonality is pivotal to prevent interference (or crosstalk caused by heating), which is enabling the linear addition of intensity from the N inputs at the photodetector. It is worth noting that the N weighted signals at the optical output are only multiplexed, not summed. They are still parallel in the optical domain. The summation only happens upon photodetection, where the intensities are finally summed up to provide one single physical quantity. This intensity-only operation considerably facilitates system control, owing to its robustness against random phase fluctuations or thermal crosstalk. However, this incoherent scheme introduces a notable challenge in terms of scalability. Given that N orthogonal properties are leveraged for a single N-to-1 mapping, the value of N is inherently constrained by the achievable orthogonality within photonics. A rudimentary estimation would yield 100 wavelengths, 4 modes, and 2 polarizations, generating a maximum conceivable N value of 800. In this scenario, achieving further optical parallelization becomes challenging as all orthogonal properties are exhausted. Concurrently, the prospect of all-optical cascadability presents an additional hurdle, given that the required number of orthogonal properties scales with N^L, where L represents the number of layers. A plausible resolution to enhance cascadability involves the incorporation of O-E-O (Optical-Electrical-Optical) conversion, which serves to replenish the orthogonal properties following each layer. However, this remedy requires an additional energy overhead and augments the system complexity due to the recurrent integration of electronic components.

It therefore becomes imperative to explore avenues to augment the number of accessible orthogonal optical properties, such as expanding the wavelength range by harnessing a broader bandwidth or exploring additional modes by delving into the orbital angular momentum of light [62]. A more innovative and promising approach would entail the utilization of a singular orthogonal optical property to execute intensity-only N-to-1 mapping.

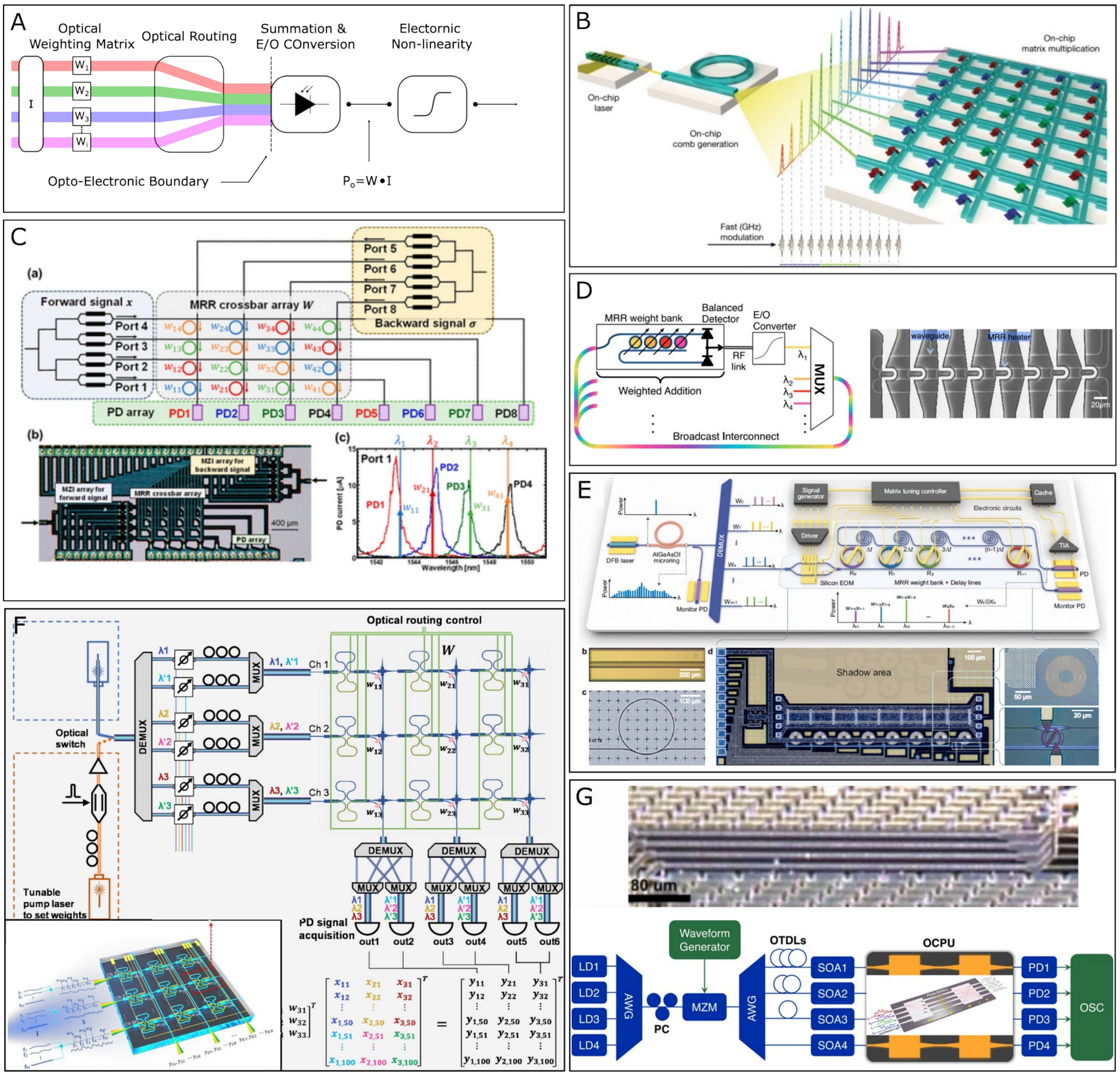


Figure 2: Photonic accelerators based on incoherent accumulation **A)** Schematic of WDM and weighting. Individually weighted signals are multiplexed into a bus waveguide and detected indiscriminately by the photodiode. **B)** Photonic crossbar array with wavelength division multiplexing and microcomb input [36]. **C)** Microring resonator crossbar array for on-chip inference and training of the optical neural network [51] **D)** Microring wight banks [37] **E)** Microcomb-based integrated photonic processing unit [44]. **F)** Crossbar array architecture with RF time division multiplexing [63]. **G)** Accelerator based on multimode interferometers [15].

Coherent Optical Accelerators

Coherent optical accelerators fundamentally use a single source which is distributed, routed and weighted by coherent optical interference throughout the network. These approaches are largely based on the concept of an arbitrary linear optical device described in the works of Reck et al. and Miller [64, 65], stating and proving that any NxN unitary matrix can be decomposed and represented in hardware via networking 2 by 2 reprogrammable cells [66, 67]. The 2 by 2 building blocks rely on interference to achieve independent amplitude and phase modulation at the output of each of the blocks; the values of each programmable unit can be self-configured using transparent detectors according to Miller's scheme [64, 68]. We consider these approaches "more optical" since linear MVM operations are performed all-optically at a single carrier frequency without the need for multiple wavelength or optical-electrical conversion making the system "more neuromorphic" and easier to tile and cascade [69, 70]. The enhanced interconnectivity of these approaches is well suited to both feedforward and recirculating architectures [71].

Feedforward meshes for inferencing

Photonic meshes are largely based on the works of Reck [65], Miller [64] and later on Clements [67, 69] with the realisation that any unitary linear transformation between a set of optical channels could be achieved with an architecture consisting of such building blocks through singular value decomposition [31]. The first experimental demonstration of an integrated neural network for deep learning based on these concept was implemented in 2017 by Shen et al (figure 3e)[72] with many following suit [69, 73, 74]. In their approach 56 MZIs were used to perform the equivalent of a 4 x 4 matrix multiplication through single value decomposition. Training was performed off-chip with standard backpropagation and gradient descent using a nonlinear function modelled after a saturable absorber. On-chip thermo-optic phase shifters were used to program the matrix which were later replaced by micromechanical shifters for 100 times faster reconfiguration [75] while implementations using electro-absorption modulators were shown to reach 50GHz reconfiguration [76]. A key advantage of photonic meshes is the ability to dynamically reconfigure the system to perform multiple functions akin to an optical FPGA, where multichannel optical switching, optical multiple-input-multiple-output descrambling, and filtering have been demonstrated [69, 74, 77, 78]. While several approaches have been explored to multiplex in time and wavelength [79], this remains a considerable limitation due to the spectral sensitivity of the interferometers [74, 80]. The unfavourable scaling in the number of reprogrammable units with the size of the matrix has since been tackled by several works which use pruning as a means to reduce the size [31, 81]

The in-situ training of an artificial neural network is important to achieve accurate inferencing. While initial demonstrations of photonic meshes performed off-chip training with traditional hardware, photonic analogues of the backpropagation algorithm were short after conceptualised for in situ training [82-84]. Recently, Pai et. al, demonstrated a three-layer, four-port silicon photonic neural network with programmable phase shifters and optical power monitoring to solve classification tasks using on-chip, in-situ backpropagation (figure 3c)[85].

Diffractive neural networks

On-chip diffractive neural networks have emerged as a natural extension to the work conducted by the free-space photonics community primarily based on spatial light modulators [86-88]. Integrated metasystems based on on-chip dielectric metasurfaces have demonstrated to perform multiplexed matrix-vector multiplications and image classifications offering a transition between freespace and integrated modalities (fig. 3d) [89, 90]. Vertical cavity surface emitting lasers (VCSELs) were also recently demonstrated to form a semi-integrated, highly scalable solution by vertical stacking to form an energy efficient deep neural network

[91]. A fully integrated diffractive neural network was demonstrated by Zhu et al. in 2022 which used two diffractive cells and *N* MZIs to implement parallel Fourier transforms and convolutions (figure 3f)[92]. This system benefits form linear scaling with the input data dimension to achieve a ~10-fold reduction in footprint [92, 93].

Conclusions on coherent optical accelerators

Optical accelerators based on coherent computing schemes have been the most popular in the literature for several reasons. Reconfigurable units of optical phase shifters can be easily fabricated on university fabrication facilities and scaled up at foundries. Although limited by the slow thermal time constants, thermo-optic effects have been popular owing to their reliability, achievable precision (>10bit) and low loss while there is a clear path to overcome these using MEMS or nonlinear shifters [94-97]. Recent works have additionally explored coherent crossbar designs based on arbitrary modulator technologies [98, 99]. Some approaches have also investigated lithography-free neural networks by free-space optical projection [100]. Furthermore, exploiting synthetic frequency dimensions has been recently explored to extend the length of operand vectors within individual components [101, 102].

Importantly, coherent approaches, in principle, require a single optical source irrespective of the matrix size which can be beneficial for scaling to larger architectures. All-optical cascading becomes straightforward here since the whole network is implemented by one optical seed which can in principle be amplified, interconnected and even recirculated through the network. This theoretical capability for scaling up and cascading represents a considerable advantage of the coherent approach over the incoherent approach, as previously delineated in the remarks on incoherent methodologies. Nonetheless, while scaling and cascading are theoretically feasible, their practical realization presents substantial challenges. The number of interferometers scales with N^2 for an N-to-N mapping. For larger architectures, the cumulative impact of phase errors, induced by factors such as random phase noise, thermal crosstalk, thermal fluctuations, and fabrication imperfections, becomes increasingly pivotal and could significantly compromise the accuracy of coherent addition. Although challenges posed by thermal fluctuations and fabrication imperfections can be mitigated through feedback control and trimming techniques, the eradication of accumulated random phase noise and thermal crosstalk remains a formidable task [103]. This has severe implications on the trainability of the network, particularly doing this off-chip [5]. Several approaches to overcome these limitations have been proposed on an architecture and device level [32, 76, 104-106]. Finally, combining additional orthogonal degrees of freedom becomes non-trivial for coherent approaches due to the narrowband characteristics of the devices. While a precise estimation of the parallelisation capacity via wavelength or mode division multiplexing in each individual technology is currently premature, it is evident that parallelisation becomes increasingly challenging with the size of the network, particularly for MZI-based approaches. Dense multiplexing in diffractive neural networks remains to be demonstrated but could provide a promising outlook for the field.

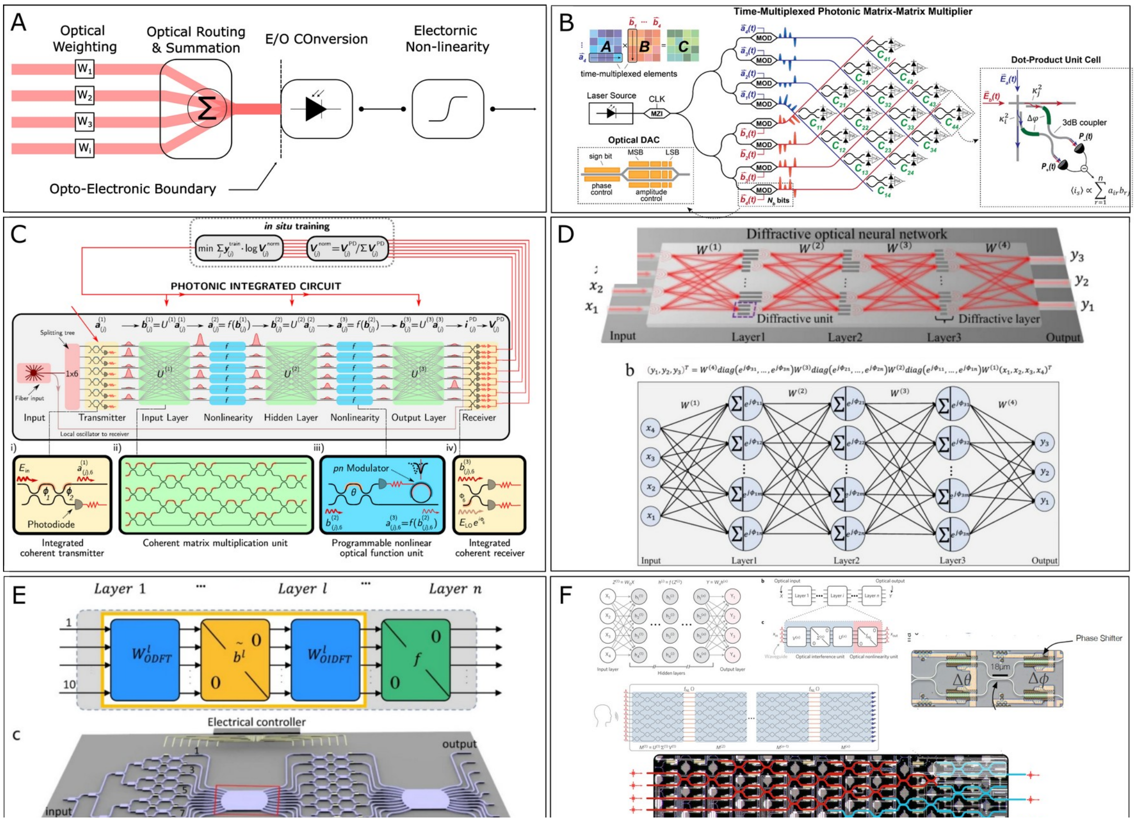


Figure 3: Photonic accelerators based on coherence A) Schematic of coherent superposition of optical signals. B) Coherent photonic crossbar arrays for large-scale matrix-matrix multiplication [107] C) Single chip photonic deep neural network with accelerated training [108]. D) Photonic machine learning with on-chip diffractive optics [89] E) Space-efficient optical computing with an integrated chip diffractive neural network [92]. F) Deep learning with coherent nanophotonic circuits [72]

Optical Nonlinear Neurons

While photonics provides clear advantages in terms of latency, compute density and in certain cases in energy for linear operations, incorporating efficient all-optical nonlinearities remains a considerable challenge. A compromise needs to be made in most cases between operating in the optical domain and exploiting weak optical nonlinear effects or employing the non-linear activation functions electronically. In the latter case, high responsivity photodetectors and high efficiency modulators are needed or else transimpedance gain is required to amplify the signal from the photodiode [109]. A choice also has to be made on whether to convert the signal back to the optical domain or proceed with CMOS.

All-Optical Nonlinearities

Optical nonlinearities rely on the interaction of the optical field with the material. The nonlinear effects are typically relatively weak and therefore amplification of their effect is achieved through resonators and interferometers. Some examples of nonlinearity sources include saturable absorbers, free-carrier dispersion and optical Kerr effects [109], semiconductor optical amplifiers (SOAs) [110] and lasers [111-113], erbium-doped amplifiers (EDFAs) [21, 114] as well as material transformations in phase-change materials (PCMs) [115]. Prucnal and colleagues demonstrated an all-optical neuron in 2020 which comprised of a cavity-loaded MZI device on a silicon photonic platform which exploited the free carrier dispersion effect to demonstrate sigmoid, radial-basis, clamped rectified linear unit, and softplus functions, with tunable thresholds [109]. (figure 4 A) Phase change materials have also been demonstrated as all-optical non-linear activation functions. Here the material performs a thresholding of an input signal as it experiences a phase transition above the crystallisation temperature or melting point for ultrashort pulses which induce amorphization [115, 116] [117]. This approach was recently employed to show multilevel switching between 45 unique synaptic weights for long-term depression (LTD) and long-term potentiation (LTP) [118]. The transient properties of dynamic switching in PCMs was also explored for all-optical filtering in the time domain [119].

Opto-electronic nonlinearities

Although potentially slower and energy intensive, converting the intensity of an optical signal to the electronic domain can provide a pathway for implementing stronger non-linearities.

One of the most prevalent approaches is the use of electrooptical modulators to operate on the signal of a photodiode [120-124]. In most cases, an amplification stage is required for the optical signal of the photodetector in order to re-apply the signal to the neuron. Finally, VCSELs, can be employed as they react nonlinearly to both optical and electronic injection in addition to high wall efficiency (30%), multi-GHz modulation speeds and fJ energy consumption. [35].

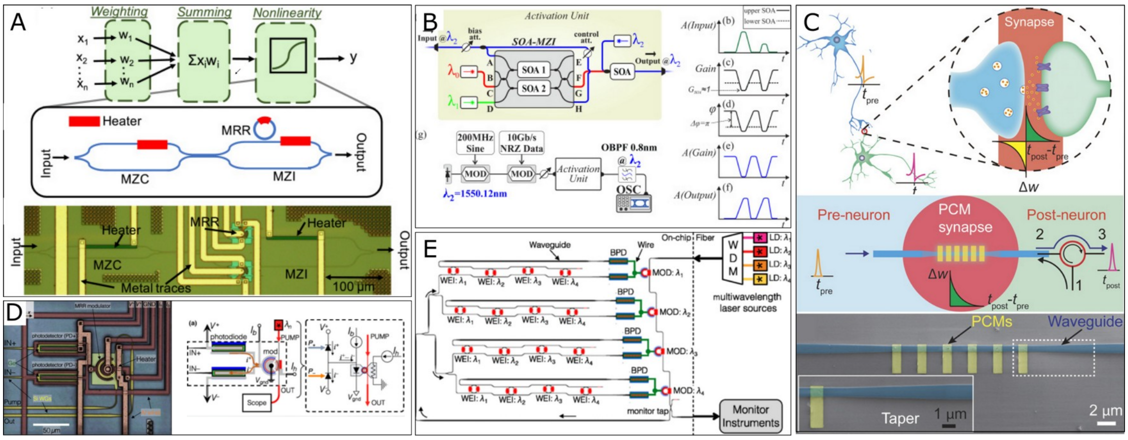


Figure 4: Optical nonlinearity implementations in photonic ANNs A) Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics [109]. B) An all-optical neuron with sigmoid activation function [110] C) On-chip photonic synapse based on PCMs [113, 115]. D,E) Silicon Photonic Modulator Neuron [122]

Reservoir Computing

Photonic reservoir computing has attracted significant attention as it reduces the complexity by no longer requiring to control all individual nodes of a neural network while maintaining computational speeds. From a systems perspective reservoir computing (RC) makes up a unique sub-category of recurrent neural networks, where the “trainer” is agnostic to the physical processes occurring in the inner layers of the system. Here, the reservoir is treated as a black box, with random, temporally fixed nonlinear nodes wherein training they system entails the modification of a set of nodes that form the output layer [125]. By virtue of the recurrence, the output of the reservoir depends on the previous time-steps thereby possessing characteristics of short-term leaky memory [126]. Any highly interconnected, stable, nonlinear dynamical system that exhibits the characteristics of fading memory can be exploited as a reservoir [127]. Owing to the lower complexity of RC schemes, several photonic implementations have naturally emerged, particularly in free-space applications where augmenting the complexity of the reservoir is not spatially bound [128][129]. In guided optical systems, this was demonstrated in rudimentary form in a system consisting of a single nonlinear node originating from OEO conversions and a delay line [130, 131]. A similar approach was undertaken by Larger et al., demonstrating spoken digit recognition and time series prediction tasks [132]. Vandoorne and co-workers explored an integrated silicon photonic reservoir employing a semiconductor optical amplifier (SOA) as the source of nonlinearity [133] with an experimental demonstration that followed [134]. Several approaches of reservoirs consisting of a coupled grid of microring resonators have also been explored [135-138] as well as alternative sources of nonlinearity which aim at reducing the consumption of the SOA while maintaining the required strength [128, 139]. A large-scale integrated RC approach was demonstrated by Nakajima et al., where both the input signals and reservoir weights were optically encoded in the spatiotemporal domain with parallel processing capabilities based on wavelength multiplexing [140]. In this implementation, passive, linear optical components were employed while the nonlinearity was applied by the photodetector upon EO conversion.

Photonic Ising Machines

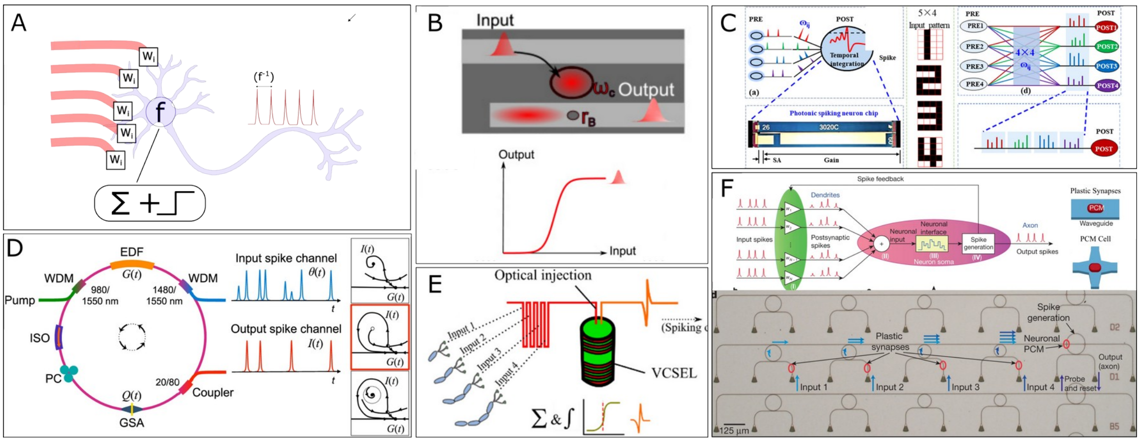
Photonic Ising machines (PIM) have recently emerged as an alternative computing modality to tackle computational tasks for which classical Von-Neumann architectures are inherently ill-suited [141]. A common example which can be mapped in such hardware is the travelling salesman problem which requires the shortest possible route between multiple cities to be computed [142]. This optimisation task along with many other nondeterministic polynomial-time–hard (NP-hard) problems become exponentially more computationally intensive as the size of the input is expanded [143, 144]– in this case with the number of cities. Approaches employing the Ising model – a statistical model of interacting bistable states called “spins” – map the problem to the total energy of the system wherein the ground state represents the solution. In the presence of energy minima, Ising machines typically do not solve for the absolute ground state, but rather yield an approximate solution at dramatically improved speeds. A proliferation of photonic Ising machines appeared in the literature in the middle of the previous decade primarily employing networks of coupled optical parametric oscillators (OPOs) [141, 145]. Initially, demonstrations were limited to a hundred spins and were soon after expanded to networks of 2000 spins [146, 147] . Integrated photonic approaches to implement part of the Ising machine have since been explored [148, 149]. More recently, freespace implementations of photonic Ising machines using spatial light modulators have attracted significant attention primarily owing to their ability to scale up the number of nodes [150-152].

Spiking Neural Networks

Biological neurons exploit the temporal dynamics of signals to process information encoded in short pulses termed spikes (fig. 5a). Spiking implies digitisation of the signals in amplitude and analogue processing in the temporal domain. Spiking Neural Networks (SNNs) operate on abrupt, short pulses with consistent duration and amplitude. This approach can be well suited or photonic implementations because the short spikes or pulses are not inhibited by capacitive delays that are otherwise common in electronic implementations. In a SNN, each neuronal unit remains active solely when receiving or emitting spikes, rendering idle any neurons that do not encounter any events [4]. This inherent sparsity in the computational task is partly responsible for the energy efficiency in biological brains [153]. Yet, as a highly nonlinear process, the generation of- and processing via- spikes remains challenging to efficiently emulate by all-optical circuits. Additionally, the generation of the spikes needs to be accompanied by the inhibitory dynamics found in biological neurons. The majority of recent efforts have exploited inherent system instabilities operating in the excitable regime or have implemented an active medium such as a PCM.

The most common approach for optical SNNs employing VCSELs and operating in an excitable regime where the dynamical system properties induce all-or-none responses. This can be achieved by integrating the VCSEL with a saturable absorber such as graphene or operating in an injection-locked excitable state [35, 154, 155]. Initial investigations of spiking dynamics were realised by Hurtado et. al, where polarisation switching was exploited to mimic phase (single spike generation) and tonic spiking (generation of serial spikes).[111, 156-158] The authors demonstrated the generation of sub nanosecond pulses for orthogonal and parallel injection. An alternative approach to produce spikes is through the integration of a saturable absorber with the laser. Optical pumping of micropillar VCSEL arrays were shown to exhibit spiking dynamics and explored for absolute and relative refractory times, spike latency, temporal summation and computations. [159-161]. In 2020, ultra-fast optical spikes were demonstrated for edge detection in response to optical inputs, demonstrating all-optical spike-based computing [162, 163]. While applications of VCSELs to SNN is promising, an important challenge is their integration with integrated waveguides. Here, additive manufacturing of polymer photonic interconnects to connect the two platforms is an exciting prospect [164-168]. Monolithic integration of nano lasers on the other hand could be an exciting prospect [169]. Recently two emerging technologies have been explored for spiking networks, particularly, optical parametric oscillators (which were covered in the context of Ising machines) [170] and resonant tunnelling diodes (RTDs) [171]. Although RTDs do not provide a homogeneously optical solution, their structural simplicity could provide highly complex non-linear behaviour [172-174].

An alternative approach to produce time dependent spike dynamics is through the use of PCMs. Feldmann et. al. demonstrated in 2019 a spiking neuron where input spikes are



weighted using PCM cells and summed up using a multiplexer. The device then performs the integration of the power and a spike is generated when the power is above a certain threshold required to induce a phase transition [175]. One particular challenge with this approach is that although reversible, the material does not spontaneously return to its initial state due to the nonvolatility of PCM used. An additional operation is required to bring the material to it’s initial state for the neuron to fire again which could however be achieved using a constant repetition amorphization signal. Alternative materials with tuneable volatility such as vanadium dioxide (VO_2) [176] and elemental antimony (Sb) [177] could resolve this limitation.

Figure 5: Spiking and photonic SNNs A) Schematic. B) All-optical non-linear activation function for neuromorphic photonic computing using semiconductor Fano lasers [113]. C) Photonic computing with a spiking neuron chip based on an integrated Fabry–Perot laser with a saturable absorber [178]. D) Graphene laser-based all-optical fiber neurons. [179] E) Ultrafast neuromorphic photonic image processing with a VCSEL neuron [180]F) All-optical spiking neurosynaptic networks with self-learning capabilities [175]

Box 1: Integrated Photonics Going 3D

Since early works in optical computing and photonic neural networks first appeared, there has been a dichotomy between integrated and freespace photonic implementations [64, 181-183]. Guided optical systems in principle provide a clearer path to miniaturisation and integration while freespace optical systems provide greater flexibility and additional spatial degrees of freedom. Recent advances in nanofabrication, primarily using 3D printing of transparent polymers [184, 185] have led to the conception of devices, systems and networks that in principle combine the advantages of both (fig. 6) [186-188]. Photonic 3D printing for interconnects in particular has provided a clear pathway to merging different material platforms by seamlessly coupling between heterogenous chiplets [167, 189]. A multitude of 3D components including 1to N cascadable couplers have already been demonstrated with excellent transmission and broadband characteristics which could help augment the multiplexing capacity of PICs (fig. 6b) [164, 190]. One particular challenge moving forward, is transitioning form passive to active 3D photonic components. Achieving this will naturally require innovations from both the optics, nanofabrication and materials communities in order to embed information storage, optical modulation, detection and amplification capabilities.

While it remains to be implemented at scale, an exciting prospect for the field is the integration of conventionally freespace active components such as VCSEL arrays, SLMs and volume storage and diffractive elements with integrated circuits [191-193].

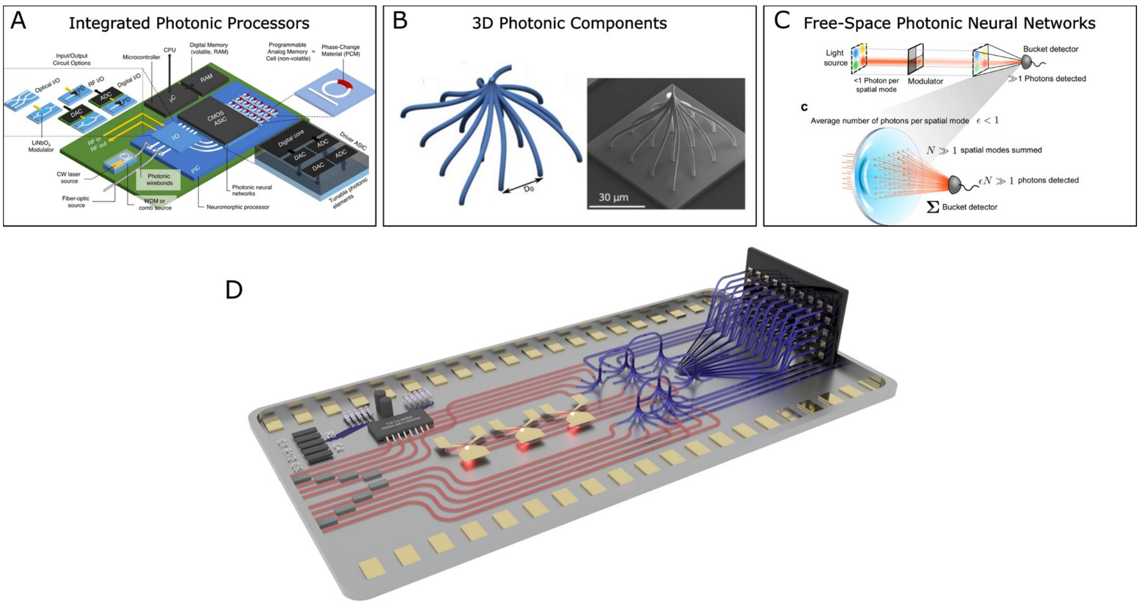


Figure 6: Photonic wirebonding could be pivotal in merging integrated and free-space optical computing A) Illustration of the future integrated optoelectronic processors [11]. B) 3D printed multimode photonic interconnects[164]. C) A free-space optical neural network using less than 1 photon per multiplication [194]. D) Illustration of a photonic processor incorporating integrated, freespace and 3D optical elements

Analysis and Conclusions

We conclude this review by ofthe recent advancements in photonic neuromorphic computing hardware, with a particular emphasis on pinpointing the specific computing step at which the conversion from optical to electrical takes place. Any operations that occur before this pivotal step are indicative of the tasks that are transferred from electronics to photonics, capitalizing on the superior processing bandwidth, energy efficiency, parallelism, and reduced latency inherent to photonics. This pivotal step itself manifests as a bottleneck, marking the intersection of the optical and electrical domains, necessitating a matching of their respective operating bandwidths and requiring conversion and control overheads within the system. Subsequent steps, implemented electronically, elucidate the prevailing challenges and guide the trajectory for future advancements, with the objective of integrating an increasing number of neuromorphic computing steps into the photonic domain. We foresee that significant impact resides in augmenting the scalability and cascadability of photonics and in amplifying the nonlinear optical responses. The enhancement of scalability and cascadability is pivotal, aspiring to confer to photonic neuromorphic computing a level of interconnectivity that is closer and closer to that of the brain. Three key aspects to realize heightened scalability and cascadability include the minimization of optical losses, the optimization of control and feedback systems and algorithms, and the efficacious utilization of optical degrees of freedom. The amplification of nonlinear optical responses is instrumental in enabling all-optical spiking neural networks to leverage the efficiency gains associated with temporal processing. Given that the conventional methodologies for harnessing optical nonlinearity are frequently inefficient, we anticipate that advancements at the material level for higher optical nonlinearity, coupled with device-level designs to actualize elevated nonlinear response, will emerge as promising solutions.

While the development of optical nonlinear neurons and their associated neural networks remains predominantly at the research stage, the evolution of photonic accelerator technologies—which facilitate linear operations such as coherent and incoherent MVMs—has demonstrated substantial industrial relevance. These technologies are nearing a stage where they can make a tangible impact in practical applications. In the near-term, harmonizing both photonic and electronic elements, is realized. In this paradigm, photonic accelerators would supplement the electronic processor by managing the substantial linear matrix-vector-multiplication (MVM) operations, while the electronic processor would be responsible for executing nonlinear activation functions, logical operations, control, and synchronization. To appraise the efficacy of any proposed photonic accelerator, we have delineated four key Figures of Merit (FOMs):

i. Throughput, measured in tera-operations per second (TOPS),
ii. Latency, specific to each MVM operation
iii. Power efficiency, measured in TOPS/W,
iv. Compute density, measured in TOPS/mm^2.

A comprehensive derivation and critical analysis pertaining to the evaluation of these FOMs are discussed in [33]. For the sake of brevity and focused discourse in this context, we will primarily underscore the major outcomes of the evaluation. We consider a scenario where the photonic accelerator is engaged in processing Q parallel input vectors of dimension M using N kernels, operating at an input sampling rate of f with 4-bit resolution. For a moderate scaling scenario where M=10, N=10, Q=4, and f=10 GSa/s, the photonic accelerator exhibits latency that is enhanced by more than seven orders of magnitude across all architectures when compared to the latest Google TPU v4, an ASIC specifically designed for tensor operations. This substantial reduction in latency can be attributed to the absence of capacitive delay, a primary bottleneck in electronic systems. Latency in photonic accelerators is exclusively determined by the optical path length. While the throughput is observed to be lower due to the small scale of photonic accelerators, the power efficiency demonstrates an improvement of around one order of magnitude, and the compute density remains comparable.

Projecting into a future scaling scenario with parameters M=100, N=100, Q=16, and f=50 GSa/s, the advantage in latency remains very significantly. Employing architectures such as the MZI mesh, MRR array, or crossbar array results in the throughput, power efficiency, and compute density all surpassing Google TPU v4 by two orders of magnitude. However, when utilizing the charge accumulation method, these three FOMs are observed to be comparable to Google TPU v4. In the charge accumulation method, the entries of a vector are successively input at the sampling rate f, resulting in a decrease in overall throughput, and consequently impacting the power efficiency and compute density. In contrast, the other three architectures facilitate the simultaneous input of the entries of a vector by leveraging the spatial degree of freedom within one sampling duration, culminating in higher throughput and enhanced performance across the other FOMs. One must note that moving from moderate scaling to future scaling is not as simple as putting small modules together to make a larger architecture. Optical losses, especially insertion losses should be reduced considerably to ensure that the optical power is above the electrical noise at the receivers; robust integrated (de)multiplexing modules accommodating a large optical bandwidth and increased modes should be designed to accommodate the larger number of parallel computing channels; and high-speed high-density electro-optic interfaces should be developed for large scale/high-speed fan-in/fan-out photonic accelerators. Nonetheless, an encouraging result obtained from the above-mentioned evaluation suggests that even a small 6×6 photonic tensor core operating at Q=4, and f=25 GSa/s can achieve power efficiency and compute density comparable to the latest Google TPUv4 [63].

It is noteworthy that, across both moderate and future scaling scenarios, the three distinct architectures — the MZI mesh, MRR array, and crossbar array — exhibit comparable compute densities and power efficiencies. A closer inspection of the power and area consumption patterns within the photonic accelerators provides clarity on this similarity. In each of the aforementioned architectures, the dominant contributors to power consumption are the transmitter, receiver, and ADC. Area consumption is primarily dictated by the transmitter and ADC. Given that these components are essentially the electrical constituents tasked with Electro-Optic-Electro (E-O-E) conversion within the photonic accelerators, their specifications remain consistent across the three architectures. This results in similar power efficiencies and compute densities. These observations underscore the importance of optimizing the power and area consumption of the electrical components, specifically the transmitter, receiver, and ADC, in the quest to engineer photonic accelerators with superior power efficiency and compute density. While fine-tuning the photonic processor will undeniably yield enhanced throughput and reduced latency, the broader impact lies in addressing these electrical elements. Moreover, it is imperative to highlight that as we look towards further scaling, optical loss is poised to become the prevailing bottleneck constraining power efficiency. In such scenarios, the charge accumulation method stands out with a discernible advantage. Its ability to aggregate multiple low-power optical signals, thereby producing an output that surpasses the detection threshold, effectively ameliorates the demands on optical power.

Acknowledgements: The authors acknowledge discussions with A.Ne. **Funding**: This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101098717, (Hybrain Project). The European Union's Horizon Europe research and innovation programme under grant agreements \ No. 101017237, PHOENICS Project and No. 101098717 RESPITE Project). This research was also supported by EPSRC via grant EP/W022931/1 and EP/W034387/1

Author Contributions: NF, BD and HB wrote the manuscript. All authors provided in-depth discussions and suggestions at all stages of the work and discussed the results. HB led the work.

Competing Interests: The authors declare no conflict of interest.

References

1. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, 2012. **25**.
2. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.
3. Amir, A., et al. *Cognitive computing programming paradigm: a corelet language for composing networks of neurosynaptic cores*. in *The 2013 International Joint Conference on Neural Networks (IJCNN)*. 2013. IEEE.
4. Roy, K., A. Jaiswal, and P. Panda, *Towards spike-based machine intelligence with neuromorphic computing.* Nature, 2019. **575**(7784): p. 607-617.
5. Wetzstein, G., et al., *Inference in artificial intelligence with deep optics and photonics.* Nature, 2020. **588**(7836): p. 39-47.
6. Mead, C., *Neuromorphic electronic systems.* Proceedings of the IEEE, 1990. **78**(10): p. 1629-1636.
7. Moore, G.E., *Cramming more components onto integrated circuits.* Proceedings of the IEEE, 1998. **86**(1): p. 82-85.
8. Waldrop, M.M., *The chips are down for Moore's law.* Nature News, 2016. **530**(7589): p. 144.
9. Koomey, J., et al., *Implications of historical trends in the electrical efficiency of computing.* IEEE Annals of the History of Computing, 2010. **33**(3): p. 46-54.
10. Brown, R.E., et al., *Report to congress on server and data center energy efficiency: Public law 109-431*. 2007, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).
11. Shastri, B.J., et al., *Photonics for artificial intelligence and neuromorphic computing.* Nature Photonics, 2021. **15**(2): p. 102-114.
12. Davies, M., et al., *Loihi: A neuromorphic manycore processor with on-chip learning.* Ieee Micro, 2018. **38**(1): p. 82-99.
13. Merolla, P.A., et al., *A million spiking-neuron integrated circuit with a scalable communication network and interface.* Science, 2014. **345**(6197): p. 668-673.
14. Furber, S.B., et al., *The spinnaker project.* Proceedings of the IEEE, 2014. **102**(5): p. 652-665.
15. Meng, X., et al., *Compact optical convolution processing unit based on multimode interference.* Nature Communications, 2023. **14**(1): p. 3000.
16. Nahmias, M.A., et al., *Photonic multiply-accumulate operations for neural networks.* IEEE Journal of Selected Topics in Quantum Electronics, 2019. **26**(1): p. 1-18.
17. Xiang, C., et al., *3D integration enables ultralow-noise isolator-free lasers in silicon photonics.* Nature, 2023. **620**(7972): p. 78-85.
18. He, M., et al., *High-performance hybrid silicon and lithium niobate Mach–Zehnder modulators for 100 Gbit s− 1 and beyond.* Nature Photonics, 2019. **13**(5): p. 359-364.
19. Rahim, A., et al., *Taking silicon photonics modulators to a higher performance level: state-of-the-art and a review of new technologies.* Advanced Photonics, 2021. **3**(2): p. 024003-024003.
20. Atabaki, A.H., et al., *Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip.* Nature, 2018. **556**(7701): p. 349-354.
21. Liu, Y., et al., *A photonic integrated circuit–based erbium-doped amplifier.* Science, 2022. **376**(6599): p. 1309-1313.
22. Youngblood, N., et al., *Integrated optical memristors.* Nature Photonics, 2023: p. 1-12.
23. Farmakidis, N., et al., *Electronically reconfigurable photonic switches incorporating plasmonic structures and phase change materials.* Advanced Science, 2022. **9**(20): p. 2200383.
24. Farmakidis, N., et al., *Plasmonic nanogap enhanced phase-change devices with dual electrical-optical functionality.* Science Advances, 2019. **5**(11): p. eaaw2687.
25. Ríos, C., et al., *Integrated all-photonic non-volatile multi-level memory.* Nature photonics, 2015. **9**(11): p. 725-732.

26. Dai, D., *Silicon nanophotonic integrated devices for on-chip multiplexing and switching*. Journal of Lightwave Technology, 2016. **35**(4): p. 572-587.
27. Hochberg, M. and T. Baehr-Jones, *Towards fabless silicon photonics.* Nature photonics, 2010. **4**(8): p. 492-494.
28. Xiang, S., et al., *A review: Photonics devices, architectures, and algorithms for optical neural computing*. Journal of Semiconductors, 2021. **42**(2): p. 023105.
29. Prucnal, P.R. and B.J. Shastri, *Neuromorphic photonics*. 2017: CRC Press.
30. Xu, X., et al., *11 TOPS photonic convolutional accelerator for optical neural networks.* Nature, 2021. **589**(7840): p. 44-51.
31. Peserico, N., B.J. Shastri, and V.J. Sorger, *Integrated photonic tensor processing unit for a matrix multiply: a review.* Journal of Lightwave Technology, 2023.
32. Al-Qadasi, M., et al., *Scaling up silicon photonic-based accelerators: Challenges and opportunities*. APL Photonics, 2022. **7**(2).
33. Dong, B., *A perspective on photonic accelerators.*
34. Capmany, J. and D. Pérez, *Programmable Integrated Photonics*. 2020: Oxford University Press.
35. Skalli, A., et al., *Photonic neuromorphic computing using vertical cavity semiconductor lasers.* Optical Materials Express, 2022. **12**(6): p. 2395-2414.
36. Feldmann, J., et al., *Parallel convolutional processing using an integrated photonic tensor core.* Nature, 2021. **589**(7840): p. 52-58.
37. Tait, A.N., et al., *Microring weight banks.* IEEE Journal of Selected Topics in Quantum Electronics, 2016. **22**(6): p. 312-325.
38. Ling, Q., et al., *On-chip optical matrix-vector multiplier based on mode division multiplexing*. Chip, 2023: p. 100061.
39. Lu, L., et al., *16× 16 non-blocking silicon optical switch based on electro-optic Mach-Zehnder interferometers.* Optics express, 2016. **24**(9): p. 9295-9307.
40. Yang, L., et al., *On-chip CMOS-compatible optical signal processor.* Optics express, 2012. **20**(12): p. 13560-13565.
41. Tait, A.N., et al., *Continuous calibration of microring weights for analog optical networks.* IEEE Photonics Technology Letters, 2016. **28**(8): p. 887-890.
42. Marquez, B.A., et al., *Fully-integrated photonic tensor core for image convolutions.* Nanotechnology, 2023.
43. Tait, A.N., et al., *Broadcast and weight: an integrated network for scalable photonic spike processing*. Journal of Lightwave Technology, 2014. **32**(21): p. 3427-3439.
44. Bai, B., et al., *Microcomb-based integrated photonic processing unit.* Nature Communications, 2023. **14**(1): p. 66.
45. Farmakidis, N., et al., *Scalable High-Precision Trimming of Photonic Resonances by Polymer Exposure to Energetic Beams.* Nano Letters, 2023.
46. Zhang, Y., et al., *Transient tap couplers for wafer-level photonic testing based on optical phase change materials.* ACS Photonics, 2021. **8**(7): p. 1903-1908.
47. Chen, R., et al., *Non-volatile electrically programmable integrated photonics with a 5-bit operation*. Nature Communications, 2023. **14**(1): p. 3465.
48. Ríos, C., et al., *In-memory computing on a photonic platform.* Science advances, 2019. **5**(2): p. eaau5759.
49. Brückerhoff-Plückelmann, F., et al., *Event-driven adaptive optical neural network.* Science Advances, 2023. **9**(42): p. eadi9127.
50. Lee, J.S., et al., *Spatio-spectral control of coherent nanophotonics.* Nanophotonics, 2024(0).
51. Ohno, S., et al., *Si microring resonator crossbar array for on-chip inference and training of the optical neural network.* ACS Photonics, 2022. **9**(8): p. 2614-2622.
52. Li, X., et al. *On-chip phase change optical matrix multiplication core*. in *2020 IEEE International Electron Devices Meeting (IEDM)*. 2020. IEEE.

53. Brückerhoff-Plückelmann, F., et al., *Broadband photonic tensor core with integrated ultra-low crosstalk wavelength multiplexers.* Nanophotonics, 2022. **11**(17): p. 4063-4072.
54. Xiao, X., et al. *Wavelength-Parallel Photonic Tensor Core Based on Multi-FSR Microring Resonator Crossbar Array*. in *Optical Fiber Communication Conference*. 2023. Optica Publishing Group.
55. Wendland, D., et al., *Coherent dimension reduction with integrated photonic circuits exploiting tailored disorder.* JOSA B, 2023. **40**(3): p. B35-B40.
56. Aggarwal, S., et al., *Reduced rank photonic computing accelerator.* Optica, 2023. **10**(8): p. 1074-1080.
57. Zhou, W., et al., *In-memory photonic dot-product engine with electrically programmable weight banks.* Nature Communications, 2023. **14**(1): p. 2887.
58. Wei, M., et al., *Electrically programmable phase-change photonic memory for optical neural networks with nanoseconds in situ training capability.* Advanced Photonics, 2023. **5**(4): p. 046004.
59. Brückerhoff-Plückelmann, F., et al. *Hybrid Electro-Optic Crossbar Array for Matrix-Vector Multiplications*. in *CLEO: Science and Innovations*. 2023. Optica Publishing Group.
60. Qu, Y., et al., *Inverse design of an integrated-nanophotonics optical neural network.* Science Bulletin, 2020. **65**(14): p. 1177-1183.
61. Lee, J.S., et al., *Polarization-selective reconfigurability in hybridized-active-dielectric nanowires.* Science Advances, 2022. **8**(24): p. eabn9459.
62. Alam, M.S., et al., *Photonic Integrated Circuit for Rapidly Tunable Orbital Angular Momentum Generation Using Sb2Se3 Ultra-Low-Loss Phase Change Material.* Advanced Optical Materials, 2022. **10**(20): p. 2200098.
63. Dong, B., et al., *Higher-dimensional processing using a photonic tensor core with continuous-time data.* Nature Photonics, 2023. **17**(12): p. 1080-1088.
64. Miller, D.A., *Self-configuring universal linear optical component.* Photonics Research, 2013. **1**(1): p. 1-15.
65. Reck, M., et al., *Experimental realization of any discrete unitary operator.* Physical review letters, 1994. **73**(1): p. 58.
66. Ribeiro, A., et al., *Demonstration of a 4× 4-port universal linear circuit.* Optica, 2016. **3**(12): p. 1348-1357.
67. Clements, W.R., et al., *Optimal design for universal multiport interferometers.* Optica, 2016. **3**(12): p. 1460-1465.
68. Miller, D.A., *Self-aligning universal beam coupler.* Optics express, 2013. **21**(5): p. 6360-6370.
69. Mennea, P.L., et al., *Modular linear optical circuits.* Optica, 2018. **5**(9): p. 1087-1090.
70. Zhou, Y.K., et al., *Thermal performance and optimized thickness of active shape-stabilized PCM boards for side-wall cooling and under-floor heating system.* Indoor and Built Environment, 2016. **25**(8): p. 1279-1295.
71. Harris, N.C., et al., *Linear programmable nanophotonic processors.* Optica, 2018. **5**(12): p. 1623-1631.
72. Shen, Y., et al., *Deep learning with coherent nanophotonic circuits.* Nature Photonics, 2017. **11**(7): p. 441.
73. Pai, S., et al., *Parallel programming of an arbitrary feedforward photonic network.* IEEE Journal of Selected Topics in Quantum Electronics, 2020. **26**(5): p. 1-13.
74. Bagherian, H., et al., *On-chip optical convolutional neural networks.* arXiv preprint arXiv:1808.03303, 2018.
75. Demirkiran, C., et al., *An electro-photonic system for accelerating deep neural networks.* ACM Journal on Emerging Technologies in Computing Systems, 2023. **19**(4): p. 1-31.
76. Giamougiannis, G., et al., *Analog nanophotonic computing going practical: silicon photonic deep learning engines for tiled optical matrix multiplication with dynamic precision.* Nanophotonics, 2023. **12**(5): p. 963-973.

77. Moralis-Pegios, M., et al., *Neuromorphic Silicon Photonics and Hardware-Aware Deep Learning for High-Speed Inference.* Journal of Lightwave Technology, 2022. **40**(10): p. 3243-3254.
78. Zhou, H., et al., *Self-configuring and reconfigurable silicon photonic signal processor.* Acs Photonics, 2020. **7**(3): p. 792-799.
79. Totovic, A., et al., *Programmable photonic neural networks combining WDM with coherent linear optics.* Scientific Reports, 2022. **12**(1).
80. Choutagunta, K., et al., *Adapting Mach–Zehnder mesh equalizers in direct-detection mode-division-multiplexed links.* Journal of Lightwave Technology, 2019. **38**(4): p. 723-735.
81. Ding, C., et al. *Circnn: accelerating and compressing deep neural networks using block-circulant weight matrices*. in *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture*. 2017.
82. Nakajima, M., et al., *Physical deep learning with biologically inspired training method: gradient-free approach for physical hardware.* Nature Communications, 2022. **13**(1): p. 7847.
83. Cruz-Cabrera, A.A., et al., *Reinforcement and backpropagation training for an optical neural network using self-lensing effects.* IEEE transactions on neural networks, 2000. **11**(6): p. 1450-1457.
84. Hughes, T.W., et al., *Training of photonic neural networks through in situ backpropagation and gradient measurement.* Optica, 2018. **5**(7): p. 864-871.
85. Pai, S., et al., *Experimentally realized in situ backpropagation for deep learning in photonic neural networks.* Science, 2023. **380**(6643): p. 398-404.
86. Lin, X., et al., *All-optical machine learning using diffractive deep neural networks.* Science, 2018. **361**(6406): p. 1004-1008.
87. Qian, C., et al., *Performing optical logic operations by a diffractive neural network.* Light: Science & Applications, 2020. **9**(1): p. 59.
88. Zhou, T., et al., *Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit.* Nature Photonics, 2021. **15**(5): p. 367-373.
89. Fu, T., et al., *Photonic machine learning with on-chip diffractive optics.* Nature Communications, 2023. **14**(1): p. 70.
90. Wang, Z., et al., *Integrated photonic metasystem for image classifications at telecommunication wavelength.* Nature communications, 2022. **13**(1): p. 2131.
91. Chen, Z., et al., *Deep learning with coherent VCSEL neural networks.* Nature Photonics, 2023: p. 1-8.
92. Zhu, H., et al., *Space-efficient optical computing with an integrated chip diffractive neural network.* Nature communications, 2022. **13**(1): p. 1044.
93. Liao, K., et al., *Integrated Photonic Neural Networks: Opportunities and Challenges.* ACS Photonics, 2023.
94. Seok, T.J., et al., *Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers.* Optica, 2016. **3**(1): p. 64-70.
95. Sattari, H., et al., *Silicon photonic MEMS phase-shifter.* Optics Express, 2019. **27**(13): p. 18959-18969.
96. Sun, H., et al., *Silicon photonic phase shifters and their applications: A review.* Micromachines, 2022. **13**(9): p. 1509.
97. Qi, Y. and Y. Li, *Integrated lithium niobate photonics.* Nanophotonics, 2020. **9**(6): p. 1287-1320.
98. Youngblood, N., *Coherent Photonic Crossbar Arrays for Large-Scale Matrix-Matrix Multiplication.* Ieee Journal of Selected Topics in Quantum Electronics, 2023. **29**(2).
99. Giamougiannis, G., et al., *A Coherent Photonic Crossbar for Scalable Universal Linear Optics.* Journal of Lightwave Technology, 2023. **41**(8): p. 2425-2442.
100. Wu, T., et al., *Lithography-free reconfigurable integrated photonic processor.* Nature Photonics, 2023: p. 1-7.

101. Fan, L., et al., *Experimental Realization of Convolution Processing in Photonic Synthetic Frequency Dimensions*. arXiv preprint arXiv:2305.03250, 2023.
102. Zhao, H., et al., *Enabling scalable optical computing in synthetic frequency dimension using integrated cavity acousto-optics.* Nature Communications, 2022. **13**(1): p. 5426.
103. Mourgias-Alexandris, G., et al., *Noise-resilient and high-speed deep learning with coherent silicon photonics.* Nature Communications, 2022. **13**(1).
104. Varri, A., et al., *Scalable Non-Volatile Tuning of Photonic Computational Memories by Automated Silicon Ion Implantation.* Advanced Materials, 2023: p. 2310596.
105. Jayatilleka, H., et al., *Post-fabrication trimming of silicon photonic ring resonators at wafer-scale.* Journal of Lightwave Technology, 2021. **39**(15): p. 5083-5088.
106. Giamougiannis, G., et al., *Universal linear optics revisited: new perspectives for neuromorphic computing with silicon photonics.* IEEE Journal of Selected Topics in Quantum Electronics, 2022. **29**(2: Optical Computing): p. 1-16.
107. Youngblood, N., *Coherent photonic crossbar arrays for large-scale matrix-matrix multiplication.* IEEE Journal of Selected Topics in Quantum Electronics, 2022. **29**(2: Optical Computing): p. 1-11.
108. Bandyopadhyay, S., et al., *Single chip photonic deep neural network with accelerated training*. arXiv preprint arXiv:2208.01623, 2022.
109. Jha, A., C. Huang, and P.R. Prucnal, *Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics.* Optics letters, 2020. **45**(17): p. 4819-4822.
110. Mourgias-Alexandris, G., et al., *An all-optical neuron with sigmoid activation function.* Optics express, 2019. **27**(7): p. 9620-9630.
111. Hurtado, A., et al., *Investigation of vertical cavity surface emitting laser dynamics for neuromorphic photonic systems.* Applied Physics Letters, 2012. **100**(10).
112. Crnjanski, J., et al., *Adaptive sigmoid-like and PReLU activation functions for all-optical perceptron*. Optics Letters, 2021. **46**(9): p. 2003-2006.
113. Rasmussen, T.S., Y. Yu, and J. Mork, *All-optical non-linear activation function for neuromorphic photonic computing using semiconductor Fano lasers.* Optics Letters, 2020. **45**(14): p. 3844-3847.
114. Rönn, J., et al., *Ultra-high on-chip optical gain in erbium-based hybrid slot waveguides.* Nature communications, 2019. **10**(1): p. 432.
115. Cheng, Z., et al., *On-chip photonic synapse.* Science advances, 2017. **3**(9).
116. Zhou, W., et al., *Phase-change materials for energy-efficient photonic memory and computing*. MRS Bulletin, 2022. **47**(5): p. 502-510.
117. Feldmann, J., et al., *All-optical spiking neurosynaptic networks with self-learning capabilities.* Nature, 2019. **569**(7755): p. 208-214.
118. Zhou, W., et al., *Artificial Biphasic Synapses Based on Nonvolatile Phase-Change Photonic Memory Cells.* physica status solidi (RRL)–Rapid Research Letters, 2022. **16**(9): p. 2100487.
119. Aggarwal, S., et al., *All optical tunable RF filter using elemental antimony.* Nanophotonics, 2024.
120. Pappas, C., et al., *Programmable Tanh-, ELU-, Sigmoid-, and Sin-Based Nonlinear Activation Functions for Neuromorphic Photonics.* Ieee Journal of Selected Topics in Quantum Electronics, 2023. **29**(6).
121. Fard, M.M.P., et al., *Experimental realization of arbitrary activation functions for optical neural networks*. Optics Express, 2020. **28**(8): p. 12138-12148.
122. Tait, A.N., et al., *Silicon photonic modulator neuron.* Physical Review Applied, 2019. **11**(6): p. 064043.
123. George, J.K., et al., *Neuromorphic photonics with electro-absorption modulators.* Optics express, 2019. **27**(4): p. 5181-5191.
124. Amin, R., et al., *ITO-based electro-absorption modulator for photonic neural activation function.* APL Materials, 2019. **7**(8).

125. Jaeger, H. and H. Haas, *Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication*. science, 2004. **304**(5667): p. 78-80.
126. Jaeger, H., *Short term memory in echo state networks*. 2001.
127. Van der Sande, G., D. Brunner, and M.C. Soriano, *Advances in photonic reservoir computing*. Nanophotonics, 2017. **6**(3): p. 561-576.
128. Brunner, D. and I. Fischer, *Reconfigurable semiconductor laser networks based on diffractive coupling*. Optics letters, 2015. **40**(16): p. 3854-3857.
129. Rafayelyan, M., et al., *Large-scale optical reservoir computing for spatiotemporal chaotic systems prediction.* Physical Review X, 2020. **10**(4): p. 041037.
130. Larger, L., et al., *High-speed photonic reservoir computing using a time-delay-based architecture: Million words per second classification.* Physical Review X, 2017. **7**(1): p. 011015.
131. Paquot, Y., et al., *Optoelectronic reservoir computing*. Scientific reports, 2012. **2**(1): p. 287.
132. Larger, L., et al., *Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing*. Optics express, 2012. **20**(3): p. 3241-3249.
133. Vandoorne, K., et al., *Parallel reservoir computing using optical amplifiers.* IEEE transactions on neural networks, 2011. **22**(9): p. 1469-1481.
134. Vandoorne, K., et al., *Experimental demonstration of reservoir computing on a silicon photonics chip.* Nature communications, 2014. **5**(1): p. 3541.
135. Boikov, I., D. Brunner, and A. De Rossi, *Direct coupling of nonlinear integrated cavities for all-optical reservoir computing*. arXiv preprint arXiv:2307.10950, 2023.
136. Staffoli, E., et al., *Nonlinear response of Silicon Photonics microresonators for reservoir computing neural network.* arXiv preprint arXiv:2302.10203, 2023.
137. Denis-Le Coarer, F., et al., *All-optical reservoir computing on a photonic chip using silicon-based ring resonators.* IEEE Journal of Selected Topics in Quantum Electronics, 2018. **24**(6): p. 1-8.
138. Mesaritakis, C., V. Papataxiarhis, and D. Syvridis, *Micro ring resonators as building blocks for an all-optical high-speed reservoir-computing bit-pattern-recognition system.* JOSA B, 2013. **30**(11): p. 3048-3055.
139. Takano, K., et al., *Compact reservoir computing with a photonic integrated circuit.* Optics express, 2018. **26**(22): p. 29424-29439.
140. Nakajima, M., K. Tanaka, and T. Hashimoto, *Scalable reservoir computing on coherent linear photonic processor*. Communications Physics, 2021. **4**(1): p. 20.
141. McMahon, P.L., et al., *A fully programmable 100-spin coherent Ising machine with all-to-all connections*. Science, 2016. **354**(6312): p. 614-617.
142. Roques-Carmes, C. and M. Soljačić, *Photonic Ising machines go big*. Physics, 2019. **12**: p. 61.
143. Roques-Carmes, C., et al., *Heuristic recurrent algorithms for photonic Ising machines.* Nature communications, 2020. **11**(1): p. 249.
144. Hamerly, R., et al., *Experimental investigation of performance differences between coherent Ising machines and a quantum annealer.* Science advances, 2019. **5**(5): p. eaau0823.
145. Marandi, A., et al., *Network of time-multiplexed optical parametric oscillators as a coherent Ising machine.* Nature Photonics, 2014. **8**(12): p. 937-942.
146. Inagaki, T., et al., *A coherent Ising machine for 2000-node optimization problems.* Science, 2016. **354**(6312): p. 603-606.
147. Inagaki, T., et al., *Large-scale Ising spin network based on degenerate optical parametric oscillators.* Nature Photonics, 2016. **10**(6): p. 415-419.
148. Prabhu, M., et al., *Accelerating recurrent Ising machines in photonic integrated circuits.* Optica, 2020. **7**(5): p. 551-558.
149. Vázquez, M.R., et al., *Optical NP problem solver on laser-written waveguide platform.* Optics Express, 2018. **26**(2): p. 702-710.

150. Pierangeli, D., et al., *Noise-enhanced spatial-photonic Ising machine.* Nanophotonics, 2020. **9**(13): p. 4109-4116.
151. Pierangeli, D., G. Marcucci, and C. Conti, *Adiabatic evolution on a spatial-photonic Ising machine.* Optica, 2020. **7**(11): p. 1535-1543.
152. Pierangeli, D., G. Marcucci, and C. Conti, *Large-scale photonic Ising machine by spatial light modulation.* Physical review letters, 2019. **122**(21): p. 213902.
153. Owen-Newns, D., et al., *Photonic spiking neural networks with highly efficient training protocols for ultrafast neuromorphic computing systems.* Intelligent Computing, 2023. **2**: p. 0031.
154. Robertson, J., et al., *Controlled inhibition of spiking dynamics in VCSELs for neuromorphic photonics: theory and experiments.* Optics letters, 2017. **42**(8): p. 1560-1563.
155. Jha, A., et al., *Photonic spiking neural networks and graphene-on-silicon spiking neurons.* Journal of Lightwave Technology, 2022. **40**(9): p. 2901-2914.
156. Hurtado, A., et al., *Nonlinear dynamics induced by parallel and orthogonal optical injection in 1550 nm vertical-cavity surface-emitting lasers (VCSELs).* Optics express, 2010. **18**(9): p. 9423-9428.
157. Al-Seyab, R., et al. *Controlled single-and multiple-pulse excitability in VCSELs for novel spiking photonic neurons*. in *2014 International Semiconductor Laser Conference*. 2014. IEEE.
158. Hurtado, A. and J. Javaloyes, *Controllable spiking patterns in long-wavelength vertical cavity surface emitting lasers for neuromorphic photonics systems.* Applied Physics Letters, 2015. **107**(24).
159. Pammi, V.A., et al., *Photonic computing with single and coupled spiking micropillar lasers.* IEEE Journal of Selected Topics in Quantum Electronics, 2019. **26**(1): p. 1-7.
160. Selmi, F., et al., *Spike latency and response properties of an excitable micropillar laser.* Physical Review E, 2016. **94**(4): p. 042219.
161. Barbay, S., R. Kuszelewicz, and A.M. Yacomotti, *Excitability in a semiconductor laser with saturable absorber.* Optics letters, 2011. **36**(23): p. 4476-4478.
162. Robertson, J., et al., *Image edge detection with a photonic spiking VCSEL-neuron.* Optics Express, 2020. **28**(25): p. 37526-37537.
163. Hejda, M., et al., *Neuromorphic encoding of image pixel data into rate-coded optical spike trains with a photonic VCSEL-neuron.* APL Photonics, 2021. **6**(6).
164. Moughames, J., et al., *3D printed multimode-splitters for photonic interconnects.* Optical Materials Express, 2020. **10**(11): p. 2952-2961.
165. Grabulosa, A., et al., *Additive 3D photonic integration that is CMOS compatible.* Nanotechnology, 2023. **34**(32): p. 322002.
166. Billah, M., et al. *Multi-chip integration of lasers and silicon photonics by photonic wire bonding*. in *CLEO: Science and Innovations*. 2015. Optica Publishing Group.
167. Lindenmann, N., et al., *Photonic wire bonding: a novel concept for chip-scale interconnects.* Optics express, 2012. **20**(16): p. 17667-17677.
168. Lindenmann, N., et al., *Connecting silicon photonic circuits to multicore fibers by photonic wire bonding*. Journal of lightwave Technology, 2015. **33**(4): p. 755-760.
169. Crosnier, G., et al., *Hybrid indium phosphide-on-silicon nanolaser diode.* Nature Photonics, 2017. **11**(5): p. 297-300.
170. Inagaki, T., et al., *Collective and synchronous dynamics of photonic spiking neurons.* Nature communications, 2021. **12**(1): p. 2325.
171. Romeira, B., et al., *Excitability and optical pulse generation in semiconductor lasers driven by resonant tunneling diode photo-detectors.* Optics express, 2013. **21**(18): p. 20931-20940.
172. Hejda, M., et al., *Artificial optoelectronic spiking neuron based on a resonant tunnelling diode coupled to a vertical cavity surface emitting laser.* Nanophotonics, 2022. **12**(5): p. 857-867.

173. Hejda, M., et al., *Resonant tunneling diode nano-optoelectronic excitable nodes for neuromorphic spike-based information processing*. Physical Review Applied, 2022. **17**(2): p. 024072.

174. Lourenco, J., et al. *Resonant Tunnelling Diode–Photodetectors for spiking neural networks*. in *Journal of Physics: Conference Series*. 2022. IOP Publishing.

175. Feldmann, J., et al., *All-optical spiking neurosynaptic networks with self-learning capabilities*. Nature, 2019. **569**(7755): p. 208-214.

176. Ryckman, J.D., et al., *Ultra-compact silicon photonic devices reconfigured by an optically induced semiconductor-to-metal transition*. Optics express, 2013. **21**(9): p. 10753-10763.

177. Aggarwal, S., et al., *Antimony as a programmable element in integrated nanophotonics*. Nano Letters, 2022. **22**(9): p. 3532-3538.

178. Xiang, S., et al., *Hardware-algorithm collaborative computing with photonic spiking neuron chip based on an integrated Fabry–Perot laser with a saturable absorber*. Optica, 2023. **10**(2): p. 162-171.

179. Shastri, B.J., et al. *Graphene excitable laser for photonic spike processing*. in *2013 IEEE Photonics Conference*. 2013. IEEE.

180. Robertson, J., et al., *Ultrafast neuromorphic photonic image processing with a VCSEL neuron*. Scientific reports, 2022. **12**(1): p. 4874.

181. Li, H.-Y.S., Y. Qiao, and D. Psaltis, *Optical network for real-time face recognition*. Applied Optics, 1993. **32**(26): p. 5026-5035.

182. Psaltis, D., et al., *Holography in artificial neural networks*. Nature, 1990. **343**(6256): p. 325-330.

183. Farhat, N.H., et al., *Optical implementation of the Hopfield model*. Applied optics, 1985. **24**(10): p. 1469-1475.

184. Gissibl, T., et al., *Two-photon direct laser writing of ultracompact multi-lens objectives*. Nature photonics, 2016. **10**(8): p. 554-560.

185. Ristok, S., et al., *Stitching-free 3D printing of millimeter-sized highly transparent spherical and aspherical optical components*. Optical Materials Express, 2020. **10**(10): p. 2370-2378.

186. Dinc, N.U., et al., *Computer generated optical volume elements by additive manufacturing*. Nanophotonics, 2020. **9**(13): p. 4173-4181.

187. Porte, X., et al., *Direct (3+ 1) D laser writing of graded-index optical elements*. Optica, 2021. **8**(10): p. 1281-1287.

188. Dinc, N.U., et al., *From 3D to 2D and back again*. Nanophotonics, 2023. **12**(5): p. 777-793.

189. Dietrich, P.-I., et al., *In situ 3D nanoprinting of free-form coupling elements for hybrid photonic integration*. Nature Photonics, 2018. **12**(4): p. 241-247.

190. Moughames, J., et al., *Three-dimensional waveguide interconnects for scalable integration of photonic neural networks*. Optica, 2020. **7**(6): p. 640-646.

191. Madrid-Wolff, J., et al., *Controlling light in scattering materials for volumetric additive manufacturing*. Advanced Science, 2022. **9**(22): p. 2105144.

192. Hunter, S., et al., *Potentials of two-photon based 3-D optical memories for high performance computing*. Applied Optics, 1990. **29**(14): p. 2058-2066.

193. Li, Q., et al., *Direct 3D-printing of microlens on single mode polarization-stable VCSEL chip for miniaturized optical spectroscopy*. Journal of Optical Microsystems, 2023. **3**(3): p. 033501.

194. Wang, T., et al., *An optical neural network using less than 1 photon per multiplication*. Nature Communications, 2022. **13**(1): p. 123.

Short Summary

Neuromorphic photonics has emerged as a platform which aims to address the growing computational demands of modern society. Here, we review recent advances in integrated synaptic optical devices and on-chip photonic neural networks and discuss challenges associated with electro-optical conversions, implementations of optical nonlinearity, amplification and processing in the time domain.

Key Points

- In this review article, photonic computing hardware is broadly categorized based on the depth at which opto-electronic conversions take place. These include (i) incoherent electro-optical processors, (ii) coherent electro-optical processors and (iii) all-optical neural networks including photonic spiking neural networks
- Looking into the future of photonic computing, we project that coherent and incoherent approaches alike have the potential to surpass state of the art electronics hardware (e.g. Google TPU v4) by two orders of magnitude in throughput, power efficiency, and compute density.
- Three-dimensional photonic components produced by 3D printing technologies are bridging free-space and in-fibre photonic neural networks with integrated photonic accelerators providing an exciting outlook into the future.
- Although the field of integrated photonic spiking neural networks is still at its infancy, SNNs are particularly well suited to photonics as photonic SNNs do not suffer from capacitive effects present in their electronic counterparts. As the individual devices are only active during a spike, SNNs offer energy-efficient computing modality.