SCIENCE CHINA
Information Sciences

June 2020, Vol. 63 160403:1–160403:14
https://doi.org/10.1007/s11432-020-2872-3

• REVIEW •
Special Focus on Photonics in AI

Towards silicon photonic neural networks for artificial intelligence

Bowen BAI[1,2], Haowen SHU[1,2], Xingjun WANG[1,2*] & Weiwen ZOU[3]

[1]*State Key Laboratory of Advanced Optical Communications System and Networks, School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871*, China;*
[2]*Frontiers Science Center for Nano-optoelectronics, Peking University, Beijing* 100871*, China;*
[3]*State Key Laboratory of Advanced Optical Communication Systems and Networks, Intelligent Microwave Lightwave Integration Innovation Center* (*iMLic*)*, Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai* 200240*, China*

Received 15 February 2020/Revised 18 March 2020/Accepted 13 April 2020/Published online 9 May 2020

Abstract Brain-inspired photonic neural networks for artificial intelligence have attracted renewed interest. For many computational tasks, such as image recognition, speech processing and deep learning, photonic neural networks have the potential to increase the computing speed and energy efficiency on the orders of magnitude compared with digital electronics. Silicon Photonics, which combines the advantages of electronics and photonics, brings hope for the large-scale photonic neural network integration. This paper walks through the basic concept of artificial neural networks and focuses on the key devices which construct the silicon photonic neuromorphic systems. We review some recent important progress in silicon photonic neural networks, which include multilayer artificial neural networks and brain-like neuromorphic systems, for artificial intelligence. A prototype of silicon photonic artificial intelligence processor for ultra-fast neural network computing is also proposed. We hope this paper gives a detailed overview and a deeper understanding of this emerging field.

Keywords photonic neural networks, artificial intelligence, deep learning, photonic computing accelerator, silicon photonics

Citation Bai B W, Shu H W, Wang X J, et al. Towards silicon photonic neural networks for artificial intelligence. Sci China Inf Sci, 2020, 63(6): 160403, https://doi.org/10.1007/s11432-020-2872-3

1 Introduction

Artificial intelligence (AI) has already widely used in almost every aspect of our daily lives and shown remarkably good performance in computational tasks, such as natural language processing and visual object recognition [1, 2]. Inspired by parallel signal processing in human brain and benefited from the explosion of data, AI has revitalized and attracted the interest of researchers recently. Intel [3], IBM [4], and Google [5], have all made AI the most important strategic development direction. Deep learning [6] with artificial neural networks (ANNs) is the key driving force promoting the explosive growth of AI. Neural network algorithm contains massive multiply accumulate computations (MAC), while the central processing units (CPUs) designed for traditional von Neumann architecture are laborious to perform these operations. The memory for computing and signal processing is physically separated in von Neumann scheme and the CPUs operating in a sequential way. The data flows between memory and processor limit the computing efficiency when implementing massively parallel signal processing. The one-size-fits-all approach is no longer capable of AI computing tasks. Therefore, researchers focus their attention on

* Corresponding author (email: xjwang@pku.edu.cn)

new hardware architectures (such as graphical processing units (GPUs), field-programmable gate arrays (FPGAs)) specifically for ANNs and deep learning. Although GPUs, FPGAs and even neuromorphic electronics, including IBM TrueNorth [7] and Google TPU [5], have improved both energy efficiency and speed for inference (learning) tasks, the end of Moore's Law [8] radically impedes the further development of electronic processors. The fundamental energy consumption and bandwidth limitation [9] of electrical interconnection have become the major bottlenecks in present AI hardware. In some power-critical situations, such as unmanned aerial vehicle (UAV), smart phone and 'edge computing', these issues become more intractable.

1.1 Photonics powers AI chips

Current microelectronic chips, regardless of the based technical framework, are designed and manufactured using traditional microelectronic process. With the improvement of microelectronic integration, the performance of microelectronic chips improves continuously. In the past few decades, microelectronic technology continues to advance, according to Moore's Law. However, in recent years, the development of microelectronics processes can hardly follow Moore's Law [8]. Limited by a series of problems such as cross-talk, power consumption, noise, and time delay, it has been incredibly difficult to further improve the information processing capabilities of microelectronic chips by simply improving integration density and operation frequency. Novel techniques are required to break the intrinsic limitations of conventional microelectronic computing framework.

Owing to the high speed, parallel processing capability and low energy consumption, photonics shows great potential to address these bottlenecks well. Researchers have tried to develop photonic devices to simulate neurons and synapses in biological brains to further improve the data processing and analysis capabilities. In some specific applications, such as UAV [10] and self-driving [11], rapid data analysis and situation judgment are essential concerns. These photonic neurons [12] constitute 'brain-like' photonic neural networks (PNNs) which can be integrated on a chip. On one hand, PNNs use photons to perform calculating and data exchanging, which offers a promising alternative approach for AI hardware accelerators. On the other hand, silicon photonic technology utilizes compatible mature microelectronic process to integrate photonic and electronic devices simultaneously, which is the ideal fabrication platform of photonic hardware. Nowadays, silicon PNNs have become a research hotspot in academia and industry.

1.2 Silicon photonic technology

Because of the cost and special fabrication requirements, large-scale manufacturing of traditional discrete photonic devices is quite challenging. Similar to microelectronic hardware, the improvement of computing performance for photonic chips also relies on the growing number of devices. Fortunately, the emergence of silicon photonics, which is compatible with standard complementary metal-oxide-semiconductor (CMOS) microelectronic integration technology, opens the gate for large-scale fabrication and reproducible of photonic chips. Silicon photonic technology utilizes both photons and electrons as information carriers to integrate photonic structures with electronic devices on a same silicon substrate simultaneously [13, 14]. Then, a new integrated chip with comprehensive functions needed for rapid information processing can be formed. Silicon photonics naturally has the dual advantages of electronics and photonics, enabling photonic devices to be integrated and mass-produced on the same scale as microelectronic chips, with the calculation speed of light while minimal energy consumption. Although increasing the integration density of photonic devices is quite challenging, silicon photonics has made great progress in recent years [15, 16]. The mature CMOS integration technology improves the yield of photonic chips while reduces the manufacturing costs as much as possible.

Light sources are a key ingredient in photonic neural networks. Silicon does not emit light owing to its indirect band-gap, therefore, lasers made of III-V semiconductors are usually separately packaged as external light sources. This method requires precise alignment between the lasers and waveguides, which suffers from higher coupling loss and packaging cost. Fortunately, these hurdles are being overcome with technology such as hybrid photonic integration. Interuniversity Microelectronics Centre (IMEC) and CST

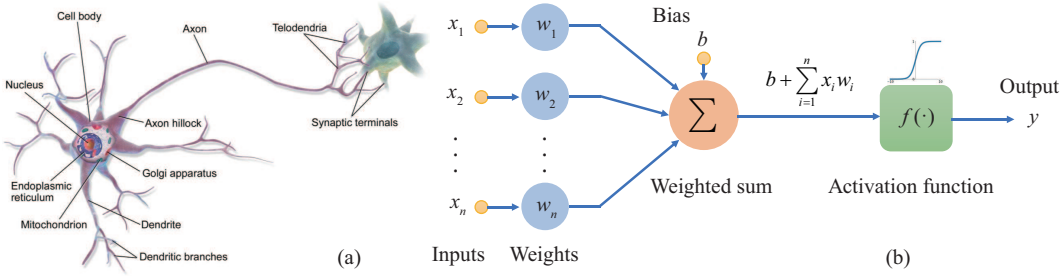

Figure 1 (Color online) (a) Schematic of a neuron. (b) An artificial neuron with simple nonlinear model: showing the input $(x_1, x_2, \ldots, x_n)$, their relevant weights $(w_1, w_2, \ldots, w_n)$, bias b and the non-linear activation function $f(x)$ applied to the weighted sum of input signals. The output is connected to other neurons through synapses (connecting links), forming a neural network.

Global have successfully integrated InP distributed feedback (DFB) lasers into photonics platform through a die-to-die bonding process. The hybrid integration property of silicon photonics can dramatically improve the energy efficiency and reduce the monetary cost of current photonic architectures, particularly in photonic neural network applications.

2 Artificial neural network

Inspired by the physical neuron system, ANN is a model that imitates the information processing of human brains. ANN can perform complex logic and non-linear operations and is capable of parallel distributed processing with high fault tolerance. These unique properties make ANN hold an important status in the field of AI.

2.1 Artificial neuron

The rapid growth of ANN for AI tasks has led to intense exploration of efficient hardware implementations to mimic the natural processing capabilities of the brain. Neuron is the basic functional unit of human brain, as shwon in Figure 1(a). Researchers strive to comprehend the complex functionality of neuron and emulate its unparalleled energy efficiency. An elementary illustration of an artificial neuron is shown in Figure 1(b). The neuron consists of three parts: the first is a set of weighted connections called synapses. The input signals $(x_1, x_2, \ldots, x_n)$ is weighted by their relevant weights $(w_1, w_2, \ldots, w_n)$; the second is a liner combiner, performing weighted addition; the third is a nonlinear activation function $f(x)$ (usually monotonic and bounded). The combined signals experience a nonlinear process and then output. $f(x)$ has a normalizing effect, which prevents the divergence of the output after several layers. Artificial neurons can be trained (rather than programmed) to execute a computing task by feeding massive of data, called learning. Nowadays, artificial neurons combined with 'deep learning' algorithms [17] have received an explosion of interest in both academia and industry for their utility in image recognition [18], language translation, decision-making problems and so on [6].

2.2 Multilayer architecture

Figure 2 shows a multilayer ANN which consists of a large number of interconnected artificial neurons. The connections between two neurons represent a set of weighted signals passing through the network. The input should be preprocessed and vectorized data, such as voice, image and language. The input layer connects to at least one hidden layer. In each layer, data experiences a linear combination (e.g., matrix multiplications) followed by an nonlinear activation. Each neuron passes data to all the neurons in the next hidden layer until the last output layer, which gives the final results. The acyclic topology means that there are no feedback connections or loops in the network. The input and output layers are the optical interfaces to the real world. ANNs can be well trained by feeding enough training data into the

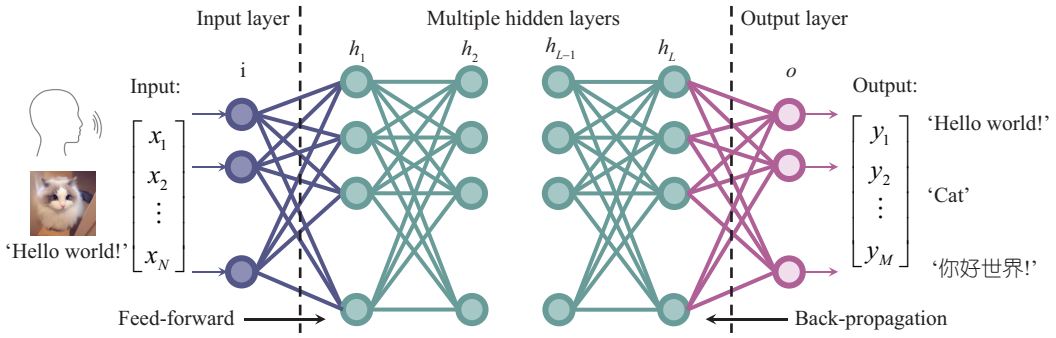

Figure 2 (Color online) Multilayer artificial neural network scheme composed of an input layer, multiple hidden layers and an output layer. The circles are neurons and each neuron is connected to all neurons in the next layer.

network and then computing the output by forward propagation; the weight parameters are subsequently optimized and tuned using standard back propagation method.

3 Silicon photonic neural networks

Silicon PNN, which consists of interconnected silicon photonic devices, is a concrete form of ANN. Recently, a number of studies on silicon PNNs to accelerate computing and reduce power consumption have been proposed, such as artificial neural networks [19–23] and brain-like neuromorphic networks [12,24–28]. Each layer of the network can be realized by using silicon Mach-Zehnder interferometer (or microring resonator) array and the silicon waveguides act as connection between two adjacent layers. Compare with electronic neural computing schemes, silicon PNNs have much faster operating speed and lower energy consumption [29]. Because of the large volume and high energy consumption, traditional discrete photonic devices are unsuitable for constructing complex systems. Therefore, photonic integration has become an inevitable choice to implement high-performance PNNs. Silicon photonics, which combines the high performance, super manufacturability, and widespread demand, well addresses these issues and has the potential to integrate large scale neural networks that vastly exceed the capabilities of electronics. This section focuses several basic cells for building a silicon PNN and describes some highlighted achievements in this nascent field.

3.1 Silicon photonic integration

By patterning silicon-on-insulator (SOI) or bulk silicon wafers using modern lithographic technology, silicon photonics enables precise alignment and cheap large-scale manufacturing of electronic-photonic chips. With the development of silicon photonic devices, such as low-loss optical waveguides [30], high-efficiency fiber-to-chip couplers [31], fast electro-optic modulators [32,33], and broadband silicon germanium photodetectors [34], it has been possible to construct high-performance integrated silicon PNNs. Some electronics foundries (e.g., Global Foundries and Taiwan Semiconductor Manufacturing Company (TSMC)) and research institutions (e.g., IMEC and Advanced Micro Foundry (AMF)) have been able to provide silicon photonics multi-project wafer (MPW) and low-volume wafer-level process service. In China, the United Microelectronics Center (CUMEC) is in the process of launching 8-inch silicon photonic manufacturing line. The emergence of commercial silicon photonics manufacturing platforms will continuously promote the commercialization of large-scale photonic parallel-computing accelerator for AI.

3.2 Silicon photonic units in neural networks

Silicon waveguides, which enable low-loss light propagation, are the fundamental building block silicon photonic devices or circuits. Waveguides are essential for communication and computing applications because they can transmit data at very high speed and they are immune to electromagnetic interference. In silicon photonic waveguide, the high-index silicon core is surrounded by low-index material (e.g., silica

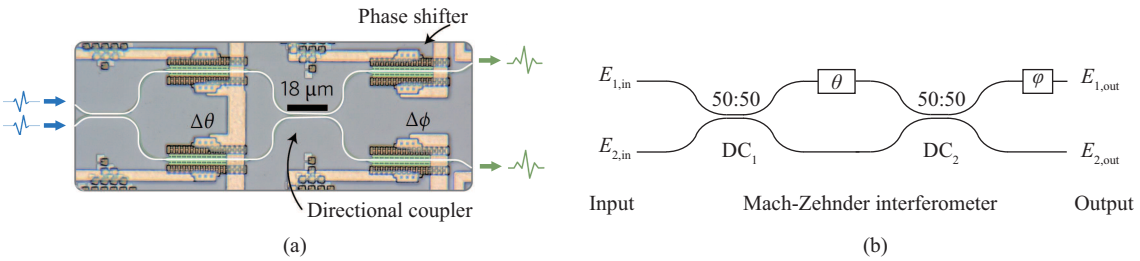

Figure 3 (Color online) (a) An individual programmable Mach-Zehnder interferometer with two thermo-optic phase shifters [21] @Copyright 2017 Springer Nature. (b) Schematic illustration of a programmable MZI, which comprises a phase shifter (θ) between two 50:50 evanescent directional couplers, followed by another phase shifter (φ).

or air), called cladding. The guided light propagates in the waveguide along the longitudinal direction (z direction) and is confined to the small region either within or adjacent to the silicon core surfaces. The properties of the waveguide are characterized by the spatially transverse distribution of the refractive index, $n(x, y)$. Owing to the high refractive index contrast between Si and SiO_2 ($n_{\text{core}} = 3.47$, $n_{\text{clad}} = 1.44$ @ wavelength $\lambda = 1.55$ μm), light is tightly confined in an extremely small effective mode area, which makes the photonic device quite compact. By solving Maxwell's equations while combining boundary conditions, the electric and magnetic field of a guide mode can be express as

$$E_\nu(r,t) = E_\nu(x,y)\cdot \mathrm{e}^{\mathrm{j}\beta_\nu z - \mathrm{j}\omega t}, \quad H_\nu(r,t) = H_\nu(x,y)\cdot \mathrm{e}^{\mathrm{j}\beta_\nu z - \mathrm{j}\omega t}, \tag{1}$$

where ν is the mode order, $E_\nu(x,y)$ and $H_\nu(x,y)$ are the transverse mode field distributions, and β_ν is the propagation constant of the mode. Another important parameter to characterize a mode is effective refractive index, written as

$$n_{\text{eff}} = \beta_\nu / k_0, \tag{2}$$

where $k_0 = 2\pi/\lambda$ is a wave vector in free space. n_{eff} signifies how strongly the light is confined to the waveguide core. Analytic methods to calculate n_{eff} are quite complex, therefore, numerical simulation is of great importance when designing photonic waveguides.

Mach-Zehnder interferometer (MZI) is the most commonly interferometer used in photonic integrated circuits, especially for electrical or thermo modulation of photonic signals. A lossless phase-modulated MZI (Figure 3) with two perfect (50:50) beam splitters performs an optical transformation, which can be described by a 2×2 unitary matrix $U(2)$. The scattering matrix of a single directional coupler DC is

$$\mathrm{DC} = \begin{pmatrix} \cos(\kappa L) & \mathrm{j}\sin(\kappa L) \\ \mathrm{j}\sin(\kappa L) & \cos(\kappa L) \end{pmatrix}, \tag{3}$$

where κ is the coupling coefficient, L is the length of the coupling region; j is the square root of -1, which indicates that the field cross the directional coupler acquires a $\pi/2$ phase shift. For 50:50 beam splitter, $\kappa L = \pi/4$, therefore, the unitary matrix $U(2)$ can be describe as

$$\begin{aligned} U(2) &= P_\varphi \cdot \mathrm{DC}_2 \cdot P_\vartheta \cdot \mathrm{DC}_1 \\ &= \begin{pmatrix} \mathrm{e}^{\mathrm{j}\varphi} & 0 \\ 0 & 1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & \mathrm{j} \\ \mathrm{j} & 1 \end{pmatrix} \begin{pmatrix} \mathrm{e}^{\mathrm{j}\vartheta} & 0 \\ 0 & 1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & \mathrm{j} \\ \mathrm{j} & 1 \end{pmatrix} = \mathrm{j}\mathrm{e}^{\mathrm{j}\vartheta/2} \begin{pmatrix} \mathrm{e}^{\mathrm{j}\varphi}\sin(\vartheta/2) & \mathrm{e}^{\mathrm{j}\varphi}\cos(\vartheta/2) \\ \cos(\vartheta/2) & -\sin(\vartheta/2) \end{pmatrix}, \end{aligned} \tag{4}$$

where P_φ and P_ϑ are the phase-shift operators, DC_1 and DC_2 are the 50:50 beam splitter operators and the same phase factor $\mathrm{j}\mathrm{e}^{\vartheta/2}$ can be ignored. Arbitrary unitary transformation U can be decomposed into sets of $U(2)$ rotations by cascading programmable MZIs (Figure 4) and the relationship between input and output optical signals is $E_{\text{out}} = UE_{\text{in}}$. Generalized discussion of using regular universal multiport interferometers can be found in [19, 35].

Microring resonator (MRR) is another important fundamental component for photonic integrated circuits. On-chip synaptic weights and fully programmable neural networks called broadcast-and-weight

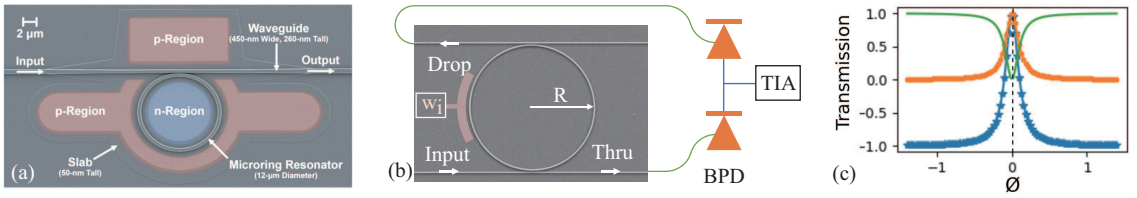

Figure 4 (Color online) (a) Top-view SEM image of a silicon all-pass microring resonator [37] @Copyright 2010 IEEE. (b) A symmetric add-drop microring resonator on SOI with O/E conversion and amplification. (c) Output of the balanced photodiode (blue triangle curve) as a function of the detuning ∅. The orange and green lines are the transmissions of drop T_{drop} and through T_{thru} ports, respectively.

schemes [24, 36] can be created using tunable silicon MRRs. An MRR consists of one or two evanescent directional couplers and a ring cavity. In steady state, the outgoing energy (absorption) of the ring equals that of the incoming energy. For all-pass type (Figure 4(a) [37]), the input light from the bus waveguide transfers into the ring cavity through the directional coupler and then recombines at the same position, thereby interfering with the input light. The power transmission between the output port and input port can be expressed as

$$T(\phi) = \frac{a^2 - 2ra\cos(\phi) + r^2}{1 - 2ra\cos(\phi) + (ra)^2} = \begin{cases} 1, & \text{if } a \to 1, \\ \frac{(r-a)^2}{(1-ra)^2}, & \text{if } \phi = 0, \end{cases} \tag{5}$$

where r is the self-coupling coefficient, a defines the propagation loss of the ring and the directional coupler. The phase ϕ depends on the wavelength λ and radius R of the MRR:

$$\phi = kL = \frac{2\pi}{\lambda} n_{\text{eff}} \cdot L, \tag{6}$$

where L is the distance around the ring cavity. The value of ϕ can be tuned by applying a current across the embedded heater, resulting in a phase shift of the transmission spectrum. For add-drop type MRR (Figure 4(b)), the power transmission of the thru port T_{thru} and drop port T_{drop} with respect to the input port are

$$T_{\text{thru}}(\phi) = \frac{(ra)^2 - 2r^2a\cos(\phi) + r^2}{1 - 2r^2a\cos(\phi) + (r^2a)^2}, \quad T_{\text{drop}}(\phi) = \frac{(1-r^2)^2 a}{1 - 2r^2a\cos(\phi) + (r^2a)^2}. \tag{7}$$

When ignoring the coupling loss and $a \to 1$, the power transmission relationship between thru port and drop port is $T_{\text{thru}} + T_{\text{drop}} = 1$. Assuming that the data signal $a_k(t)$ is normalized ($|a_k(t)| \leqslant 1$) and modulated to the amplitude of the electric field, then the time-frequency expression for the input optical signal is

$$E_{\text{in}}(\omega) = E_0 \sqrt{\frac{1 + a_k(t)}{2}} \delta(\omega - \omega_0) = E_0 \sqrt{x_k(t)} \delta(\omega - \omega_0), \tag{8}$$

where E_0 is the amplitude of the input electric field, δ is the Dirac delta function, and ω_0 is the frequency of the optical field. If the thru port and drop port are connected to a balanced photodiode (BPD) and a TIA with gain of G (Figure 4(b)), the photocurrent of the BPD after TIA amplification is described by

$$i_{\text{BPD}} = x_k(t) \underbrace{R_0 G \left[T_{\text{drop}}(\phi_k) - T_{\text{thru}}(\phi_k)\right] |E_0|^2}_{\mu_k} = \underbrace{\mu_k x_k(t)}_{\text{Dot Product}}, \tag{9}$$

where $= R(\omega_0)$ is spectral response, which can be roughly seen as a constant R_0 in the research spectral region; ϕ_k is the phase when applying a specific current w_k to the heater. It is worth noting that the value of $T_{\text{drop}}(\phi_k) - T_{\text{thru}}(\phi_k)$ is between -1 and $+1$, as shown in Figure 4(c). Such photonic framework can be used to create any kernel value and realize vector dot product [38], which is the key operation in a convolution neural network [39].

3.3 Silicon photonic artificial neural network

There are two specific forms of silicon PNNs. One is multilayer artificial neural networks using silicon photonic structures to achieve matrix multiplication and nonlinear activation; the other is brain-like neural

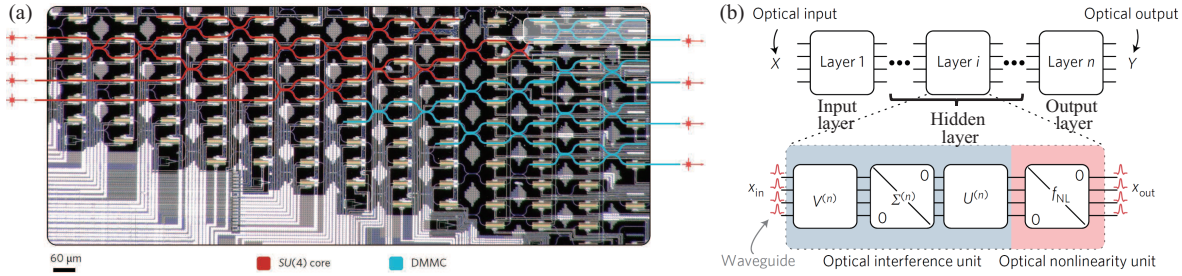

Figure 5 (Color online) (a) Optical micrograph image of the silicon photonic ANN using MZI arrays with 4 input ports and 4 output ports. (b) General PNN architecture and an individual layer consisting of optical interference and nonlinearity units. Reprinted from [21] @Copyright 2017 Springer Nature.

systems which mimic the synaptic-like connections of physical neurons. Silicon PNNs combine the parallel signal processing capabilities of neuromorphic and the high speed and bandwidth of silicon photonics, which offer a promising optoelectronic computing paradigm for AI. Monolithic silicon photonic integration technology [14] allows greater compatibility with digital electronic devices, enabling compact and powerful photonic accelerator. Neural network computing relies heavily on fixed matrix multiplication, which can be well realized by specially designed photonic structures (e.g., MZI arrays and MRRs). Different from traditional electronic hardware, PNNs transport data and perform calculations at the speed of light, namely, the training and inference tasks in AI could be fulfilled in optical domain. Once a PNN is trained, the entire structure becomes a passive system, and the energy consumption of matrix multiplication is almost zero. These unique features enable PNNs far more efficient and faster than their electronic counterparts.

A silicon photonic ANN utilizing coherent nanophotonic circuits for deep learning has been demonstrated in [21]. The reconfigurable silicon photonic ANN (Figure 5(a)), which realizes matrix multiplication (highlighted in red) and attenuation (highlighted in blue) via constructive and destructive interference effects, is constituted by tuneable MZI arrays. This structure includes input layer, hidden layers and output layer, as shown in Figure 5(b). In each layer, the input optical signals first experience a linear matrix multiplication and then pass through a nonlinearity unit. The training data is fed into the input layer, and the PNN can be trained using feed-forward and back-propagation algorithm. The weights of the matrix w_{ij} is replaced by $(\theta_{ij}, \varphi_{ij})$ of each MZI and optimized by calculating the gradient of the loss function. The method implemented in this PNN is consistent with the training strategy for traditional electronic AI chips. Each layer of the PNN consists of an optical interference unit (OIU) and a nonlinearity unit (ONU) (Figure 5(b)). The role of OIU is to construct any real-valued matrix and complete matrix multiplication. By using singular value decomposition (SVD) method [40], any matrix M can be decomposed into

$$M = U\Sigma V^*, \tag{10}$$

where U is an $m \times m$ unitary matrix, Σ is an $m \times n$ diagonal matrix and V^* is the complex conjugate of the $n \times n$ unitary matrix V. U, Σ, V^* can be implemented using photonic programable MZI arrays. The digital electrical signal to drive the heater can be converted to analog optical intensity. For example, by tuning the phase of each MZI, the π phase shift means complete destructive interference and the associated driving voltage can be set as '000'; 0 phase shift means maximum output and the driving voltage corresponds to 'FFF'. ONU can be realized using semiconductor optical amplifier (SOA) [41] or microring resonators [42]. For an input optical intensity I_{in} coded with input information X_{in}, the output intensity is

$$I_{\text{out}} = f(I_{\text{in}}), \tag{11}$$

where $f(x)$ is a nonlinear function, such as 'Sigmoid', '(Leaky) ReLu', etc. I_{out} is detected by photodetector arrays and the analog electrical signal returns the absolute value of the output information X_{out}. Ref. [19] shows a new design for universal multiport interferometers using an alternative arrangement of programable MZIs, which has a much shorter optical depth and suffers less propagation loss compared

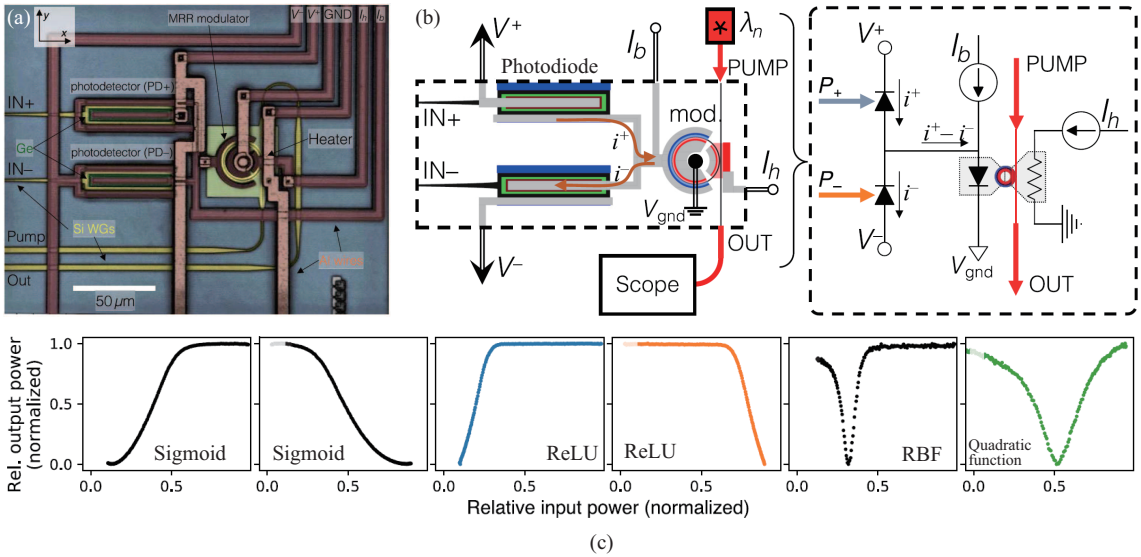

Figure 6 (Color online) (a) A micrograph image of the fabricated silicon photonic neuron. (b) Equivalent circuit diagram of the silicon MRR modulator neuron. Two photodetectors are connected to the neuron, resulting in an O/E/O nonlinear transfer function. (c) The normalized relationship between the input power and output power under different bias current I_b. Reproduced from [12] @Copyright 2019 American Physical Society.

with Reck scheme [43]. Ref. [35] proposed a gradient-based optimization method to initialize the random unitary matrices on universal photonic devices, which greatly improves convergence speed.

3.4 Silicon photonic brain-like neurosynaptic system

Compared with photonic ANNs, photonic neurosynaptic systems are the higher-level morphologies of PNNs. Photonic neurons are the fundamental elements of a brain-like PNN. An isolated photonic neuron, which is compatible with currently available silicon photonic platforms and capable of interacting with other neruons, is experimentally demonstrated in [43]. The neuron consists of a balanced photodetector connected to a tunable MRR modulator (Figure 6(a)). This photonic neuron, which behaves as a neuron participating in a network, can convert multiple weighted optical inputs into a single optical output (fan-in) and implement a nonlinear activation function to the weighted sum inputs. By showing the capacity of driving other neurons including itself (cascadability), this device shows great resemblance to physical neuron. Figure 6(b) shows the equivalent circuit diagram of the silicon neuron. The two optical inputs $(\mathrm{IN}_+, \mathrm{IN}_-)$ convert to two photocurrents (i^+, i^-) by a balanced photodetector. The output current $(i^+ - i^-)$ combined with bias I_b remodulates the new optical signal with wavelength of λ_n, which serves as the neuron's optical output. Owing to the Lorentz type transmission spectrum of MRR, the output signal is a nonlinear function of the input. The resonance wavelength of the MRR can also be tuned by an in-ring heater with a current I_h. More important, configurable optical-to-optical nonlinearity is also observed. Figure 6(c) illustrates six nonlinear function shapes under different biases. Especially the sigmoid and rectified linear unit (ReLU) nonlinear activation functions are commonly used in machine learning and convolutional neural networks (CNNs). Essentially, these nonlinear proprieties come from the modulator's transmission function. Different biases correspond to different parts of the MRR's Lorentz shape response. Besides, this photonic device is capable of inhibitory fan-in, pulse compression, time-resolved pulse processing, and indefinite cascadability, which constitutes the final piece needed to make PNNs fully integrated on silicon photonic chips.

Photonic spiking neurosynaptic networks, which mimic biological neurons and synapses, can process information more analogously to human brains in optical domain. An all-optical, integrated and scalable neuromorphic framework using phase-change material (PCM) on Silicon Nitride on Insulator (SiNOI) platform was demonstrated in [27] for the first time. PCM, such as $Ge_2Sb_2Te_5$ (GST), is commonly used in reversible optical recording medium, and exhibits a large contrast in the absorption of light between

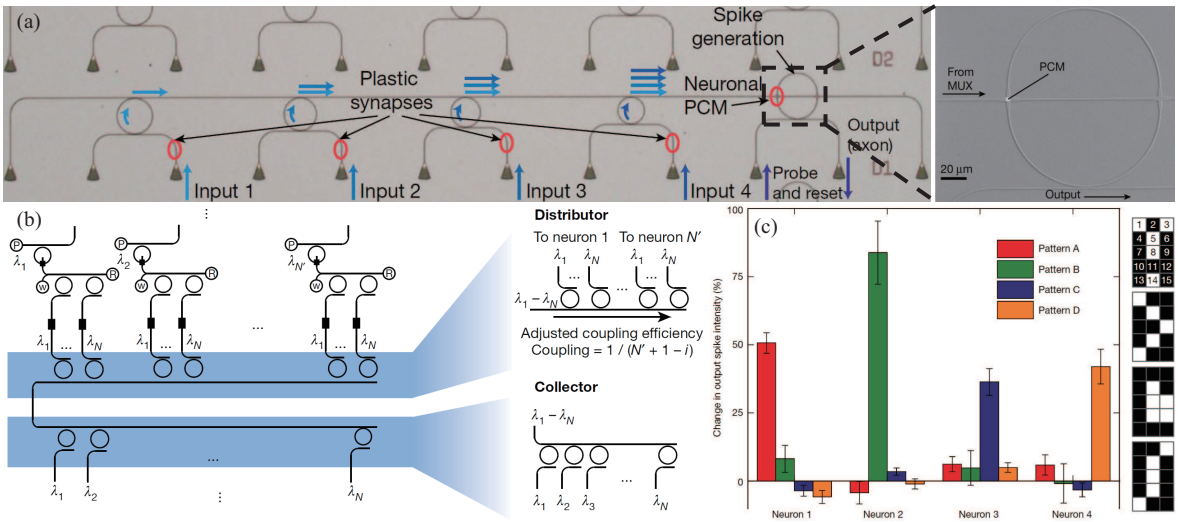

Figure 7 (Color online) (a) Optical micrograph of one complete neuron (D1) with a zoomed in ring resonator to implement activation function. (b) Schematic of a single-layer neurosynaptic system consisting of four neurons with 15 synapses each. (c) The output spike intensity of the four trained patterns (four letters A, B, C and D) illustrated on the right side. Reproduced from [27] @Copyright 2019 Springer Nature.

amorphous (low absorption) and crystalline states (high absorption) [44]. The states of the neuronal PCM cells (red circles in Figure 7(a), sputter-deposited on top of the waveguides) can be switched in a controlled manner and then the input optical signal is weighted. The weighted inputs with four different wavelengths are combined into the bus waveguide via four microrings and then propagate to the spiking neuron circuit. Only if the weighted sum of the input power exceeds a threshold, the PCM cell will be switched to low absorption state. Owing to the resonance condition variation of the larger ring, the probe pulse outputs as a neural spike. Such photonic neuron naturally emulates the basic 'integrate-and-fire functionality' of a biological neuron and can be used as a a fundamental building block for neurosynaptic PNNs. The whole PNN consists of an input, an output layer and N hidden layers. Each layer (Figure 7(b)) consists of a collector uniting the optical pulses from the previous layer using wavelength-division-multiplexing (WDM) multiplexer, a distributor with well-designed coupling efficiency distributing the input signal equally to individual neurons. A prototypical AI task of 15-pixel images pattern recognition is successfully demonstrated in the optical domain as each neuron only responds to one of the four patterns (Figure 7(c)). In addition, this PNN is capable of supervised learning and unsupervised learning as well.

4 Performance evaluation of photonic neural networks

Compared with electronics, photonics has congenital advantages—high speed, large bandwidth, and high energy efficiency. In microelectronic chips, electrons are the carrier of information transportation and most of the energy is consumed during the process of moving electrons through metal links. While photonic interconnections can transfer data at the speed of light with ultra-low energy consumption [45]. Because photons with different wavelengths do not interfere with each other, photonic chips can achieve large bandwidth density through WDM and high-speed signal transmission technology. Lately, single lane 200 G optical interconnects with silicon photonic modulator have been demonstrated [46]. If combined with dense WDM (DWDM) technology, e.g., on-chip frequency comb [47], the bandwidth density of the photonic chip will dramatically increase. Utilizing some specific optical structures, e.g., cascaded MZI [21] and meta-lens [48], photons can implement some mathematical operations (e.g., Fourier transform and MACs) at the speed of light with almost zero energy consumption. That is, when photons pass through the optical structures, the calculation process is complete. These unique properties significantly improve the computational performance of photonic chips for both energy efficiency and compute density. Therefore,

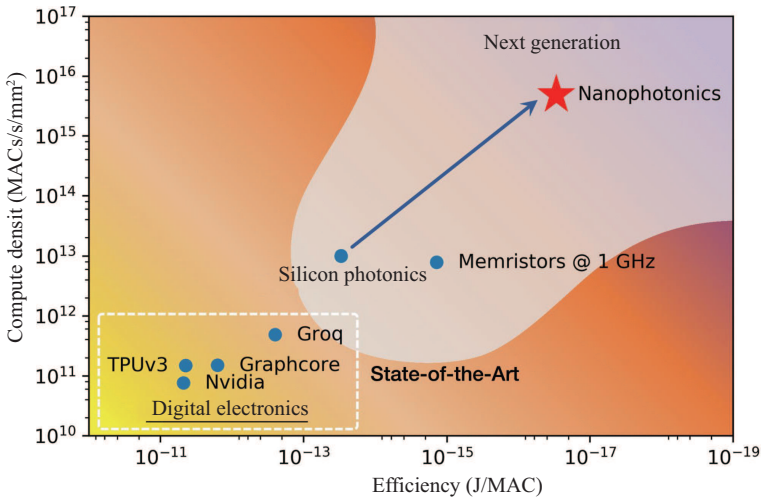

Figure 8 (Color online) Comparison of digital electronic architectures with photonic platforms for multiply-accumulate computations (MACs) which takes the form $b' \leftarrow b + w \times x$. Here b is accumulator, w is multiplier and x is input. Photonic neural networks have the potential to outperform digital hardwares. Taken from [26] @Copyright 2017 IEEE.

Table 1 Comparison of electronic AI chips with photonic neural networks. Modified from [29].

Architecture	Energy efficiency/MAC	Vector size	Latency[a)]
Google TPU [50]	0.43 pJ	256	2 μs
Flash (analog) [51]	7 fJ	100	15 ns
Hybrid laser neural networks [52]	0.22 pJ	56	<100 ps
Integrated silicon PNN [12]	2.7 fJ	148	<100 ps
Sub-λ nanophotonics (prediction)	30.6 aJ	300	<50 ps

a) Latency is the required time for completing a single MAC at the given vector size.

photonic architectures can significantly surpass electronic computing schemes, as shown in Figure 8.

The core challenge of electronic AI chips is to process massive of data at high speed with low energy consumption. A well trained ANN is able to execute what is programmed for, which is called inference. PNNs use waveguides, high-speed modulators, high-sensitivity photodetectors to implement high performance, low energy consumption computing architectures. Once the PNNs have been trained, the inference process can be passive by introducing non-volatile phase change materials [49] to maintain the phase. Then, the matrix multiplication requires almost no energy consumption. The whole energy consumption of the PNNs is mainly derived from the energy required to activate the nonlinearity unit and to obtain a high signal-to-noise ratio for the detector. Calculations show that the energy efficiency of PNNs can be at least two orders of magnitude better than digital electronic [29]. In electronic AI chips, the energy consumption is proportional to the square of the matrix dimension N. Because the optical computing process requires almost no energy consumption, PNNs is more efficient than traditional processors. In other words, the larger the neural network, the bigger advantages of using photonics.

In AI, latency is one of the important parameters to evaluate the performance of an neural network. PNNs are particularly good at reducing the latency of inference because the time required for the optical signal from input to the output is roughly determined by the speed of light. It is crucial for some applications that require fast response, such as autonomous driving and unmanned aerial vehicle. Researchers have made great efforts to develop specialized digital processors, such as GPU (NVIDIA) and TPU (Google), for calculating MACs more efficient. Recently, silicon photonics provides an alternative solution to alleviate the energy consumption of moving data via metal wires and MAC in electronic AI hardware. Table 1 [12, 50–52] gives a comparison of electronic AI chips with PNNs. Theoretically, the modulation and detection bandwidth of the PNNs is generally above 100 GHz, which is two orders of magnitude faster than the electronic ANNs. The energy efficiency of PNNs is on the order of pJ/MAC, even aJ/MAC, and the latency is below 100 ps. A concrete MAC comparison between electronic hardware

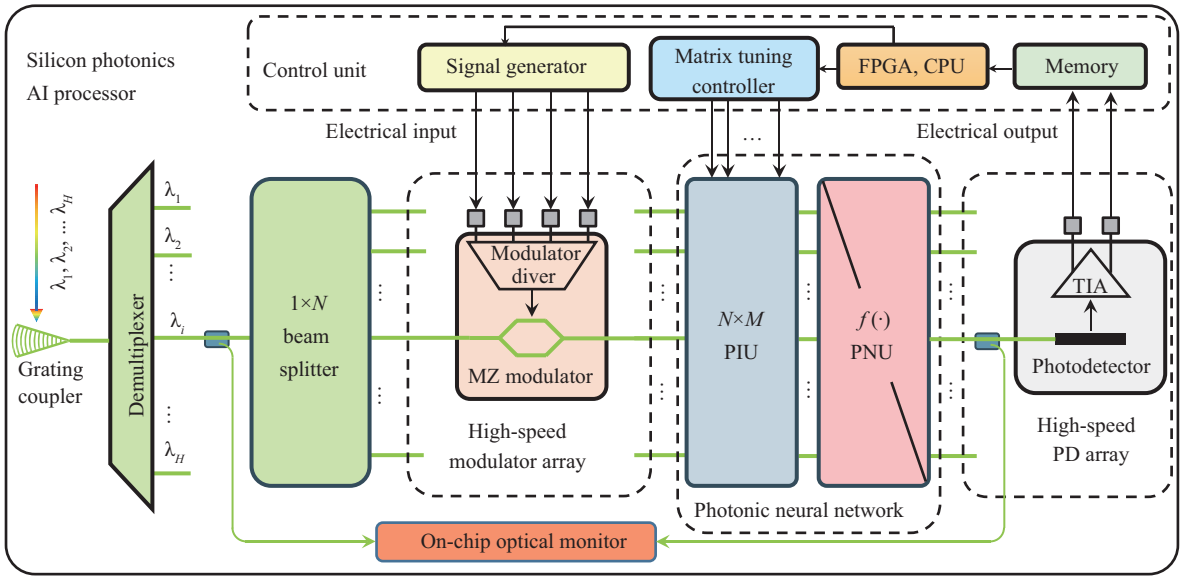

Figure 9 (Color online) Block diagram of a silicon photonic AI processor. On-chip WDM technology makes full use of the advantages of large bandwidth and parallel processing of light. The control unit allows for training the PNN.

and PNNs using several empirically validated device and system models is provided in [29].

5 Outlook for silicon photonic AI processor

Different from the general-purpose processor (e.g., CPU and GPU), silicon photonic AI processor is similar to the electronic application specific integrated circuits (ASICs), which is customized for specific computing. In AI algorithms, matrix multiplications (MACs) are the basic operations and appear everywhere. PNNs can perform MACs in the speed of light with low energy consumption, which are particularly suitable for accelerating the AI computing tasks. PNNs are appropriate for analog computation tasks (e.g., image identification and voice recognition) that the calculation results are probability distributions instead of precise numerical values. In this section, the schematic of a silicon photonic AI processor and how it can be applied to specific applications are described.

Figure 9 illustrates the internal framework of the proposed silicon photonic AI processor. The optical input with a series of wavelengths is demultiplexed by using on-chip wavelength division demultiplexer. Most of the optical energy (e.g., 99%) with specific wavelength λ_i is fed into a $1\times N$ beam splitter, followed by MZ modulators. The preprocessed input vector $(x_1, x_2, \ldots, x_N)$ are encoded in the amplitude of the optical signal via high-speed modulator array. The reconfigurable PNN consists of a photonic interference unit (PIU) that performs matrix multiplication and a photonic nonlinearity unit (PNU) that implements the nonlinear activation. The PIU can be programmed by tuning the heaters embedded in the cascaded MZ interferometers. The high-speed PD array converts the weighted, summed and nonlinearized optical signal to electrical output. The electrical I/O interface can be easily connected to the peripheral control unit or computer, increasing the portability of the processor.

To ensure the photonic AI processor is well trained and performs as expected, control unit and on-chip optical monitor are essential components. The control unit receives data and instructions from the outside and loads the preprocessed input vectors via signal generator. The matrix tuning controller gets commands from FPGA & CPU, and then adjusts the weights in PNN by manipulating the phase shifters (heaters). When training the weight parameters in PNN, the output electrical signals are stored in the internal memory for conducting back-propagation algorithm using gradient descent method. Once the training is completed, the entire computing is in optical analog domain and the results are directly output, which breaks through the Von-Neumann computing framework. Because photonic devices are sensitive to the temperature and stress variation, the information from the on-chip optical power monitor

and the memory should be also be considered in order to achieve the desired weights.

For practical applications, the fabrication variations (e.g., waveguide width variation and layer thickness fluctuation) may influence the performance of the photonic devices and result in inconsistent behavior from one device to another. Photonic neural networks are analog computation systems which are susceptible to noise and environment variation. Proper fabrication optimization and control techniques should be developed and employed. Besides, the wavelength shift, ageing of the devices, humidity of the environment, robust of the system, etc. should all be taken into account when designing the proposed processor.

6 Conclusion

Photonic neural networks, which combine the large bandwidth, high speed, low energy consumption of photonics and the efficient parallel processing capacity of neural networks, have the potential to perform ultrafast computing that exceeds electronic hardware by several orders of magnitude. Silicon photonics provides an ideal platform for large-scale photonic integration, allowing silicon PNNs to perform far more complex operations than other competitors. This paper focuses on the key photonic units and gives an overview of two specific forms of silicon PNNs, one is multilayer artificial neural networks using silicon photonic structures to realize matrix multiplication and nonlinear activation; the other is brain-like neural systems which mimics the synaptic-like connections of physical neurons. We evaluate the performance of the photonic neural networks and make a comparison with electronic AI chips. A silicon photonics AI processor for ultra-fast AI computing is also proposed as a proof-of-concept. We hope this paper provides a deeper understanding of PNNs for ultra-fast AI computing. There are many problems that need to be solved before PNNs can be implemented to real applications, including low-power, swift matrix tuning control, thermal management, monolithic integration with the digital electronic control unit, etc. Nevertheless, academia and industry have been striving to solve these problems, leading to a bright future of this emerging field.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61635001, 61822508), Beijing Municipal Science & Technology Commission (Grant No. Z19110004819006), and National Key R&D Program of China (Grant No. 2018YFB2201704).

References

1 Lane N D, Bhattacharya S, Mathur A, et al. Squeezing deep learning into mobile and embedded devices. IEEE Pervasive Comput, 2017, 16: 82–88
2 Wu N J. Neuromorphic vision chips. Sci China Inf Sci, 2018, 61: 060421
3 Davies M, Srinivasa N, Lin T H, et al. Loihi: a neuromorphic manycore processor with on-chip learning. IEEE Micro, 2018, 38: 82–99
4 Merolla P A, Arthur J V, Alvarez-Icaza R, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science, 2014, 345: 668–673
5 Graves A, Wayne G, Reynolds M, et al. Hybrid computing using a neural network with dynamic external memory. Nature, 2016, 538: 471–476
6 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444
7 Esser S K, Merolla P A, Arthur J V, et al. Convolutional networks for fast, energy-efficient neuromorphic computing. Proc Natl Acad Sci USA, 2016, 113: 11441–11446
8 Waldrop M M. The chips are down for Moore's law. Nature, 2016, 530: 144–147
9 Miller D A B. Device requirements for optical interconnects to silicon chips. Proc IEEE, 2009, 97: 1166–1185
10 Liu S H, Wang S Q, Shi W H, et al. Vehicle tracking by detection in UAV aerial video. Sci China Inf Sci, 2019, 62: 024101
11 Levinson J, Askeland J, Becker J, et al. Towards fully autonomous driving: systems and algorithms. In: Proceedings of 2011 IEEE Intelligent Vehicles Symposium (IV), 2011. 163–168
12 Tait A N, de Lima T F, Nahmias M A, et al. Silicon photonic modulator neuron. Phys Rev Appl, 2019, 11: 064043
13 Sun C, Wade M T, Lee Y, et al. Single-chip microprocessor that communicates directly using light. Nature, 2015, 528: 534–538

14 Atabaki A H, Moazeni S, Pavanello F, et al. Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip. Nature, 2018, 556: 349–354
15 Thomson D, Zilkie A, Bowers J E, et al. Roadmap on silicon photonics. J Opt, 2016, 18: 073003
16 Wang X X, Liu J F. Emerging technologies in Si active photonics. J Semicond, 2018, 39: 061001
17 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484–489
18 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, 2012. 1097–1105
19 Clements W R, Humphreys P C, Metcalf B J, et al. Optimal design for universal multiport interferometers. Optica, 2016, 3: 1460–1465
20 Ribeiro A, Ruocco A, Vanacker L, et al. Demonstration of a 4×4-port universal linear circuit. Optica, 2016, 3: 1348–1357
21 Shen Y C, Harris N C, Skirlo S, et al. Deep learning with coherent nanophotonic circuits. Nat Photon, 2017, 11: 441–446
22 Hughes T W, Minkov M, Shi Y, et al. Training of photonic neural networks through in situ backpropagation and gradient measurement. Optica, 2018, 5: 864–871
23 Chiles J, Buckley S M, Nam S W, et al. Design, fabrication, and metrology of 10× 100 multi-planar integrated photonic routing manifolds for neural networks. APL Photon, 2018, 3: 106101
24 Tait A N, Nahmias M A, Shastri B J, et al. Broadcast and weight: an integrated network for scalable photonic spike processing. J Lightw Technol, 2014, 32: 4029–4041
25 Tait A N, de Lima T F, Zhou E, et al. Neuromorphic photonic networks using silicon photonic weight banks. Sci Rep, 2017, 7: 7430
26 Peng H-T, Nahmias M A, de Lima T F, et al. Neuromorphic photonic integrated circuits. IEEE J Sel Top Quantum Electron, 2018, 24: 1–15
27 Feldmann J, Youngblood N, Wright C D, et al. All-optical spiking neurosynaptic networks with self-learning capabilities. Nature, 2019, 569: 208–214
28 Chakraborty I, Saha G, Roy K. Photonic in-memory computing primitive for spiking neural networks using phase-change materials. Phys Rev Appl, 2019, 11: 014063
29 Nahmias M A, de Lima T F, Tait A N, et al. Photonic multiply-accumulate operations for neural networks. IEEE J Sel Top Quantum Electron, 2019, 26: 1–18
30 Lee H, Chen T, Li J, et al. Ultra-low-loss optical delay line on a silicon chip. Nature Commun, 2012, 3: 1–7
31 Notaros J, Pavanello F, Wade M T, et al. Ultra-efficient cmos fiber-to-chip grating couplers. In: Proceedings of 2016 Optical Fiber Communications Conference and Exhibition (OFC), 2016. 1–3
32 Xiao X, Xu H, Li X Y, et al. High-speed, low-loss silicon Mach-Zehnder modulators with doping optimization. Opt Express, 2013, 21: 4116–4125
33 Sun J, Kumar R, Sakib M, et al. A 128 Gb/s PAM4 silicon microring modulator with integrated thermo-optic resonance tuning. J Lightw Technol, 2019, 37: 110–115
34 Vivien L, Polzer A, Marris-Morini D, et al. Zero-bias 40 Gbit/s germanium waveguide photodetector on silicon. Opt Express, 2012, 20: 1096–1101
35 Pai S, Bartlett B, Solgaard O, et al. Matrix optimization on universal unitary photonic devices. Phys Rev Appl, 2019, 11: 064044
36 Tait A N, Wu A X, de Lima T F, et al. Microring weight banks. IEEE J Sel Top Quantum Electron, 2016, 22: 312–325
37 Biberman A, Chan J, Bergman K. On-chip optical interconnection network performance evaluation using power penalty metrics from silicon photonic modulators. In: Proceedings of 2010 IEEE International Interconnect Technology Conference, 2010. 1–3
38 Bangari V, Marquez B A, Miller H, et al. Digital electronics and analog photonics for convolutional neural networks (deap-CNNs). IEEE J Sel Top Quantum Electron, 2019, 26:1–13
39 Xu S F, Wang J, Zou W W. High-energy-efficiency integrated photonic convolutional neural networks. 2019. ArXiv: 1910.12635
40 Lawson C L, Hanson R J. Solving Least Squares Problems. Philadelphia: Society for Industrial and Applied Mathematics, 1995. 15
41 Mourgias-Alexandris G, Tsakyridis A, Passalis N, et al. An all-optical neuron with sigmoid activation function. Opt Express, 2019, 27: 9620–9630
42 Coarer F D, Sciamanna M, Katumba A, et al. All-optical reservoir computing on a photonic chip using silicon-based ring resonators. IEEE J Sel Top Quantum Electron, 2018, 24: 1–8
43 Reck M, Zeilinger A, Bernstein H J, et al. Experimental realization of any discrete unitary operator. Phys Rev Lett, 1994, 73: 58–61
44 Burr G W, BrightSky M J, Sebastian A, et al. Recent progress in phase-change memory technology. IEEE J Emerg

Sel Top Circuits Syst, 2016, 6: 146–162
45 Miller D A B. Attojoule optoelectronics for low-energy information processing and communications. J Lightw Technol, 2017, 35: 346–396
46 Zhu Y X, Zhang F, Yang F, et al. Toward single lane 200G optical interconnects with silicon photonic modulator. J Lightw Technol, 2019, 38: 67–74
47 Chang L, Xie W Q, Shu H W, et al. Ultra-efficient frequency comb generation in algaas-on-insulator microresonators. 2019. ArXiv: 1909.09778
48 Jang M, Horie Y, Shibukawa A, et al. Wavefront shaping with disorder-engineered metasurfaces. Nat Photon, 2018, 12: 84–90
49 Wuttig M, Yamada N. Phase-change materials for rewriteable data storage. Nat Mater, 2007, 6: 824–832
50 Jouppi N P, Young C, Patil N, et al. In-datacenter performance analysis of a tensor processing unit. In: Proceedings of the 44th Annual International Symposium on Computer Architecture, 2017. 1–12
51 Mahmoodi M R, Strukov D. An ultra-low energy internally analog, externally digital vector-matrix multiplier based on nor flash memory technology. In: Proceedings of the 55th Annual Design Automation Conference, 2018. 1–6
52 Nahmias M A, Shastri B J, Tait A N, et al. A leaky integrate-and-fire laser neuron for ultrafast cognitive computing. IEEE J Sel Top Quantum Electron, 2013, 19: 1–12