Integrated Neuromorphic Photonics: Synapses, Neurons, and Neural Networks

Xuhan Guo, Jinlong Xiang, Yujia Zhang, and Yikai Su*

Ever-growing demands of bandwidth, computing speed, and power consumption are now accelerating the transformation of computing research, as work-at-home becomes a new normal. Brain-inspired photonic neuromorphic computing for artificial intelligence is raising an urgent need, and it promises orders-of-magnitude higher computing speed and energy efficiency compared with digital electronic counterparts. Photonic neuromorphic networks combine the efficiency of neural networks based on a non-von Neumann architecture and the benefits of photonics to constitute a new computing paradigm. Herein, some recent advances in photonic neural networks are reviewed, including the concept, principle, key photonic components, and architectures that construct the neuromorphic systems, hoping to provide a better understanding of this emerging field.

1. Introduction

Inspired by the human brain, the most important, complex, yet still mysteriously unfolded biological organ, artificial neural networks (ANNs) have enjoyed a renaissance in both research and industry since the very seminal observation of action potential spikes in neural cells in 1928.[1] It has now drastically changed our lives in almost every aspect especially in computation-consuming tasks (e.g., image recognition[2], audio processing, and deep learning[3]) and spurred many emerging application scenarios like new medical treatments, new assistive robots, and much more. Very recently, an artificial intelligence (AI) network program AlphaFold developed by DeepMind of Google almost conquered one of the grandest challenges in biology by accurately predicting a protein’s 3D shape from its amino-acid sequence.[4] It would be a brand new era for life science to understand the cells and vastly promote more advanced drug discovery and impact other research communities.

The invention of transistors[5,6] from Bell Labs in 1947 triggered the semiconductor revolution, followed by decades of information and computing domination so far from industry-led innovations, with the most successful microelectronics integration in complementary metal–oxide–semiconductor (CMOS) technology.[7] The performance and integration of micro-electronic chips have been improving continuously, until recently, when ever more signs are indicating that the transistors scaling[8] or Dennard scaling[9] that is ruled by the classic Moore’s law[10] is coming to an end.[11] This is especially true in the background of information explosion, as currently we are generating nearly 2.5 exabytes (10^{18}) of daily data in the world and it is still increasing with ongoing exponential trends,[12] which has incurred a dilemma in data exploration and processing. Under the classic Turing–von Neumann architecture for computing, the data-processing and the memory units are physically separated, and the computing process relies on the point-to-point communication methodology. Hence, the latency and power consumption are increasingly limited, such as in data-centric computation.[13] It is just becoming extraordinarily difficult to enhance the data-processing performance of electronic chips further by simply improving the integration density capability, due to a series of fundamental bottlenecks of resistor–capacitance (RC) constant, radiative physics, and the typical bandwidth–distance–energy limits in electronic links. These, however, have prompted the research community to hunt for novel information-processing frameworks to break the intrinsic limitations of the current computing architecture, such as in a massively parallel and high-throughput style yet with much less computing power. Neuromorphic AI architectures inspired by the brain are drawing increasing attention as one of the prime contenders. The human brain is a superiorly efficient information-processing ecosystem with a topological network that consists of billions of neurons connected by trillions of synapses through axons and dendrites.[14] Neurons are the primary computing elements in the brain that process information through biochemical discrete action potentials or “neuronal spikes.” Synapses act as the storage elements for memory and learning. Neurons and synapses interconnect together in the brain and form the unique spike-based temporal transferring mechanism that allows sparse and efficient information processing. Although most behaviors of the brain remain an enigma for

Prof. X. Guo, J. Xiang, Y. Zhang, Prof. Y. Su
State Key Laboratory of Advanced Optical Communication Systems and Networks
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai 200240, China
E-mail: yikaisu@sjtu.edu.cn

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adpr.202000212.

© 2021 The Authors. Advanced Photonics Research published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/adpr.202000212
human beings, there are three empirical observations from neuroscience that may lay the speculations for its remarkable capability: tremendous connectivity, structural and organizational hierarchy, and time-dependent neurosynaptic functionality.\textsuperscript{[14]} These form the unique event-driven and parallel non-von Neumann computation scheme that consumes energy only when and where the information processing occurs. The brain can perform various recognition tasks with low-frequency neuronal spiking at an ultralow-power level.\textsuperscript{[15]} For example, the power budget is only $\approx 20 \text{ W}$\textsuperscript{[16]} when performing pattern recognition, reasoning, control, or movement simultaneously, whereas the standard computer consumes power exceeding 250 W when simply undertaking the recognition task of 1000 different kinds of objects.\textsuperscript{[17]} Indeed, the brain-inspired computing or neuromorphic computing can overshadow digital computers on certain complicated computing tasks, which is believed to be a competitive way to improve the computing capacity. Actually, the concept of ANNs that mimics the brain to construct AI with electronic and photonic hardware was proposed shortly after the invention of the modern computer in the 1980s.\textsuperscript{[18]} However, as the neuromorphic processor typically requires vast interconnects or high volumes of fan-in/fan-out, the interdisciplinary implementation of electronic circuits for biological neural routines encounters physical limitations as it requires a huge multicasting workhorse with a significant processing burden. Although there are now supercomputers available to conduct millions of operations, the real-time brain-scale simulation of ANNs with the von Neumann fashion has not yet been realized and is estimated to consume at least 500 megawatts (MWs) of energy,\textsuperscript{[19]} without mentioning the hulking size. The approach of parallel computing philosophy has developed some new parallel computing technologies and multicore architectures,\textsuperscript{[20]} such as the graphical processing units (GPUs) and field-programmable gate arrays (FPGAs). These computing frameworks generally comprise some bespoke parallel processors at the digital computer front-end to specifically optimize for ANNs. Now, deep learning with ANNs,\textsuperscript{[21]} though still working with the traditional von Neumann architecture, has expanded the application scenarios such as image recognition\textsuperscript{[22]} and language translation\textsuperscript{[23]} and even overwhelmed humans at highly complex strategy games like Go\textsuperscript{[24]} and StarCraft II.\textsuperscript{[23]} The amount of computing power required to train state-of-the-art deep learning AI has been fitted to double every 3.5 months over the last 6 years.\textsuperscript{[24]} For example, AlphaGo from Google alone requires 1920 CPUs and 280 GPUs per game, consuming power at the level of MWs.\textsuperscript{[4]} Training neural networks also requires a considerable amount of computational time. For example, the residual neural networks (ResNet-200) using eight GPUs require a training time of more than 3 weeks to narrowly complete image classification tasks with error rates close to 20.7%.\textsuperscript{[2]} All these raise great challenges to the power-consuming and time-critical applications, such as supercomputers, autodriving, and “edge computing.” Although current neuromorphic electronics (e.g., IBM TrueNorth\textsuperscript{[25]} and domain-specific hardware accelerators (e.g., Google TPU\textsuperscript{[26]}), both working with non-von Neumann architectures, have significantly enhanced the energy efficiency and computing speed, the fundamental electronic RC and bandwidth–distance–energy challenges of the interconnection still pose major limitations.

Photonics, in contrast, has revolutionized the communication area. Fiber-optic communications since its birth\textsuperscript{[27]} have beaten the best electronic communication counterparts in transmission capacity over the last three decades, gradually forming the backbone of telecommunications infrastructure of the world. Distinct from the fundamental challenges in electronics, photonics suffers from less interference between different wavelengths and naturally has the advantages of large parallel operations with low latency, low power consumption, less distortion, and low jitter, offering a promising solution to address the computing drawbacks in neuromorphic hardware. The photonic waveguides can carry information at the speed of light with enormous bandwidth densities (terabits/s) that scale nearly independent of distance through high-speed signal-transmission technologies such as wavelength division multiplexing (WDM) and space division multiplexing (SDM). Records of aggregated WDM transmission capacity of 115 terabit/s over 240 km in a single-mode fiber (SMF)\textsuperscript{[28]} and 10.16 petabit/s SDM/WDM transmission using an 11.3 km 6-mode 19-core fiber\textsuperscript{[29]} have been demonstrated. The bandwidth density of the photonic chip will further increase if exploited with dense WDM (DWDM) technologies, e.g., optical frequency combs. Recently, an ultra-broadband on-chip single-soliton frequency comb spanning from 1097 to 2040 nm (126 THz) with a record-low threshold pump power of 73 μW was demonstrated in the Si$_3$N$_4$ platform.\textsuperscript{[30]} It should be noted that photonic technologies have traditionally been pursued for long-haul and high-capacity communication. Meanwhile, the most exciting part of photons in the computing field comes at the light-speed transmission with ultralow energy consumption (i.e., the associated femtojoules-per-bit\textsuperscript{[31]} energy efficiency) and ultralow latency. In other terms, the calculation process is completed when photons pass through the predefined optical structures. These unique properties lay the foundations for the superior computational performance of photonic chips. Some mathematical operations such as the differential processes, Fourier transform, and multiply-accumulate computations (MACs) have already been realized with specific photonic topological structures.\textsuperscript{[32–34]} Utilizing the high parallelism and large bandwidth of photonic devices to mimic the non-von Neumann biological neurosynaptic systems, theoretically photonic neural networks (PNNs) have the potential to surpass the state-of-the-art von Neumann electronics\textsuperscript{[35]} with at least 10 000 times faster speed, far less power consumption, and are multiple orders of magnitude faster than neuromorphic electronics.\textsuperscript{[36]} From the pioneering work of utilizing holographic materials for PNN implementation,\textsuperscript{[37]} neuromorphic photonics are booming, especially in the light of the recent progress of various optical materials\textsuperscript{[38]} and photonic integrated circuits (PICs).\textsuperscript{[39]} From the detailed computing performance, a comparison between electronic and PNNs using several experimentally verified photonic components and empirically validated network models\textsuperscript{[34]} the bandwidth of PNNs is generally above 100 GHz with the latency below 100 ps, outperforming the electronic ANNs by 2 orders of magnitude, let alone the remarkable energy efficiency on the order of pJ/MAC or even aJ/MAC.

Together with the neuroscientists, the optical research community now has been trying to develop neuromorphic photonic devices to mimic neurons and synapses to boost the prevalent computing capabilities. Utilizing the native interference
characteristics of light, the passive matrix multiplication engine has been realized in either free space or waveguide-based coherent circuitry. Nonlinear activation functions in PNNs have been implemented with creative combinations of classic photonic primitives, including microring resonators (MRRs), Mach–Zehnder interferometers (MZIs), semiconductor optical amplifiers (SOAs) and electro-optic modulators together with photodetectors. In addition, typical biological features of the spiking neuron have been successfully emulated with different types of excitable lasers, i.e., vertical-cavity surface-emitting lasers (VCSELs), micropillar lasers, distributed feedback lasers (DFBs), and quantum-dot (QDs) lasers. Meanwhile, phase-change materials (PCMs), which can be flexibly switched between multiple states of different optical properties, have emerged as a competitive alternative in hardware emulation of both the fundamental leaky integrate-and-fire (LIF) functionality of neurons and the plastic weighting process of synapses. To further obtain scalable PNNs on the mainstream silicon platform, the broadcast-and-weight (B&W) protocol has been proposed as a practical architecture to efficiently unite the well-defined neurons and synapses. However, all the optical neurons proposed so far require external active components, rendering it difficult for large-scale integrated implementations of neural networks. Although the PC-based spiking neuron is regarded as passive, still, extra laser sources with customized triggering pulse patterns are needed.

Neuromorphic photonics based on passive silicon has now become a research hotspot with a goal of the photonic neurons integrated on a single chip to form “brain-like” PNNs. Large-scale implementation and computing performance improvement for photonic chips are determined by the network complexity or integration density. Silicon photonics naturally takes advantages of both electronics and photonics, enabling the same-scale photonic integration density as microelectronics. Besides the merits mentioned earlier, the fabrication of silicon photonic devices is largely compatible with the microelectronics process using CMOS, facilitating large-scale, low-cost reproduction of photonic chips. Light sources are also indispensable in PNNs. However, silicon is not suitable for lasing due to the inherent indirect bandgap nature, silicon PICs and III–V semiconductors are normally fabricated and packaged separately, which suffer from high coupling loss and packaging cost due to the stringent alignment process between the lasers and waveguides. Fortunately, there have been growing integration/packaging technologies to tackle the difficulties, such as hybrid photonic integration and monolithic integration. Neuromorphic photonics, intersecting neuroscience and photonics, is predicted to transcend microelectronic performance by many orders of magnitude, thanks to the combination of the high efficiency of ANNs and the ubiquitous advantages of photonics. A number of neuromorphic computing concepts have been proposed, including reservoir computing, ANNs and spiking neural networks (SNNs). These promise to provide essential technological routines to support photonic neuromorphic computing. The core challenges of current AI chips are the massive and parallel data processing at fast speed with high power efficiency and low latency. The energy consumption of the electronic AI chips is proportional to the matrix dimension square $N^2$; nevertheless, the PNNs consume almost no energy in matrix processing but only in nonlinear activations. So, the larger the neural network, the more advantages PNNs will have. The comparison of deep learning computing hardware from various sources is shown in Table 1. With the functions and analogy of components in PNNs.

Figure 2a shows a schematic description of a simplified neurosynaptic system in the brain with the aforementioned fundamental elements. A mathematical model for implementing ANNs in a classic formalized artificial neuron computational element approach by McCulloch and Pitts is introduced, as shown in Figure 2b. The artificial neuron is illustrated in the following mathematical form.

$$y_j = f \left( \sum_{i=0}^{N} w_{ij} x_i \right)$$ (1)

The execution flow mainly includes two parts. First, the multiple inputs ($x$) from the other neurons are weighted ($w$) and parallelly transmitted into an artificial neuron. Then the outputs ($y$) are produced by summing up ($\Sigma$) the inputs and further trigger the other multiple neurons on axons with nonlinear activation or transfer function ($f$) such as the step function, Sigmoid function, etc.
ANNs consist of various topological architectures from the interconnections of the artificial neurons and synapses. Figure 2c shows a multilayer feedforward ANN. Once this acyclic topology is well trained by feeding enough data into the network, the information is forward processed with no feedback or loop connections. In the input layer, a collection of artificial neurons is integrated to map the real-world interface. Then one or more hidden layers are fed consecutively. Finally, the output layer provides the processed results for real-world information. The ANNs can be trained to learn a task through data sequences and subsequently optimize and tune the weight parameters using the standard backpropagation method. One typical learning process is the supervised learning method that involves weights updated through the backpropagation errors to narrow the gap between the output values and the target values, which is distinctly different from the current predesigned programming by von Neumann computer. The most typical unsupervised learning method lies in the analogy to spike-timing-dependent plasticity (STDP), which is generalized from the biological spiking nature of the neurons in neural networks as the information is encoded in the timing of the single spike.

Currently, the neuromorphic photonic approaches demonstrated are with different approaches (optoelectronic, all optical, etc.), different materials (Si, III–V, PCM, etc.), and different integration platforms (CMOS, III–V, etc.). There is a consensus among the research community to unify the gap and develop the high-performance-integrated on-chip PNNs. Therefore, great efforts have been devoted to various cointegration approaches including hybrid bonding, flipchip bonding, monolithic integration, and 2.5D or 3D integrations (i.e., interposers, through-silicon-vias). The definition of neuromorphic photonics has been widely used either with ANNs directly based on biological neurosynaptic systems or refers to photonic systems mimicking the neurosynaptic architectures with the artificial neurons. We here narrow down this Review scope to the latter by reviewing and discussing some of the most recent advances in neuromorphic computing hardware constructed with or potentially applicable to integrated photonics. This Review is neither intended to be exhaustive nor in depth but should serve as an illustration of the recent advances within the field. We focus on the key photonic neural components and give an overview of brain-like PNNs. The first part will discuss the photonic integrated synapses and the photonic integrated neurons that realize matrix multiplication and nonlinear activation, and then comes with the scalable photonic integrated neural networks to mimic the synaptic-like optical interconnections with photonic neurons. After that, we will briefly discuss the challenges of neuromorphic photonics. Finally, we will summarize the neuromorphic photonic computing with an outlook.

### Table 1. Summary of the key components of neuromorphic photonic networks.

| Components          | Functions                                           | Photonic devices                                                                 |
|---------------------|-----------------------------------------------------|----------------------------------------------------------------------------------|
| Neuron              | Summation and thresholding (nonlinear node)         | Lasers, electric-optic modulators, photodetectors, optical bistable devices, PCMs |
| Synapse             | Weighting and memory (coding scheme)                | MZIs, ring resonators, PCMs                                                     |
| Axon and dendrite   | Interconnection (network)                           | Waveguides                                                                      |

**Figure 2.** a) A schematic of the brain organizational principles. b) A mathematical model of artificial neurons and c) architecture of feedforward neural networks. Data obtained from the online image libraries.
3. Photonic Integrated Synapses

Neuromorphic computing systems, mainly consisting of neurons and synapses, can acquire powerful memory or learning capacity by modifying the synaptic weights between interconnected neurons. Various techniques that fit for on-chip integration have been utilized to realize photonic synapses including MRRs, MZI meshes, PCMs, SOAs, and so on.

The MRR bank is a natural candidate to realize tunable on-chip weighting functionality, benefited from their ubiquity of compact footprints in PICs, intrinsic WDM compatibility, and good tunability via thermal, carrier, or depletion effects.\(^{[71]}\) As shown in Figure 3a, an MRR weight bank is composed of multiple parallel-coupled microrings, each of which independently controls the transmission of exactly one WDM signal. A microring can guide the carrier signal to THRU and DROP ports by tuning on- and off-resonance, respectively, leading to complementary +/− weighting.\(^{[72]}\) However, the high sensitivity of MRRs to fabrication errors, thermal fluctuations, and thermal crosstalk poses great challenges to the control stability. Previously, great efforts have been devoted to this research field.

Tunable methods can be generally categorized into two types. 1) Feedforward control based on calibrations of concerned parameters. A single photonic weight with a continuous range from −1 to +1 and precision of 3.1 bits has been experimentally demonstrated using a 5 GHz signal.\(^{[73]}\) However, it becomes impractical for accurate weight control of an increasing channel number of \(N\) as \(O(N^2)\) calibration measurements are required in the presence of interchannel dependence. To deal with this problem, scalable calibration models of thermal crosstalk and amplifier cross-gain saturation have been developed.\(^{[73]}\) In addition, a weight accuracy of 3.8 bits has been further obtained on a four-channel MRR weight bank with the feedback control approach. It is worth mentioning that a suite of external equipment including the optical spectrum analyzer, oscilloscope, and pattern generator is necessary for these feedforward calibrations. In addition, the control technique is sensitive to temperature variations after phase calibration, which casts shadows on the feasibility of the feedback control approach. 2) Feedback control based on real-time monitoring of the related parameters, which can adapt automatically to the environment changes and provide a simplified calibration and modeling requirement. The in-ring \(N\)-doped photoconductive heaters have been used for continuous, multichannel weight control of an MRR weight bank.\(^{[74]}\) As shown in Figure 3b, the microring waveguide is lightly doped with donor carriers, thus forming an in-ring resistor.

**Figure 3.** a) A WDM weight bank implemented with MRRs. b) Photonic weight bank consisting of two \(N\)-doped MRRs. c) A simplified feedback control model starting from desired weight and ending at applied current. d) A binary search algorithm for a command transmission value in the feedback control rule. e) Layouts of an MRR add-drop demultiplexer (left) and an MRR weight bank (right). f) Coherent paths in MRR weight bank with 1-pole filters (left) and 2-pole filters (right). (a) Reproduced with permission.\(^{[73]}\) Copyright 2016, Optical Society of America. (b–d) Reproduced with permission.\(^{[74]}\) Copyright 2018, Optical Society of America. (e–f) Reproduced with permission.\(^{[78]}\) Copyright 2016, IEEE.
Applying current will heat up the MRR, resulting in a resonance peak shift. Meanwhile, electron–hole pairs will be generated due to the absorption of a portion of circulating optical power, which, in turn, increases the conductivity of the heater. Therefore, the amount of light circulating in the MRR can be detected by applying current and measuring the voltage (or vice versa). A simplified feedback control model is schematically shown in Figure 3c, where \( D \) and \( \hat{D} \) are actual and estimated DROP port transmissions, respectively. \( \delta \) and \( P \) are normalized and nominal electrical power, respectively. It begins with the target weights and ends with the applied currents. In the feedback control rule shown in Figure 3d, the controller performs a binary search algorithm for a target transmission value, and the commanded value for the thermoelectric control step is determined by the converged value of \( \delta \). Utilizing this feedback control procedure, two-channel photonic principal component analysis (PCA) and independent component analysis (ICA) have been experimentally demonstrated with on-chip MRR weight banks, which can extract the principal components and recover the independent components solely based on the statistical information of the weighted addition output.[75,76] Furthermore, by simplifying the transmission edge calibration procedure, a record high accuracy of 6.6 dB and precision of 6.5 dB with negligible interchannel crosstalk are achieved for the four-channel weight bank control,[77] significantly outperforming the digital weight resolution (5 bits) used in the TrueNorth architecture.[25] Apart from the stable control of the MRR weight bank, channel density is another important factor for large-scale networks. Conventional analyses of the MRR devices for (de)multiplexing and modulating WDM signals have predicted a maximum channel count of \( \approx 36 \),[55] where the channel spacing is mainly limited by interchannel crosstalk (Xtalk). As shown in Figure 3e, each WDM channel is coupled to one distinct bus waveguide. Xtalk will be induced when a wavelength channel couples partially with an adjacent MRR filter and exits the incorrect DROP port, which sets a limitation on the minimum channel spacing. However in an MRR weight bank, the MRRs are parallelly coupled with exactly two same bus waveguides and the outputs are still multiplexed, resulting in the breakdown of crosstalk between neighboring channels and giving rise to the multiresonator coherent interaction. Figure 3f shows the optical coherent paths for an MRR weight bank: resonator like for 1-pole configuration and interferometer like for 2-pole configuration. Taking the coherent phenomena into consideration, an original simulation technique which combines parametric programming with generalized transmission theory is introduced, indicating a channel count near 108.[78] Further, the coherent interchannel effects in 1-pole and 2-pole MRR weight bank have been experimentally explored.[79]

Utilizing the native interference characteristics of light, MZI meshes have been developed to implement totally passive optical matrix multiplication units for PNNs. So far, three architectures, made up from meshes of \( 2 \times 2 \) interferometers, have been proposed to realize arbitrary unitary transforms between an input vector and a corresponding output vector for coherent light at a given wavelength: a “triangular mesh” architecture (see Figure 4a),[80–84] a “cascaded binary tree” architecture.
(see Figure 4b)\cite{82}, and a “rectangular mesh” architecture (see Figure 4c)\cite{83}. Among them, both “triangular mesh” and “cascaded binary tree” architectures can be automatically configured using “training” vectors of inputs and simple progressive algorithms based on detection and simple one- or two-parameter feedback minimization processes.\cite{82,83} To further realize arbitrary nonunitary transforms, two architectures have been reported: an architecture based on the singular value decomposition (SVD) of the desired matrix\cite{84} and one based on the use of a \(2N \times 2N\) unitary matrix to implement an \(N \times N\) nonunitary transform by operator dilation.\cite{85} It is worth mentioning that the SVD approach is trainable and requires minimum number of phase shifters. In practice, it can be implemented with two unitary transforms illustrated earlier together with an additional row of modulators. Detailed information about principles, architectures, algorithms, development routines, and applications of MZI meshes can be found in these reviews.\cite{86,87}

A fully hardware implementation of the SVD architecture is schematically shown in Figure 4d, where Unit (2) and Unit (4) can implement an arbitrary unitary matrix transformation individually, whereas Unit (3) can implement an arbitrary diagonal matrix transformation. The four MZIs in Unit (1) are used to tune the input power. The programmable photonic processor can perform fundamental matrix computation including \(XB = C\), \(AB = X\), and \(AX = C\) where \(A, B, C\) are known matrices, and \(X\) is the matrix to be solved. Moreover, it can be trained to demonstrate the optical PageRank algorithm.\cite{88}

PCM can be reversibly switched between its crystalline state and amorphous state by electrical or optical excitations.\cite{89} Dramatically high contrasts in the optical and electrical properties can be exhibited between these two states from the visible to infrared spectrum region, which may find applications in optical switching, memory, and calculating. In addition, intermediate states between the two phases can be obtained by carefully controlling the excitation signals, resulting in multilevel operations.

An innovative scheme to implement photonic synapses with Ge\textsubscript{2}Sb\textsubscript{2}Te\textsubscript{5} (GST) has been reported in 2017.\cite{90} As shown in Figure 5a–c, discrete GST islands deposited with ITO films are patterned on a tapered waveguide structure. The device is crystallized completely beforehand, and the initial optical transmission \((T_0)\) is defined as the baseline of the readout and assigned to a synaptic weight “0.” In addition, the relative change of optical readout \((\Delta T = T - T_0)\) in percentage to the baseline \((\Delta T/T_0)\) is regarded as the change of synaptic weight (see Figure 5d), which can be achieved by sending optical pulses with a fixed duration and fixed energy in the waveguide. More importantly, arbitrary synaptic weight levels can be achieved with a series of known pulses without any prior knowledge of the current actual weight. For instance, weight “0,” “1,” “2,” “3” in Figure 5d can be switched from any previous weight with 1000, 100, 50, and 1 pulses, respectively (each pulse is of 243 pJ with 50 ns length). In addition, the repeatability of the weighting process is examined over multiple switching cycles in Figure 5e. Further investigations prove that the synaptic weight change depends on the pulse number exponentially and monotonically, as shown in Figure 5f. Moreover, all-optical STDP can be simply implemented with photonic integrated-circuit techniques. As shown in Figure 5g, the presynaptic signal is divided into two equal parts with one-half coupled into a photonic synapse and the other half \((P_{in})\) connected to an interferometer via a phase modulator. Similarly, the postsynaptic signal is also separated into two beams, with one-half transmitted and another half \((P_{in})\) fed back to the interferometer. The output power of the interferometer \((P_{out})\) can be tuned between 0 and \(P_{in1} + P_{in2}\) by modifying the phase modulator, which is utilized to update the synaptic weight. The power of pre- and postsynaptic signals is set between \(P_{th}\) and 2 \(\times P_{th}\), where \(P_{th}\) is the threshold optical power to switch GST. By intentionally setting the pulse widths and repetition rates of pre- and postsynaptic signals differently, there is only a single pulse larger than \(P_{th}\) in the net output power, when there is no time delay between pre- and postsynaptic spikes \((\Delta t = 0)\) (see Figure 5h). The number of pulses with net power above \(P_{th}\) can be progressively increased to 2, 3, 4, 5 by increasing the time delay (see Figure 5i–l). By properly engineering the pre- and postsynaptic signals, the number of switching pulses shows a linear dependence on \(\Delta t\), which further leads to the required exponential dependence of the weight change on \(\Delta t\). Consequently, STDP can be emulated simply and effectively.

Another route toward photonic synapses has been reported with a PCM element embedded on one arm of a single-bus microring waveguide,\cite{90} schematically shown in Figure 5m. The output power can be expressed as the product of the transmission at the resonant wavelength and the input power. Therefore, the synaptic weight can be determined by the transmission at the resonant wavelength, which can be flexibly tailored by varying the crystallization degree of GST. Furthermore, a photonic dot-product engine can be implemented between the input spikes and synaptic weights by leveraging WDM technology. As shown in Figure 5n, multiple microrings with increasing radius are arranged in a row to represent different synapses. The amplitude of input WDM spikes will be selectively modulated by the GST element on each microring; thus, a multiwavelength spike comprising different \(T_{ij}P_{j}\) products can be obtained at the output port. Finally, the output spike is fed into a photodiode array, generating a current determined by the sum of all the amplitudes, which is exactly the dot-product of the input vector and weight vector.

In addition, a novel matrix operation scheme has been proposed with a Kerr frequency microcomb source by simultaneously combining wavelength and time multiplexing.\cite{91} As shown in Figure 6, each wavelength generated from the microcomb source functions as a single synapse and its weight is assigned by adapting its output power with an optical spectral shaper. The raw input image is preprocessed and flattened into a 1D vector, which is then sequentially multiplexed in the time domain via a high-speed electrical digital-to-analog converter. Next, the electrical waveform is multicast on all wavelength channels simultaneously with an electro-optic modulator, thus generating an identical replica of the temporal input signal on each comb line. Then, progressive time delays are introduced to the weighted replicas with a dispersive fiber for the purpose of lining up all the diagonal elements into the same timeslot. Finally, the weighted and summed output is obtained by detecting the optical intensity of the aligned timeslot. In the proof-of-concept experiment, a high single-unit throughput speed of 95.2 Gbps is achieved by mapping the synapses of a single perceptron onto 49 wavelengths of the microcomb.
Some other techniques have been proposed to implement the weight addition functionality. As shown in Figure 7a, arrayed waveguide gratings (AWGs) combined with SOA technologies are utilized to realize cross connectivity and WDM operations. The SOAs are utilized to set the weight matrix and provide the on-chip gain for scalability, whereas the AWGs are implemented for filtering out the out-of-band noise accumulated by multiple cascaded stages of SOAs, as well as for demultiplexing the input data channels. On-chip amplitude-based weighted addition operation of four input channels per neuron for a total number of...
Figure 6. Principle of microcomb-based photonic matrix operations. Reproduced with permission.\(^{[91]}\) Copyright 2020, Wiley-VCH.

Figure 7. a) SOAs-based weight addition processor with eight WDM inputs and outputs. b) Composite microscope image of the fabricated PIC. Layout of the: c) dual-IQ modulator, d) \(n = 2^m\) optical linear algebra unit (OLAU) circuit, e) \(n = 2^{m+1}\) OLAU circuit, and f) \(n = 2^m\) OLAU along with the bias branch, forming a coherent optical linear neuron. (a,b) Reproduced with permission.\(^{[92]}\) Copyright 2020, IEEE. (c–f) Reproduced with permission.\(^{[93]}\) Copyright 2020, IEEE.
In addition, a nested intermodulator to determine a continuous set of real-valued outputs. The third generation of neurons referred to as spiking neurons, functions as a LIF unit and exhibits all-or-none outputs in response to input spikes. Actually, modern deep learning networks are all based on the second generation of neural networks,[97] and current photonic implementations of ANNs also only fall into the last two categories. To better review and clarify the activation differences between various photonic neurons, we here adopt a commonly acceptable convention of the ANN definition to the second generation of neural networks in the following part. With this definition, the ANN discussed here refers to a more “old fashioned” activation scheme (the sigmoid and ReLu). Then, SNN with spiking neuron behaviors will be discussed afterward.

The past decade has witnessed the phenomenal success in the deep learning field utilizing ANNs based on analog information processing. However, the training and inferencing of such multilayer neural networks will become extremely expensive in terms of energy and memory with the explosive growth of computational burden and complexity. SNNs exploring the spiking behaviors in a more biologically computational fashion have leapfrogged as a promising alternative that bridges the gap between neuroscience and machine learning. Due to the event-driven nature and sparse-information encoding, SNNs are intrinsically more viable for energy-efficient neuromorphic computing. The following sections will discuss the hardware implementations of photonic neurons in detail.

4. Photonic Integrated Neurons

The research of ANNs has gone through multiple phases of evolution over long-time development and can be normally categorized into three generations in terms of their neural information-processing mechanisms.[94] The first generation of neurons conducts a thresholding operation with binary-valued outputs, which returns 1 when the potential exceeds a fixed threshold and 0 otherwise. The second generation of neurons works with continuous nonlinearity using a smooth mathematically defined activation function, typically a logistic sigmoid unit,[95] or a rectified linear unit (ReLU).[96] to determine a continuous set of real-valued outputs. The third generation of neurons, referred

Table 2. Recent advances and key performance of typical photonic synapses.

| Type                        | Precision | Footprints | Power | CMOS compatibility |
|-----------------------------|-----------|------------|-------|--------------------|
| MRR[92]                     | >7 bits   | Medium     | Medium| Yes                |
| MZI[82]                     | 16 bits   | Large      | High  | Yes                |
| PCMs[54-73]                 | >5 bits   | Small      | Low   | No                 |
| Microcomb + modulator[31]   | 8 bits    | Large      | High  | Yes*               |
| SOA[92]                     | 6 bits    | Large      | High  | No                 |
| Dual-IQ modulator[93]       | N.A.      | Large      | High  | Yes                |

*aThis discrete scheme can be potentially integrated on the chip.

four neurons per layer is experimentally demonstrated (see Figure 7b), which is further used to build a multilayer neural network for image classification.[92] In addition, a nested interferometric arrangement of dual-IQ modulation cell structures has been proposed to realize weighting addition at a single wavelength, supporting both positive and negative value encoding through the electric field phase information.[93] The architecture is schematically shown in Figure 7c–f. The dual-IQ modulator cell consists of a typical IQ modulator structure, where each interferometric branch contains both the I- and Q-modulation stage, thus virtually merging the two IQ modulators into the same MZI structure. As a result, each MZI arm functions as a complete photonic synapse where the I-modulation stage realizes the electrical-to-optical conversion of the input signal, whereas the Q-modulation stage implements the weighting functionality. With this configuration, the amplitude weighting and weight sign encoding processes are separated.

Table 2 shows the recent typical types of photonic synapses. The MRR weight bank and MZI meshes with dual-IQ modulator take great advantages with good compatibility with the standard CMOS fabrication process. Compared with the MZI-based configurations, MRR weight banks have compact footprints but sacrifice the precision due to their high sensitivity to thermal crosstalk. Beneficial from the nonvolatile property of GST, the PCM-based weighting scheme exhibits the lowest power consumption. However, its scalability is greatly limited by the complicated deposition process. Regarding the SOA-implemented weight addition operation, monolithic integration or cointegration processes are needed within a passive chip.

4. Photonic Integrated Neurons

The research of ANNs has gone through multiple phases of evolution over long-time development and can be normally categorized into three generations in terms of their neural information-processing mechanisms.[94] The first generation of neurons conducts a thresholding operation with binary-valued outputs, which returns 1 when the potential exceeds a fixed threshold and 0 otherwise. The second generation of neurons works with continuous nonlinearity using a smooth mathematically defined activation function, typically a logistic sigmoid unit,[95] or a rectified linear unit (ReLU).[96] to determine a continuous set of real-valued outputs. The third generation of neurons, referred
saturated differentially biased regime, whereas the followed SOA operates in its small-signal gain region and functions as a cross-gain modulation-wavelength converter (XGM-WC). The nonlinear transfer function can be modified with the control attenuation factor and the bias attenuation factor. Figure 8c shows an O/E/O modulator neuron fabricated with a conventional silicon photonic process. The optical neuron is composed of two photodetectors electrically connected to an MRR modulator, which subtracts the photocurrents from two inputs and remodulates a signal on a new wavelength. Depending on the wavelength offset between the pump signal and the microring resonance, the modulator neuron can exhibit a variety of nonlinear transfer functions, including the sigmoid, ReLu, and radial basis functions (RBFs). In addition, other important properties to act as a network-compatible neuron-like fan-in, inhibition, time-resolved pulse processing and cascadability have also been demonstrated. A monolithically integrated SOA-based photonic neuron has been experimentally demonstrated, shown in Figure 8d, where gain variations of multiple SOAs are used to weight the input signals and the activation function is implemented with an SOA-based WC (SOA-WC). However, the SOA-WC can only provide an inverted signal at its output. Via the free-carrier dispersion (FCD) effect, an all-optical neuron is implemented with an MRR-loaded MZI structure, as shown in Figure 8e. By varying the coupling ratio of the Mach–Zehnder coupler (MZC) and the wavelength detuning between the input signal and the MRR resonance, the optical neuron can be programmed to exhibit the sigmoid, radial basis, clamped ReLu, and softplus response with tunable thresholds. A novel on-chip electro-optic circuit is reported to realize arbitrary nonlinear activation function. As shown in Figure 8f, a small portion of the input optical signal is converted into an electrical signal, which then can modulate the intensity of the remaining optical signal. Actually, only a limited range of activation functions can be realized with the electrical signal directly used to modulate the remaining optical signal. Implementing a look-up table in a microcontroller can map the photogenerated current to any desired modulating voltage, thus realizing arbitrary nonlinear activation functions. However, the usage of a microcontroller will limit the operation speed of the optical neuron to sub-GHz range.

Recent advances and the key performance for different types of photonic neurons in ANNs are shown in Table 3. The MRR

![Figure 8. Optical implementations of nonlinear activation function in ANNs. a) Theoretical model for electro-optic absorption modulators. b) An all-optical neuron with a sigmoid transfer function based on SOA-MZI. c) An InP photonic integrated neuron implemented with SOA-WC. d) A microring modulator neuron. e) MRR-loaded MZI structure. f) An electro-optic circuit to obtain arbitrary optical nonlinear activation function. (a) Reproduced with permission. (b) Reproduced with permission. (c) Reproduced with permission. (d) Reproduced with permission. (e) Reproduced with permission. (f) Reproduced with permission.](https://www.advancedsciencenews.com/adv-photonics-research)
Table 3. Recent advances and key performance of photonic neurons in ANNs.

| Type                  | Power  | Speed | Footprints | Reconfigurability | CMOS compatibility |
|-----------------------|--------|-------|------------|-------------------|--------------------|
| SOA-MZI               | mW     | sub-ns| Large      | No                | No                 |
| SOA-WC                | mW     | ns    | Large      | No                | No                 |
| MRR modulator         | mW     | ns    | Small      | Yes               | Yes                |
| MRR-loaded MZI        | mw     | ns    | Medium     | Yes               | Yes                |
| Electro-optic modulator| mw    | ns    | Large      | Yes               | No                 |

modulator and MRR-loaded MZI show great competitiveness with fast operation speed on the order of GHz, reconfigurable activation functions, and good CMOS compatibility. In contrast, the electro-optic modulator is more complicated with optic-to-electro (O/E) conversion circuits.

4.2. Integrated Neurons in SNNs

In SNNs, information is not contained in the shape or amplitude of spikes, but in their timing of presence or absence, which can overcome the noise accumulation problem inherent in purely analog computation. From the perspective of computability and complexity theory, the leaky LIF neuron can work as a powerful computational primitive to simulate both Turing machines and traditional neural networks. The standard LIF neurons can be treated as electrical devices and typically characterized with the following equations:

$$C_m \frac{dV_m(t)}{dt} = - \frac{1}{R_m} (V_m(t) - V_L) + I_{app}(t)$$  \hspace{1cm} (2)

if $V_m(t) > V_{\text{thresh}}$, then release a spike and set $V_m(t) \rightarrow V_{\text{reset}}$  \hspace{1cm} (3)

where $V_m(t)$ is membrane potential, $C_m$ and $R_m$ are the membrane capacitance and resistance respectively, $V_L$ is the resting potential, and $I_{app}$ is the input electrical current.

The LIF neuron has multiple unique properties. 1) Excitability threshold: The LIF neuron can only fire an output spike when its potential exceeds a certain threshold. Moreover, there usually exists a steep transition in the response when below and above the threshold. 2) Leaky integrating property: The LIF neuron can be excited by integrating several closely spaced sub-threshold inputs, which can be used to realize the temporal spike pattern recognition task. 3) Refractory period: There is a time period immediately following an excitation, during which it is impossible to excite another spike. This is the well-known absolute refractory period. After this period, the neuron usually does not recover completely and its excitability threshold will increase. As a result, it will be more difficult to excite the neuron. This is called the relative refractory period. It is noted that the refractory period directly determines how fast the neural network can work. 4) Cascadability. The output spike of a neuron should be strong enough to excite neurons in the next stage, which guarantees the successful communication between interconnected neurons and is necessary to form a multilayer SNN. 5) Inhibition. The arrival of an inhibitory stimulus can prevent the neuron from spiking activity. Actually, inhibition together with excitation lays the foundation of STDP synaptic learning, which determines the synaptic weight modification in reliance on provisional correlations between the spikes of pre- and postsynaptic neurons.

There has been a long history of human exploration in appropriate photonic spiking computational primitives. The early attempts mainly focus on discrete nonlinear fiber components, which can hardly be scaled to a few neurons due to the complex circuits, high power consumption, and large footprints. Although some biological spiking features have been demonstrated with the excitability in semiconductor lasers close to $Z_2$-symmetry, and the polarization switching (PS) effect in VCSELs, it was not until 2013 that a general excursion laser model capable of implementing cortical-inspired algorithms was proposed to behave analogously to an LIF neuron. As shown in Figure 9a, the two-section excitable laser consists of a gain section with a saturable absorber (SA) and mirrors for cavity feedback. The inputs perturb the gain section selectively. Serving as a temporal integrator, the time constant of the gain medium equals the carrier recombination lifetime. However, an SA becomes transparent with increasing optical intensity, thus acting as a thresholder by gating the light intensity accumulated in the cavity with the gain medium (see Figure 9b). Starting from the Yamada model with some reasonable assumptions like

Figure 9. a) A simple schematic of a two-section SA laser. b) Simulation of an SA excitable laser as an LIF neuron. (a,b) Reproduced with permission. Copyright 2013, IEEE.
the SA that yields a fleeting relaxation time on the order of the cavity intensity, the following instantaneous pulse-generation model can be obtained by compressing the internal dynamics in this system.\[102\]

\[
\frac{dG(t)}{dt} = -\gamma_C(G(t) - A) + \theta(t) \quad (4)
\]

if \( G(t) > G_{\text{thresh}} \), then release a pulse and set \( G(t) \rightarrow G_{\text{reset}} \) \( (5) \)

where \( G(t) \) models the gain, \( \gamma_C \) is the relaxation rate of the gain, \( A \) is the bias current of the gain, \( \theta(t) \) is the input term, \( G_{\text{thresh}} \) is the gain threshold, and \( G_{\text{reset}} \approx 0 \) is the gain at transparency. The conditional statements result in the fast dynamics of the system that occur on time scales of order \( 1/\gamma_I \) (\( \gamma_I \) is the inverse photon lifetime) and assure that \( G_{\text{thresh}}, G_{\text{reset}}, \) and the pulse amplitude remain constant. It is obvious that there exist great similarities between this model and the typical LIF model. For instance, the gain of the excitable laser \( G(t) \) can be regarded as a virtual membrane potential. Since then, various biological properties and learning tasks of spiking neurons have been demonstrated experimentally and numerically. Fiber-based graphene excitable lasers (GELs)\[8,112,113\] micropillar lasers with integrated SAs\[48,114–116\] DFBs\[117,118\] and VCSEL-SAs\[102,119–122\] all fall under this category. In addition, it is worth mentioning that a circuit model has been proposed by mapping the rate equations of excitable lasers with an embedded SA to an equivalent circuit; therefore, the signal-processing behaviors such as excitation and inhibition can be efficiently and accurately simulated with the SPICE engine.\[123\]

In 2016, a fiber-based GEL (Figure 10a) and an integrated GEL (Figure 10b) on a hybrid silicon III–V platform were

![Figure 10](image-url)

**Figure 10.** a) A fiber-based GEL. b) Structure of an integrated GEL. Experimental results of c) spiking dynamics including leaky integrating, refractory period, excitability threshold, and pulse energy encoding. d) All-optical DTS conversion. e) Simultaneous excitation and inhibition. (a–c) Reproduced with permission.\[8\] Copyright 2016, Springer Nature. (d) Reproduced with permission.\[112\] Copyright 2017, Optical Society of America. (e) Reproduced with permission.\[113\] Copyright 2018, Optical Society of America.
experimentally demonstrated and theoretically proposed, respectively.\textsuperscript{[8]} The integrated GEL consists of electrically pumped QWs (gain section), two sheets of graphene (SA section), and DFB grating (section). The integrated device can be characterized with the same theoretical model as the fiber prototype, thus exhibiting the same behaviors with lower pulse energies on a much hasty timescale. The fiber-based GEL can demonstrate the typical features of biological neurons including excitability threshold, leaky integrating, and refractory period (see Figure 10c).\textsuperscript{[8]} Moreover, the cascadability and pulse generation are presented with a self-referenced connection. Based on this computational prototype, an all-optical digital-to-spike (DTS) conversion scheme has been proposed without clock signal synchronization.\textsuperscript{[112]} As shown in Figure 10d, a conversion rate of 40 Kbps and a bit error rate (BER) of $10^{-9}$ were achieved with an on-off keying input power of $-24$ dBm in the proof-of-concept experiment. Further extended work by experimentally presenting simultaneous excitation and inhabitation in a fiber-based GEL was demonstrated in 2018.\textsuperscript{[113]} As shown in Figure 10e, a single excitatory input can activate the GEL. The output spike will be gradually suppressed to null as the inhibitory stimulus moves toward the excitatory stimulus. In addition, the output spike will recover quickly as the inhibitory input moves away from the excitatory input. The micropillar laser has been experimentally confirmed to demonstrate similar spiking properties with GELs by responding to perturbations with subnanosecond spikes and exhibit both absolute and relative refractory periods.\textsuperscript{[114]} Also, the nonlinear timing and response properties of a micropillar laser under both coherent and incoherent perturbations are analyzed in detail. The spike latency of the laser has been shown to provide temporal encoding.\textsuperscript{[115]} Moreover, the leaky integrating capacity of a micropillar laser has been verified by studying its response to consecutive subthreshold stimuli.\textsuperscript{[116]} Furthermore, it has been proved that a chain of coupled micropillar lasers has the computing ability to implement spike-based logic gates and temporal pattern recognition.\textsuperscript{[48]}

In 2015, a two-section DFB excitable laser neuron was theoretically proposed in a hybrid III–V/silicon platform.\textsuperscript{[117]} More recently, an integrated photonic spiking processor directly implementing the analog O/E/O link was demonstrated.\textsuperscript{[49]} As shown in Figure 11a, the integrated chip mainly consists of nine two-section DFB laser neurons, pairs of high-speed balanced photodetectors (BPDs), together with connecting metal wires. Each DFB laser is composed of a large gain section and a small absorber section, which are optically coupled but electrically isolated. The BPDs are only connected to the gain section. In addition, the PD1/PD2 functions as an inhibitory/excitatory synapse with the generated current flowing out/into the gain section. Apart from demonstrations of the fundamental spiking dynamics including the excitability threshold (Figure 11b), refractory period (Figure 11c), and leaky integrating (Figure 11d), XOR classification is successfully simulated with a configuration compatible with the current hardware architecture.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure11.png}
\caption{a) The micrograph of a photonic laser neuron. b) Simulated (green) and measured (red) response of a spiking laser neuron to input pulses with increasing width. c) Response of the refractory period to successive input pulses as a function of their separation. d) Response to two pulses with varying center-to-center separation. (a–d) Reproduced with permission.\textsuperscript{[49]} Copyright 2020, IEEE.}
\end{figure}
The first compact VCSEL-SA that could mimic the spiking dynamics of a LIF neuron has been reported as early as 2013.\(^{[102]}\) Furthermore, VCSEL-SAs in combination with VCSOAs have been numerically investigated and shown their potential in different spiking-processing applications including spike encoding, spike memory, STDP learning mechanism, and pattern recognition with both supervised and unsupervised learning.\(^{[119–122]}\)

The PS effect and nonlinear dynamics induced by polarized optical injection of VCSELs have been widely explored to emulate the basic functionalities of spiking neurons. The operational principles can be explained as follows.\(^{[110]}\) As schematically shown in Figure 12a,b, the VCSEL’s state in a certain biasing condition is determined by two orthogonal polarizations of the fundamental transverse mode. Without optical injection, the parallel polarization is solely conductive to the lasing mode, whereas the orthogonally polarized mode is suppressed. When applied to powerful enough external optical injection with orthogonal polarization, the lasing mode of the VCSEL will be switched from parallel polarization to orthogonal polarization, as shown in Figure 12c. Therefore, the orthogonal polarization mode is activated, which is equivalent to the activation of spiking neurons. In addition, the VCSEL can be switched back to its initial state with a subsequent input at the other polarization (Figure 12d). In a more practical configuration, the VCSEL is usually biased with a single optical input, and perturbation is induced in the form of power drops. In this situation, the initial input power level is adequately enough for inducing stable injection locking in either of the two polarization modes. The arrival of perturbations will trigger the VCSEL out of its locking range, thus firing different spike patterns that are dependent on the original conditions and the perturbation characteristics. Using this scheme, an all-optical DTS format conversion has been realized in a 1310 nm VCSEL.\(^{[124]}\) A fully controllable and reproducible inhibition response was reported in 2017.\(^{[125]}\) In addition, the controllable propagation of spiking patterns between two interconnected VCSELs has been experimentally demonstrated,\(^{[126]}\) where the receiver VCSEL responded the same as the transmitter VCSEL with the equal number of spikes, same-spike and inter-spike temporal durations, and similar spike intensity properties. Moreover, it has been proved that perturbations encoded in the applied bias current can excite the VCSEL to fire controllable spiking patterns.\(^{[127]}\) Utilizing a biasing tee, the perturbation current signal can be either an electrical or an optical (wavelength-independent)

**Figure 12.** a,b) Optical spectrum of the 1550 nm VCSEL with and without orthogonally polarized optical injection producing PS and injection locking. Operation principle of the VCSEL neuron under c) an excitatory stimulus (orthogonally polarized signal) and d) an inhibitory stimulus (parallel polarized signal). (a,b) Reproduced with permission.\(^{[110]}\) Copyright 2011, American Physical Society. (c,d) Reproduced with permission.\(^{[110]}\) Copyright 2010, Optical Society of America.
signal with varying amplitude. Recently, multiple photonic processing tasks including spiking memory module, emulation of neuronal circuits in the retina, and pattern recognition were experimentally presented with VCSEL spiking neurons.\textsuperscript{128,129} In addition, there have been multiple interesting numerical investigations on different interconnectivity architectures, learning algorithms, and network frameworks based on the PS operation of VCSEL neurons.\textsuperscript{47,130–136}

PCM can not only implement the plastic weighting operation of synapses but also the LIF functionality of spiking neurons. The spiking neuron prototype based on GST-embedded MRR resonators (Figure 13a,b) was first proposed in 2018.\textsuperscript{137} Following this idea, an all-optical spiking neurosynaptic system has been experimentally presented, which successfully demonstrates pattern recognition tasks directly in the optical domain with both unsupervised and supervised learning.\textsuperscript{54} As shown in Figure 13c, the resonance condition and propagation loss of the MRR are influenced by the state of the GST element. When the GST cell is initially in the crystalline state, the probe pulse sent along the output waveguide will be strongly coupled into the MRR. Consequently, no spike can be observed at the output port. However, when the combined membrane potential of the weighted inputs from presynaptic neurons is high enough to cross the threshold, the GST cell will be switched to its amorphous state. In this case, the propulse will be no longer on-resonance with the MRR and an output spike will be generated. Figure 13d shows the activation function of the spiking neuron obtained by measuring the output transmission in responding to different excitation pulses at a fixed wavelength.

Although impressive and encouraging results have been reported, there is still a long way to go before the implementation of a fully integrated neuromorphic system with PCM. Currently, extended lasers are used to produce output pulses. In addition, to update the synaptic weights in unsupervised learning, the output pulses are amplified off-chip in the feedback loops. More importantly, reset pulses have to be sent manually after each activation to switch the GST back to its initial crystalline state, which sets great limitations on large-scale integration.

There have been some exciting researches on exploiting the excitability of passive microcavities to emulate the spiking dynamics of biological neurons.\textsuperscript{103,138} There are various nonlinear effects, e.g., free-carrier absorption (FCA), FCD, thermo-optic (TO) effect, interacting with each other in a microcavity. Among them, the FCD effect leads to a blue shift in the resonance wavelength, whereas the TO effect induces a red shift in the resonance. Typically, the heat relaxes at least one order magnitude slower than the free carriers. As a result, excitability can be observed in microcavities due to the difference between the fast free-carrier dynamics and the slow heating effects, as schematically shown in Figure 14a. The nonlinear dynamics of a passive side-coupled microresonator can be characterized by a set of coupled-mode theory (CMT) equations:\textsuperscript{103}

\[
\frac{da_\pm}{dt} = j(\omega_+ - \omega_-) - \frac{1}{2\tau_{\pm,\text{total}}} a_\pm + \kappa_\pm S_\pm
\]

\[
\frac{dN}{dt} = -\frac{N}{\tau_{\text{ec}}} + \frac{\Gamma_{\text{FCA}} \beta_{\text{FCA}} c^2}{2\hbar \omega V^2_{\text{FCA}}} \left( |a_+|^4 + 4|a_+|^2|a_-|^2 + |a_-|^4 \right)
\]

Figure 13. a) A bipolar LIF neuron based on GST-embedded MRR devices containing both the integration and firing units. b) Timing diagram showing the integration of the membrane. c) Scanning electron micrograph of a PCM-based photonic neuron. d) Normalized transmission to the output at a fixed wavelength. (a,b) Reproduced with permission.\textsuperscript{137} Copyright 2018, Springer Nature. (c,d) Reproduced with permission.\textsuperscript{54} Copyright 2019, Springer Nature.
\[
\frac{d\Delta T}{dr} = -\frac{\Delta T}{\tau_{\text{th}}} + \frac{\Gamma_{\text{th}} P_{\text{abs}}}{\rho_{\text{Si}} C_{\text{p, Si}} V_{\text{cavity}}} \tag{8}
\]

where \(a_x\) is the complex amplitude of the forward and backward propagation mode, respectively, and \(|a_x|^2\) stands for the corresponding mode energy in the microresonator. \(S_x\) represents the complex amplitude of the pump light and the perturbation signal injected in the opposite direction. \(\omega_x\) is the frequency of input light in the waveguide, and \(\omega'\) denotes the shifted resonance frequency of the microresonator. \(\kappa_x\) is the coupling coefficient between the waveguide and the microresonator. \(N\) is the concentration of free carriers in the microresonator and \(\Delta T\) is the mode-averaged temperature difference with the surroundings. \(\tau_{\text{th}}\) is the relaxation time for temperature and \(\tau_{\text{fc}}\) is the effective free-carrier decay rate accounting for both recombination and diffusion. \(\beta_{\text{Si}}\) is the constant governing TPA, \(\rho_{\text{Si}}\) is the density of silicon, \(C_{\text{p, Si}}\) is the thermal capacity of silicon, and \(V_{\text{cavity}}\) is the volume of the microresonator. \(n_g\) is the group index. \(V_{\alpha}\) and \(\Gamma_{\alpha}\) denote the effective mode volume and confinement coefficient.

As shown in Figure 14b–d, the excitability, refractory period, and cascadability of MRRs have been experimentally demonstrated.\(^\text{138}\) However, leaky integrating, inhibitory properties of nanobeams are further numerically simulated, as shown in Figure 14e–f.\(^\text{103}\) The passive nanobeam neuron has the potential to be a competitive alternative in implementing large-scale...
an complex-valued weight matrix is used to conduct the linear weighting operation, whereas the nonlinear activation function is implemented with laser-cooled atoms with electromagnetically induced transparency. More recently, a purely passive metaneural network with a small footprint has been proposed by analyzing acoustic scattering,[140] which is able to converge the scatter energy from the scattered object into the corresponding region on the detection plane (see Figure 15c). The phase shifts of metaneurons are chosen as the trainable parameters in training and real-time object recognition of handwritten digits, and misaligned orbital angular momentum beams are successfully demonstrated.

Even though impressive performance has been demonstrated with spatial PNNs, an overarching theme of PNNs has been the on-chip PNNs, which are anticipated to revolutionize the current computing architecture and even unravel the elusiveness of the brain with the merits of large scalability, high energy efficiency, ultracompact footprints, and high programmability. Previously, great efforts have been made to build scalable photonic integrated neural networks. Based on whether the information is always processed in the optical domain or switched from optic to electrical and back forth, PNNs can be generally classified into two categories: all optical and O/E/O.[141]

All-optical links can be realized with either a coherent scheme or an incoherent scheme. The coherent system denotes the single-wavelength operation, and the signals can only be distinguished by their physical paths. Currently, their scalability is restricted by the incompatibility with WDM technology. Other dimensions of light such as mode or polarization may potentially be used to alleviate this emergency of parallel channels. Figure 16a shows the architecture of a fully integrated coherent PNN.[46] The input information (an image or a vowel) is mapped to a high-dimensional vector first with standard algorithms in a computer. Then, the preprocessed signals are encoded and denoted in optical pulses with different amplitudes propagating in the PIC, which implements a multilayer PNN. Each layer of the PNN consists of an optical interference unit (OIU) that utilizes MZI meshes to conduct matrix multiplication and an optical nonlinear unit (ONU) that uses optical nonlinearity such as saturable absorption to implement the nonlinear activation function. As shown in Figure 16b, the fabricated processor realizes both matrix multiplication (highlighted in red) and attenuation (highlighted in blue) completely in the optical domain, which obtains a correctness of 76.7% in the vowel recognition task. Notably, the necessary global phase control for coherent systems poses some challenges in information synchronization, which, in turn, offers extra freedom of degree for flexible weights configuration. Recently, an optical complex-valued neural network has been proposed by encoding and manipulating light signals with both magnitude and phase information during the whole input preparation and network evolution process.[140] Figure 17a shows the architecture of the optical neural chip, which integrates all the input preparation, weight multiplication, and coherent detection modules. The red-marked MZIs realize the separation and modulation of the input light signals, whereas the MZI marked in green separates the reference light for the use of coherent detection. The $6 \times 6$ complex-valued weight matrix is implemented with the MZIs marked in blue. The remaining gray-marked MZIs with a configurable interference stage are designed to switch between intensity and coherent detection.

### Table 4. Recent advances and key performance of spiking neurons in SNNs.

| Type            | Power | Speed   | Cascadability | Footprints | CMOS compatibility |
|-----------------|-------|---------|---------------|------------|--------------------|
| VCSEL-SA[102,119-122] | mW sub-ns | Yes | Large | No |
| Micropillar laser[48,114-116] | mW sub-ns | Yes | Large | No |
| DFB laser[117,118] | mW sub-ns | Yes | Large | No |
| PS-based VCSEL[110,125-129] | mW sub-ns | Yes | Large | No |
| QD laser[50,51] | mW sub-ns | Yes | Large | No |
| PCM[54,117] | mW ns | No | Small | No |
| Microring[120] | mW sub-ns | Yes | Medium | Yes |
| Nanobeam[103] | μW ns | Yes | Medium | Yes |

on-chip neuromorphic processors with operation speed on the timescale of nanosecond and power consumption on the order of microwatt.

Some other types of excitable lasers have been used to emulate spiking neurons. A Fano laser in a photonic crystal platform has been numerically proposed to obtain an all-optical nonlinear activation function with a refractory period of nanoseconds.[139] Similar to the PS effect in VCSELs, the waveband switching effect in QD mode-locked lasers has been utilized to obtain both excitatory and inhibitory spiking neurons.[51] In addition, an inhibitory LIF neuron has been numerically demonstrated with a single-section QD InAs/GaAs laser.[50]

Table 4 shows a brief summary of the performance comparison between different kinds of spiking neurons. In general, the excitable laser takes great advantages of fast operation speed and moderate power consumption. However, the hybrid monolithic integration poses great challenges on its scalability with the rapid growth in the number of neurons. In contrast, the passive micromachined beams, especially for the nanobeam, shows great potential for large-scale neural networks with low power consumption, moderate operation speed, and ultracompact footprints without sacrificing the important CMOS compatibility. However, more experimental results have to be demonstrated to further validate its feasibility.

### 5. Scalable Photonic Integrated Neural Networks

Synapses and neurons can only show their computational capacities when interconnected in a neural network. Previously, the native interference nature of light has been developed to implement PNNs in free space. Following Huygens’ principle, 3D-printed diffractive deep neural networks have been created for image classification of handwritten digits.[40] As shown in Figure 15a, each point on a given transmissive/reflective layer works as a neuron with a complex-valued transmission/reflection coefficient, which can be trained with classical deep learning methods to obtain a specific map between the input and output panel of the network. In addition, a fully functional all-optical PNN with the capability of categorizing different phases for the Ising model has been proposed.[43] As shown in Figure 15b, the spatial light modulators and Fourier lenses are used to conduct the linear weighting operation, whereas the nonlinear activation function is implemented with laser-cooled atoms with electromagnetically induced transparency. More recently, a purely passive metaneural network with a small footprint has been proposed by analyzing acoustic scattering,[140] which is able to converge the scatter energy from the scattered object into the corresponding region on the detection plane (see Figure 15c). The phase shifts of metaneurons are chosen as the trainable parameters in training and real-time object recognition of handwritten digits, and misaligned orbital angular momentum beams are successfully demonstrated.

Even though impressive performance has been demonstrated with spatial PNNs, an overarching theme of PNNs has been the on-chip PNNs, which are anticipated to revolutionize the current computing architecture and even unravel the elusiveness of the brain with the merits of large scalability, high energy efficiency, ultracompact footprints, and high programmability. Previously, great efforts have been made to build scalable photonic integrated neural networks. Based on whether the information is always processed in the optical domain or switched from optic to electrical and back forth, PNNs can be generally classified into two categories: all optical and O/E/O.[141] All-optical links can be realized with either a coherent scheme or an incoherent scheme. The coherent system denotes the single-wavelength operation, and the signals can only be distinguished by their physical paths. Currently, their scalability is restricted by the incompatibility with WDM technology. Other dimensions of light such as mode or polarization may potentially be used to alleviate this emergency of parallel channels. Figure 16a shows the architecture of a fully integrated coherent PNN.[46] The input information (an image or a vowel) is mapped to a high-dimensional vector first with standard algorithms in a computer. Then, the preprocessed signals are encoded and denoted in optical pulses with different amplitudes propagating in the PIC, which implements a multilayer PNN. Each layer of the PNN consists of an optical interference unit (OIU) that utilizes MZI meshes to conduct matrix multiplication and an optical nonlinear unit (ONU) that uses optical nonlinearity such as saturable absorption to implement the nonlinear activation function. As shown in Figure 16b, the fabricated processor realizes both matrix multiplication (highlighted in red) and attenuation (highlighted in blue) completely in the optical domain, which obtains a correctness of 76.7% in the vowel recognition task. Notably, the necessary global phase control for coherent systems poses some challenges in information synchronization, which, in turn, offers extra freedom of degree for flexible weights configuration. Recently, an optical complex-valued neural network has been proposed by encoding and manipulating light signals with both magnitude and phase information during the whole input preparation and network evolution process.[140] Figure 17a shows the architecture of the optical neural chip, which integrates all the input preparation, weight multiplication, and coherent detection modules. The red-marked MZIs realize the separation and modulation of the input light signals, whereas the MZI marked in green separates the reference light for the use of coherent detection. The $6 \times 6$ complex-valued weight matrix is implemented with the MZIs marked in blue. The remaining gray-marked MZIs with a configurable interference stage are designed to switch between intensity and coherent detection.
The application of nonlinearity and the calculation of loss function are completed via the electrical interface. The optical complex-valued neural network is benchmarked in four practical tasks: 1) the implementation of elementary logic gates with a single complex-valued neuron, including the nonlinear XOR gate which usually requires a three-layered real-valued neural network; 2) the classification of Iris species using a single complex layer with an accuracy up to 97.4%; 3) the classification of the nonlinear dataset Circle and Spiral. The model constructions and the visualization of decision boundaries obtained from simulations and experimental measurements are shown in Figure 17b–e from left to right. In contrast to the straight lines for the simulated real model, the decision boundaries of the simulated complex model are nonlinear curves which perfectly match the entangled shape of the datasets. In addition, the measured accuracies are 98% for the Circle and 95% for the Spiral, respectively. 4) The recognition of handwritten digits with a complex multilayer perceptron network, which improves the testing accuracy from 82.0% of the real-valued counterpart to 90.5%.

In typical incoherent configurations, the output wavelength usually has a strong dependence on the input wavelength. For instance, to guarantee the fiber-based GEL works correctly, the input wavelength should be shorter than the output wavelength to realize carrier population inversion. Therefore, novel networking schemes are highly desired.

On the contrary, it is quite straightforward to build integrated PNNs with the O/E/O link in WDM frameworks, where the O/E subcircuit uses photodetectors to provide the weighted addition of multiple WDM signals, whereas the E/O subcircuit is responsible for the nonlinear functionality.
Figure 16. a) General architecture of an all-optical fully integrated neural network. b) Optical micrograph illustration of a fabricated OIU. (a,b) Reproduced with permission.\textsuperscript{[42]} Copyright 2017, Springer Nature.

Figure 17. a) The schematic of the complex-valued neural network chip. For the circle (top) and spiral (down) dataset, the subfigures successively present b) the model constructions, the decision boundaries obtained from c) the simulated real model, d) the simulated complex model, as well as e) the hardware implementation of the complex model, respectively. (a–e) Reproduced with permission.\textsuperscript{[140]} Copyright 2021, Springer Nature.
been proposed to build practical scalable PNNs, which enable dense connections between neurons based on WDM technology. As shown in Figure 18a, each neuron is assigned with a unique wavelength and broadcasts its output signal into the bus waveguide, which guarantees efficient information communication between any two neurons. Independent weighting is accomplished by a series of tunable spectra filters, typically MRRs in the optical domain. The weighted signals are not demultiplexed in the network but directly fed to photodetectors to generate a spatial sum of the input channels. Finally, laser neurons transform the electric signals into corresponding optical outputs. Based on the B&W protocol, a neuromorphic system composed of two off-chip Mach–Zehnder modulators (MZMs) and two on-chip MRR weight banks each with four MRR weights (Figure 18b) is experimentally validated to be isomorphic to a continuous-time recurrent neural network (CTRNN) model.

Although the B&W protocol is originally designed for SNNs implemented with the O/E/O link, the design principle has been extended to obtain an all-optical neuromorphic system with self-learning capacity based on PCM. As shown in Figure 19a, the whole PNN is composed of one input layer, multiple hidden layers, and one output layer. Each layer of the network consists of a collector that unites all the outputs from the previous layer, a distributor that distributes the input signal equally to each

![Figure 18](image_url)

**Figure 18.** a) B&W architecture utilizing the O/E/O link for scalable PNNs. b) Optical micrograph of a four-node recurrent B&W network. (a) Reproduced with permission.[78] Copyright 2016, IEEE. (b) Reproduced with permission.[142] Copyright 2017, Springer Nature.

![Figure 19](image_url)

**Figure 19.** PCM-implemented all-optical spiking neurosynaptic system. a) The general neural network architecture. b) Photonic implementation of a single layer from the network. c) Optical micrograph of three fabricated neurons (B5, D1, and D2). d) The whole system consisting of four photonic neurons, each with 15 synapses. e) The change in output spike intensity for the four trained patterns illustrated on the right-hand side. (a–e) Reproduced with permission.[54] Copyright 2019, Springer Nature.
individual neuron within the layer, and a set of neurosynaptic elements. Figure 19b shows the photonic hardware implementation of each layer. MRRs are utilized to multiplex the outputs from a previous layer into a single bus waveguide (the collector) and demultiplex the signal to each neuron equally by carefully engineering the coupling gap between the bus waveguide and MRRs (the distributor). Figure 19c,d shows the experimental implementation of such a single-layer SNN consisting of 4 neurons with 15 synapses each. Each pixel of a 3 × 5 image is mapped on one synapse and encoded in the wavelength corresponding to the ring multiplexer (see numbering in Figure 19e). After training, the system can successfully distinguish four 15-pixel images as each neuron can only be activated by one of the input patterns (Figure 19e).

Over the past decade, the optical frequency combs in the integrated platforms have burgeoned with various exciting demonstrations,[143–147] featuring compact footprints, high scalability, and reliable performance. There are mainly two different underlying principles for optical frequency combs: the processes of supercontinuum generation (SCG) in optical waveguides (Figure 20a) and Kerr-comb generation (KCG) in microresonators (Figure 20b).[148] The rapid development of microcombs has enabled many fundamental breakthroughs in diverse research areas including quantum information processing (Figure 20c),[143] massively parallel communications (Figure 20d),[147] optical frequency synthesis (Figure 20e),[145] massively parallel light detection and ranging (Figure 20f),[146] advanced microwave photonics,[149] etc. More detailed discussions on microcombs can be found in these impressive reviews,[148,150–152] More recently, the microcomb has emerged as a powerful tool for large-scale PNNs, and several extraordinary works have been reported.

As shown in Figure 21, the design concept of the single-neuron perceptron based on the microcomb has been extended to implement a photonic vector convolutional accelerator (VCA).[153] The weight matrices of ten 3 × 3 kernels are rearranged into a combined weight vector and then mapped onto the power intensity of 90 comb lines by a single waveshaper. Meanwhile, a classical 500 × 500 image is electrically flattened into an input vector and encoded onto the intensity of 250,000 temporal symbols. After the process of broadcast, weighting, and progressive delay, the 90 wavelengths are demultiplexed into 10 sub-bands and detected individually, with each sub-band containing nine wavelengths and corresponding to one kernel. In the last stage, the ten electrical waveforms undergo digital signal processing including analog-to-digital conversion and resampling. As a result, each timeslot of each individual electrical waveform exactly corresponds to the convolutional result between the input image and one of the kernel matrices within a sliding window. Eventually, ten feature maps containing the
extracted hierarchical features of the raw input data can be effectively acquired from the resulting waveforms. Notably, the simultaneous interleaving of temporal, wavelength, and spatial dimensions yields a total computing speed of 11.3 trillion operations per second (TOPS) for the VCA. Furthermore, an optical convolutional neural network (CNN) consisting of a front-end convolutional processor and a fully connected layer is sequentially formed, successfully achieving a recognition accuracy of 88% on the handwritten digit images. Given that a set of components like the microcomb, modulator, dispersive media, demultiplexer, and photodetector have already been realized in integrated forms, it is expected that the photonic VCA can be realized with on-chip integration of much higher levels.

With the combination of a chip-based soliton microcomb and a PCM memory array, a computationally specific photonic tensor core has also been experimentally demonstrated to operate at speeds of trillions of MAC operations per second.\(^\text{[154]}\) As shown in Figure 22a, the pixels of the input image are flattened into matrices of dimension \(1 \times (k^2 \times d_{in})\) and stacked row by row, thus forming a large input matrix of dimension \((n - k + 1)^2 \times (k^2 \times d_{in})\). Meanwhile, a large filter matrix of dimension \((k^2 \times d_{in}) \times d_{out}\) is also obtained by stacking all kernel matrices column by column, which is implemented with a waveguide crossbar array with additional directional couplers (DCs). To conduct a matrix–vector multiplication (MVM), each time, a single row of the input matrix is encoded onto the individual

**Figure 21.** Architecture of the photonics convolutional accelerator. Reproduced with permission.\(^\text{[153]}\) Copyright 2021, Springer Nature.
comb teeth via high-speed electro-optic modulators and fed into the filter matrix. The horizontal DCs will equally distribute the power of the input vector to different columns of the filter matrix, with each column representing one individual image kernel. However, the vertical DCs will combine the input light after interaction with the PCM elements, thus conducting the accumulation operation. It is worth mentioning that each input vector only interacts with a single PCM cell per matrix column. The output power at each column represents the result of one MAC operation between the input vector and a kernel multiplied by a certain fixed factor of $1/(m \times n)$, depending on the matrix size. In this way, the convolitional operations between the input data and all kernels can be conducted simultaneously. Furthermore, the WDM technology can be utilized to realize parallel MVM by properly (de)multiplexing and combining the wavelengths corresponding to different input vectors. Processing four input vectors in parallel with four image kernels of dimension $4 \times 4$ has been experimentally demonstrated. Figure 22b shows

![Figure 22](image-url)

**Figure 22.** a) Architecture of photonic tensor cores using an on-chip microcomb and PCMs. b) Original input images. c–f) Output images corresponding to four different kernels. g) Combined output images from (c–f) highlighting edges successfully. (a–g) Reproduced with permission. Copyright 2021, Springer Nature.
the original input images. The vertical edges are shown in Figure 22c,f, whereas the horizontal edges are shown in Figure 22d,e. The combined images in Figure 22g successfully highlight all edges. Due to the nonvolatile nature of PCM and the WDM capability provided by the microcomb, the photonic tensor core can conduct massively parallel in-memory computing in theory at the speed of light with ultralow power consumption.

6. Discussion and Challenges

The intersection of neuroscience and photonics is now burgeoning for neuromorphic photonics. It can combine both neuroscience information-processing capacity with non-von Neumann architectures and unique properties of photonics: the essentially unlimited bandwidths, high speed, high power efficiency, multidimensional multiplexing capabilities, and fundamentally immunity to electromagnetic interference. The cointegration feasibility with mature microelectronics in CMOS makes neuromorphic photonics a real hit in future neuromorphic computing hardware options. Neuromorphic photonics generally use low-loss waveguides, high-efficiency couplers, high-speed modulators, and high-sensitivity photodetectors to facilitate high performance and energy-efficient computing architectures.\(^{[39]}\)

The energy consumption mainly lies in the process of input preparation, weight adjustment, nonlinearity activation, and output detection. For the integrated nonvolatile memory,\(^{[38]}\) it consumes almost no power for weight maintenance or phase control once the PCM elements are trained, and the matrix multiplication can be passive and conducted at the speed of light in theory. However, currently, there are still many scientific and technological challenges to overcome envisioning a fully practical integrated neural network.\(^{[39]}\) First and foremost, as the complete neuromorphic photonic computing is essentially an ecosystem with light sources, passive and active components, and transistors working together, there is no single commercial fabrication platform offering such on a single die simultaneously. Current on-chip optical light or gain sources involve either cointegration of III–V using heterointegration,\(^{[56]}\) cointegration with precise assembly approach,\(^{[155]}\) or direct epitaxy approach with QD lasers on silicon,\(^{[58]}\) but the fabrication process is either complicated or with the reliability not up to commercial standards. Cointegrating the photonic systems with transistors and low-power CMOS controllers to implement electrical control/feedback/stabilization should also be crucial for robust PNNs. State-of-the-art photonic solutions aforementioned use different photonic materials (e.g., Si, III–V, PMCs, etc.) with mostly incompatible foundry processes. Nevertheless, synergistic efforts of various cointegration approaches are being actively explored toward a final goal of on-chip PNNs. Second, low power consumption and nonvolatile photonic storage and weighting are necessities to ensure neuromimetic functionalities. Neural nonlinearities are already demonstrated on mainstream platforms using electro-optic transfer functions,\(^{[42,43]}\) PMCs,\(^{[38]}\) and laser spiking behaviors,\(^{[62]}\) but energy efficiency and fast switching of new integrable materials are still at large to promise enhanced performance opportunities. Third, desires for fully reconfigurable integrated PNNs to conduct full ANN operations are booming. Silicon photonics is emerging as an ideal platform for integrating these components, while offering a combination of foundry compatibility and device compactness with low costs, thus allowing for the realization of on-chip scalable neuromorphic photonic systems. The prevailing silicon photonics device development\(^{[56]}\) makes it possible for the construction of high-performance integrated silicon PNNs with aforementioned highly functional optical components. A growing number of electronic communications foundries (Global Foundries, TSMC, etc.) and research institutions (IMEC, AFM, CUMEC, etc.) have released the multi-project wafer (MPW) for wafer-level process services in silicon photonics, and some are even developing the cointegration process of photonics and electronics, which greatly promotes the potential of fully integrated PNNs. Last but not least, to harness the power of both photonics and neuroscience and push these platforms into real applications, there should be significant advances to bridge the current neural network algorithms with PNNs’ physical response. So far in the literature, only few proof-of-concept PNNs with limited control units and neural algorithms are demonstrated for some simple recognition scenarios. It is much more preferable that the neural network programming tools can be compatible with the electronic AI, to directly reconfigure a large-scale neuromorphic photonic processor in the near future.\(^{[62]}\)

Ultimately, the self-contained PNNs have to face and compete with the real high-performance computers, so the robustness to the environment, universal algorithms, and interface with the electrical processors would also be the spotlights.

7. Conclusion and Future Perspectives

The physical and technological limits of electronic integration and the post-Moore era are pressing the community to search for new candidates for next-generation computing architectures. Neuromorphic photonics is a formidable yet competitive candidate, which combines the merits of photonics and the capacity of neural networks and has the potential to conduct any computable tasks by ANNs in principle. There are still many obstacles ahead to be addressed before neuromorphic photonics can be practically implemented, including new materials, fabrication platforms, cointegration with the electronic control units, etc. The booming silicon photonics manufacturing platforms are also continuously promoting the commercialization of large-scale neuromorphic photonics for ANNs. We envisage the development of this emerging field to accelerate as the neuromorphic photonic computing scheme makes further leaps toward AI and expand our understanding of neuroscience. Furthermore, what is noticeable is that there are other promising non-neuromorphic technologies that leverage the splendid properties of photonics for efficient and high-speed information processing or computing, like photonic Ising machines\(^{[136]}\) and the latest extraordinary work of quantum computing using photons.\(^{[157]}\) This Review is not intended to be comprehensive, but we hope to focus on the integrated neuromorphic photonics area and provide some vision for the future development of neuromorphic computing. As the photonic technology matures, augmenting neuromorphic photonic computing technology with increasingly high-performance integrated photonic components will further unleash the potential of ANNs to continue even after the end.
of Moore’s law and beyond the von-Neumann plateau, leading to a bright future for neuromorphic photonics in the years ahead.

Acknowledgements

X.G. and J. X. contributed equally to this work. The work was supported in part by the National Key R&D Program of China under grant 2019YFB2203104, the National Natural Science Foundation of China (NSFC) (61805137 and 61835008), and the Natural Science Foundation of Shanghai, China (19ZR1475400).

Conflict of Interest

The authors declare no conflict of interest.

Keywords

artificial intelligence, deep learning, neuromorphic networks, photonic integrated circuits, silicon photonics

Received: December 30, 2020
Revised: March 2, 2021
Published online: May 7, 2021

[1] E. D. Adrian, The Basis of Sensation, W. W. Norton & Co, New York 1928.
[2] K. He, X. Zhang, S. Ren, J. Sun, presented at Proc. IEEE Conf. Computer Vision Pattern Recognit (CVPR), Las Vegas, NV, USA, June 2016.
[3] D. Silver, A. Huang, C. J. Maddison, A. Guez, V. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, H. Hassabis, Nature 2016, 529, 484.
[4] E. Callaway, https://www.nature.com/articles/d41586-020-03348-4 (accessed: November 2020).
[5] J. Bardeen, W. H. Brattain, US Patent 2524035, Bell Telephone Laboratories, 1950.
[6] W. Shockley, US Patent 2569347, Bell Telephone Laboratories, 1951.
[7] F. Wanlass, C. Sah, presented at the 1963 IEEE Int. Solid-State Circuits Conf., Philadelphia, PA, USA, February 1963.
[8] B. J. Shastry, M. A. Nahmias, A. N. Tait, A. W. Rodriguez, B. Wu, P. R. Prucnal, Sci. Rep. 2016, 6, 19126.
[9] R. H. Dennard, IEEE Trans. Electron Devices 1984, 31, 1549.
[10] G. E. Moore, Electronics 1965, 38, 114.
[11] M. M. Waldrop, Nature 2016, 530, 144.
[12] N. Jones, Nature 2018, 561, 163.
[13] M. M. S. Aly, M. Gao, M. Shulaker, T. F. Wu, M. Asgheri, J. Bokor, F. Schmitt, K. E. Goodson, C. Kozyrakis, I. Markov, K. Olukotun, L. Pileggi, E. Pop, J. Rabaey, C. Ré, H. P. Wong, S. Mitra, Computer 2015, 48, 24.
[14] E. Bullmore, O. Sporns, Nat. Rev. Neurosci. 2012, 13, 336.
[15] D. Ielmini, Microelectron. Eng. 2018, 190, 144.
[16] David D. Cox, T. Dean, Curr. Biol. 2014, 24, R921.
[17] M. Milakov, https://www.numpy.org/2019/04/10, 9024 (accessed: November 2014).
[18] C. M. McMillan, Proc. IEEE 1990, 78, 1629.
[19] B. M. A. Nahmias, A. N. Tait, T. F. Wu, M. Asgheri, J. Bokor, F. Schmitt, K. E. Goodson, C. Kozyrakis, I. Markov, K. Olukotun, L. Pileggi, E. Pop, J. Rabaey, C. Ré, H. P. Wong, S. Mitra, Computer 2015, 48, 24.
[20] E. Bullmore, O. Sporns, Nat. Rev. Neurosci. 2012, 13, 336.
[21] D. Ielmini, Microelectron. Eng. 2018, 190, 144.
[22] David D. Cox, T. Dean, Curr. Biol. 2014, 24, R921.
[23] M. Milakov, https://www.numpy.org/2019/04/10, 9024 (accessed: November 2014).
[24] C. M. McMillan, Proc. IEEE 1990, 78, 1629.
[25] B. M. A. Nahmias, A. N. Tait, T. F. Wu, M. Asgheri, J. Bokor, F. Schmitt, K. E. Goodson, C. Kozyrakis, I. Markov, K. Olukotun, L. Pileggi, E. Pop, J. Rabaey, C. Ré, H. P. Wong, S. Mitra, Computer 2015, 48, 24.
[26] E. Bullmore, O. Sporns, Nat. Rev. Neurosci. 2012, 13, 336.
[27] D. Ielmini, Microelectron. Eng. 2018, 190, 144.
[28] David D. Cox, T. Dean, Curr. Biol. 2014, 24, R921.
[29] M. Milakov, https://www.numpy.org/2019/04/10, 9024 (accessed: November 2014).
[30] C. M. McMillan, Proc. IEEE 1990, 78, 1629.
Xuhan Guo received his Ph.D. in Electrical Engineering from University of Cambridge, Cambridgeshire, UK, in 2013. He worked at University of Cambridge and then joined the Shanghai Jiao Tong University as an associate professor in 2017. His research areas cover silicon photonic devices for information processing and neuromorphic photonics.

Jinlong Xiang received the B.E. in Department of Electronic Engineering from Shanghai Jiao Tong University, Shanghai, China. In September 2019, he joined the State Key Laboratory of Advanced Optical Communication Systems and Networks. His main research interests include silicon photonics devices and circuits and neuromorphic computing.

Yujia Zhang received her B.Eng. from Nanjing University in Microelectronics Science and Engineering, Nanjing, China, and obtained her M.S. from UC Davis in Electrical and Computer Engineering, Davis, USA. Now she is pursuing Ph. D. at Shanghai Jiao Tong University. Her current research interests mainly involve the design and fabrication of integrated silicon photonic devices and systems.

Yikai Su received his Ph.D. in Electrical Engineering from Northwestern University, Evanston, IL, USA, in 2001. He worked at Crawford Hill Laboratory of Bell Laboratories and he joined the Shanghai Jiao Tong University as a full professor in 2004. His research areas cover silicon photonic devices for information transmission and switching.