Computer Science and Information Systems 20(1):513–535 https://doi.org/10.2298/CSIS220131066X

Efficient Neural Network Accelerators with Optical Computing and Communication*

Chengpeng Xia[1], Yawen Chen[1]**, Haibo Zhang[1], Hao Zhang[1], Fei Dai[1], and Jigang Wu[2]

[1] Department of Computer Science, University of Otago
Dunedin 9016, New Zealand
chengpeng.xia@postgrad.otago.ac.nz
{yawen, haibo, travis}@cs.otago.ac.nz
hao.zhang@postgrad.otago.ac.nz
[2] School of Computers, Guangdong University of Technology,
Guangzhou 510006, China
asjgwuch@outlook.com

Abstract. Conventional electronic Artificial Neural Networks (ANNs) accelerators focus on architecture design and numerical computation optimization to improve the training efficiency. However, these approaches have recently encountered bottlenecks in terms of energy efficiency and computing performance, which leads to an increase interest in photonic accelerator. Photonic architectures with low energy consumption, high transmission speed and high bandwidth have been considered as an important role for generation of computing architectures. In this paper, to provide a better understanding of optical technology used in ANN acceleration, we present a comprehensive review for the efficient photonic computing and communication in ANN accelerators. The related photonic devices are investigated in terms of the application in ANNs acceleration, and a classification of existing solutions is proposed that are categorized into optical computing acceleration and optical communication acceleration according to photonic effects and photonic architectures. Moreover, we discuss the challenges for these photonic neural network acceleration approaches to highlight the most promising future research opportunities in this field.

Keywords: Optical neural networks, Optical interconnection networks, Neural network accelerator.

1. Introduction

The wide applications of Artificial Intelligence (AI), such as computer vision, speech recognition and language processing, call for efficient implementation of the model training and inference phases in machine learning [61]. Especially for Artificial Neural Networks (ANNs), due to the seminal work by Hinton et al. on deep learning in 2006, ANNs have reappeared in people's vision [30]. Multiple neural networks have been studied and applied in different fields. However, with large data sets and massively interconnected

* This is an extended version of The 22nd International Conference on Parallel and Distributed Computing, Applications and Technologies.

** Corresponding author

ANNs, the traditional computer architectures suffer from the efficient training and inference due to the limited device computing efficiency and energy consumption.

To increase computing performance and energy efficiency, both hardware and software acceleration have been studied extensively in academia and industry. The specifically tailored electronic solutions have been regarded as ideally suitable for ANNs training, such as Graphics Processing Units (GPU), Tensor Processing Unit (TPU) and Field Programmable Gate Arrays [77,48]. These novel electrical architectures focus on high inter-chip bandwidth for big data traffic, memory architectures for matrix multiplications and advanced numerical calculation method to support model parallelism and data reuse. Nevertheless, the demand for computing power in ANNs is continually growing, and electronic solutions are still limited by the energy consumption of physical limits [45].

Along with the development of photonic devices and integrated optics, it has been considered a possible alternative for electronic architectures in the future to use optical architectures. There are many optical solutions emerging for communications and computing acceleration of ANNs as the times require. To this aim, some studies focus on optical linear transformations in passive optical network that enable to be operated without power consumption and with minimal latency [53], and the optical logic gates has also been proposed in different structures [34]. Further, optical devices were integrated to implement ANNs, with the aim of increasing the training speed and the energy efficiency [1,58]. For accelerating the communication of ANNs, optical on/off chip network architectures with different parallelization strategies and topologies have also been designed for decreasing ANN training workload and increasing and data transmission speed.

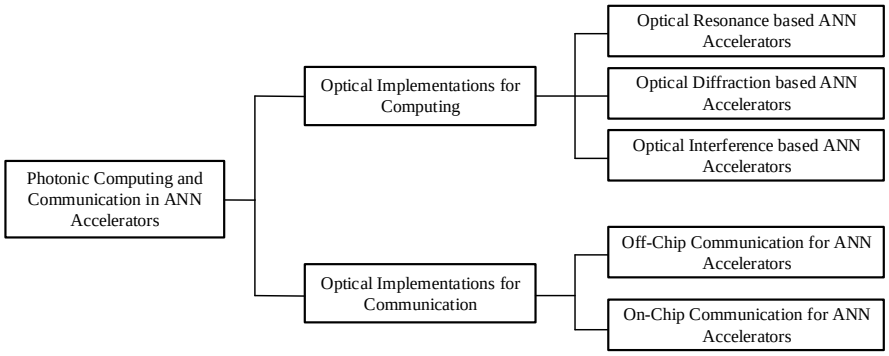


Fig. 1. Classification of Photonic Implementation in ANN Accelerators.

In this paper we present a survey of approaches for implementing optical Neural Network (ONN) accelerator. A classification of the existing solutions is proposed which includes two categories: optical computing implementations and optical communication implementations for ANN accelerators, as can be seen in Fig. 1. Existing reviews on optical ANN accelerators have either focused on reviewing performance and energy of a specific type of optical ANN computing architecture such as reservoir computing architectures

[54], [31] and Broadcast-and-Weight architectures [18], [68]), or proposed a simplified taxonomy of the realized neural network models such as CNNs, SNNs [20]. By comparison, we present a different and more comprehensive review of photonic ANN from two aspects including computing acceleration and communication acceleration approaches with a bottom-up classification across design-layer abstractions: from lower-level optical devices, to the neuron microarchitectures, and covering a variety of integrated neural network.

Recently, some works have focused on the computing acceleration in neural network and on bottlenecks of photonics technologies [42], [74]. However, these works have ignored the contribution of on-chip optical communication to neural networks acceleration. Compared to our previous work [72], this paper provides the review of optical devices from lower-level, and a more comprehensive summary of optical computing and communication acceleration neural networks. In addition, the advantages and disadvantages of existing works are summarized by comparing literature.

The remainder of this paper is organized as follows: In Section 2, we present the motivations behind ANN accelerator, and introduce a taxonomy of the approaches presented in the literature. The relevant optical devices is reviewed in terms of the application in ANN accelerators. In Section 3, we describe the optical architectures devised for the computing implementations in ANNs. In Section 4 the most relevant solutions are reviewed according to the categories of optical communication for ANNs acceleration. While in Section 5, we discuss the challenges and future research opportunities in this field, and Section 6 concludes the paper.

2. Background

2.1. Optical Neural Network Accelerators

The exiting researches of ANN accelerators are mainly focused on the development of electronic architecture specially tailored to neural network, such as GPU, TPU, FPGA and ASIC optimized neural network by adjusting the computing architecture. However, after decades of prosperous development of electronic computers, the current silicon-based computer circuits are reaching their physical limits [16]. Conventional high-performance computer architectures still cannot break through the bottleneck of the memory wall, and computing performance is limited by bandwidth and huge date processing workload and power consumption [14].

In recent years, some researches began to explore analog electronic circuits to address the memory wall challenges that can meet the ANN computation requirements. Quantum Neural Networks [23], Processing-in-Memory [13] and Memristor [2] have been specially designed to implement ANN acceleration. Processing-in-Memory employs the memory arrays themselves for computing to reduce the movement of data between CPU and memory, which obviates memory cell redesign and has low area overhead and friendly manufacture. However, The accuracy of Processing-in-Memory is limited by analog calculations [32]. Memristor-based Accelerators mainly consist of resistor array with memory and analog crossbars. The main drawback of Memristor is the concern on the high power consumption [71]. While all these three technologies lack mature development platforms and industry standards.

Photonics have demonstrated great potentialities for various application in ANN accelerators. In order to implement the functionality of neural networks in photonic networks, the present research efforts have been undertaken numerously. In comparison with state-of-the-art electrical architectures, the optical solutions are expected to enhance computational speed and energy efficiency when performing data transfer and training tasks [58]. The rationale behind optical architectures can considerably decrease the energy cost both in logical calculation and data transmission by using passive optical network to execute the linear operations in a typical ANN [28]. The application of passive components in an integrated optical circuit enables high-speed operation while consuming less than the transmitter and receiver energy limits. Hence, analog optical computing circuits are an exciting research field possibility for high-performance computing, especially for ANNs. On the other hand, optical interconnection networks have been well studied by a large amount of works, due to its unique advantage in data transmission such as high bandwidth density, low power consumption and immune to electromagnetic effect [49]. Optical network with the implementation of Wavelength-Division-Multiplexing technology (WDM), is also suitable for neural networks to accelerate on-chip or off-chip communications [17].

2.2. Photonic Devices

Over the past few years, optical communication has been generally applied in remote communications and data centers for cost-effective and high bandwidth interconnects. With the development of silicon photonics, photonic devices are becoming miniaturized and low power consumption that makes it increasingly possible to integrate photonic network architectures. The integrate optical devices are considered to develop optical computing chips with better performance. The computing platform using Wavelength-Division Multiplexing technology can carry more than 64 wavelengths of light in a single waveguide. Each optical signal enables wavelengths to carry different data at high transmission speed without any crosstalk [4]. In optical network, the multiplexed signals are switched and separated by the Microring Resonators, Mach-Zehnder Interferometers and other optical devices to achieve optical communication and computing. The optical signals are eventually converted into electrical signals for storage or further processing by photo detectors, optical to electronic converters and other devices. However, there are still several challenges for optical devices to support robust computation and communication at chip scale. Therefore, this paper reviews some basic optical devices that are used for computing and communication in optical acceleration.

Lasers The first challenging requirement for optical architecture is to develop an effective and stable chip-scale optical source. There are two types light source in the integrated optical circuit: on-chip laser and off-chip laser, and the different structures lead to different advantages and disadvantages for optical interconnections [81]. The off chip lasers can offer excellent luminous performance and stable temperature control but limited by high optical power loss because of the coupling. The on chip lasers could possibly offer a higher integrated ability and lower power loss, whereas the development of on chip lasers is limited by the low emission efficiency of silicon that is one main obstacle preventing the integration of optical interconnection [29,82].

Light sources are emitted by several ways. Firstly, Vertical-cavity surface emitting lasers (VCSELs) are now key optical sources in optical communications which the lasers

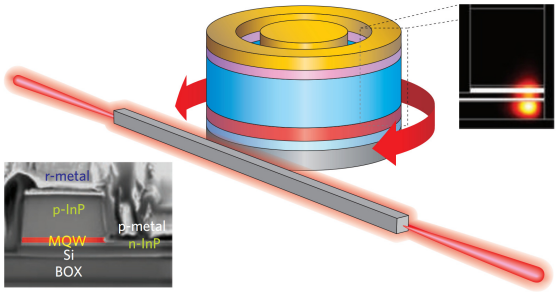


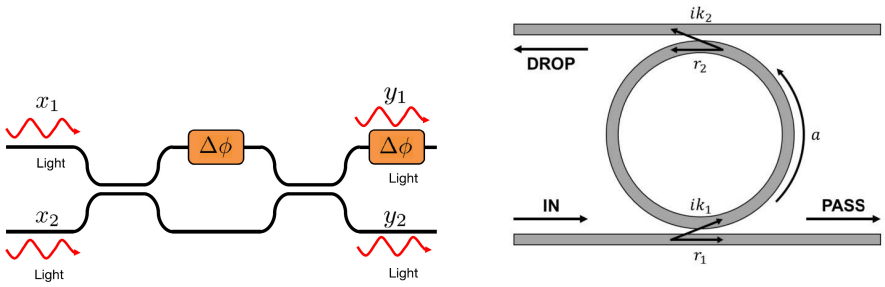
Fig. 2. A hybrid microring laser with a Si bus waveguide.

are perpendicular to the surface of substrates. VCSELs are widely utilized to the optical neural network as they meet the requires of cost-effective, integratable for array and high coupling efficiency to waveguides [73]. Meanwhile, hybrid silicon lasers with small size and a short cavity structure are verified have greater footprint efficiency. Secondly, as shown in Fig. 2 the hybrid microring lasers have been experimentally demonstrated on Si integrated platform that be contained about 400 laser devices in $1cm^2$ chip [64]. In the optical computing, lasers are employed to implement some functions of neural network. Light sources with integrated modulator can directly output light to carry data by adjusting the amplitude, power and phase of light [63]. However, the thermal impedance caused by high density lasers lying is still a major hurdle for large scale integration.

Mach-Zehnder Interferometers Mach-Zehnder Interfeometer (MZI) is broadly applied to develop optical modulators, switches and filters in photonic architectures [80,65]. A MZI is composed of two beam splitters and two phase shifters. Fig. 3(a) shows the layout of a MZI device. While the fixed 50:50 beam splitters are not configurable, the two phase shifters are configurable by adjusting the angle. The input optical signal is split proportionally as it passes through the beam splitter. By applying power to the two phase shifters, the MZI can be controlled to provide phase shifting or attenuation for the optical signals which pass through the two arms. This enables MZIs to work as the directional couplers or more simply as the optical switches. Based on the above functions, MZI has been utilized to realize fundamental logic operation by optical power level, phase adjusting and output port scheduling. In [6], authors designed a cascaded MZIs structure with interference of multiple beams that obtained various gates by detecting optical intensity at different points such as AND, OR, NAND, XNOR and so on. According to the research above, MZI-based logic gates are studied to support optical neural network computation [55]. The cascade MZIs array is used as the basic unit of matrix multiplication, which has been attracting a great deal of attention recently. The singular value segmentation is used to make it suitable for matrix multiplication operation in ANNs [58]. Such neural functions and their implementations are discussed in the next section.

Microring Resonators A Microring Resonator (MRR) can be seen as a closed waveguide circular and a common structure of MRR is that two bus waveguides border upon a ring waveguide. The ring waveguide resonates when the path-length of resonator cavity is

equal to the integer multiple of the input wavelength [7]. As shown in Fig. 3(b), the MRR contains a ring waveguide, an input bus waveguide and a drop bus waveguide. When the resonance occurs, the input lights will be passed the ring waveguide and turned to the drop waveguide, conversely, the lights will be routed to the pass waveguide. Hence, the MRRs are used as switches or filters in photonic communication, especially for Wavelength Division Multiplexing technology [76]. The WDM technology arrows optical signal with different wavelength to carry more data without interference, which increases parallel process in the optical network. Moreover, MRRs and WDM optical signal are employed to realize weight accumulation in optical neural network, in which the cascaded add-drop MRRs is grid in shape, and is called weight bank [67,78]. To sum up, MRRs and MZIs are designed as switches, filters and modulators in optical communication and computing. Compared to the MZI, since the dropped output of the MRR can be directly monitored at each wavelength of the WDM signal, it is more straightforward to set the elements of weight than using meshed-MZI [50].



(a) Mach-Zehnder Interferometer (b) Microring Resonator

Fig. 3. The structures of a Mach-Zehnder Interferometer and a Microring Resonator.

Photodetectors Photodetectors are commonly used to detect optical signals and convert optical signals into electrical signals. Therefore, the optical detector can be used as an optical signal output device to be connected at the end of the optical network [38]. In addition to being the photo-to-electric conversion devices in optical neural networks, photodetectors are also used to realize the operations of ANNs. In [59] the positive signals and negative signals are propagated and superimposed on different optical waveguides, and two balanced photodetectors are set at the end of the waveguides to detect the total optical power in positive and negative waveguides, respectively. The detected lights are converted into the currents by the photodetectors. The balanced photodetectors can calculate the difference between positive and negative currents, which enable optical neural network to realize the accumulation operation for different kernels.

The performance of a photodetector is related to the responsivity which is defined as the optical power that can be detected per unit area. An efficient photodetector with a lower responsivity have better detection accuracy for low-power input light sources, and a photodetector with high responsitivity needs higher power consumption to meet its detec-

tion accuracy demand. For driving the photodetector effectively, the optical power reaching the photodetector should be greater than the responsivity. The input optical power, energy loss and responsivity are considered to be the critical factors when choosing an input light source. Therefore, the performance of photodetector affects the bandwidth and power consumption of optical networks.

3. Optical Implementations for Computing

This section summarizes optical implementation for ANN computing acceleration. The innovation of these works mainly focus on constructing different types of neural networks, such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Spiking Neural Network (SNN) and Multi-layer Perceptron (MLP). The use of optical fundamental principles and photonics components makes matrix multiplication available. The performance of silicon photonics ANN accelerator is also affected by the optical principle and architecture. Hence, these implementations will be categorized according to the primary photonics principles.

3.1. Optical Resonance based Neural Network Accelerators

Inspiration comes from the field of neurobiology in which each neuron communicates in the way of short pulses, the resonance-based photonic neural networks have been investigated widely. Wavelength division multiplexing is applied in optical neural networks depending on the resonance modulation property and wavelength specificity of MRRs. The WDM channel transmits multiple wavelengths in the same waveguide without interference, which reduces the number of optical devices in ANNs implementation to some extent. In [66], the authors integrated several MRRs to design an optical neural network with on-chip architecture that is called Broadcast-and-Weight (BW). The input signals are transmitted parallelly in a bus waveguide, and the weights of neurons are loaded into the MRRs. Fig. 4 shows that the multiple wavelengths are aggregated in a waveguide by multiplexer, and the MRRs act as the neurons. While the passive splitters are employed at the end of the bus to broadcast data, so the output signals of the bus is connected to all neurons. The weight bank is actually a set of reconfigurable filters composed of MRRs, which can map the weight to different network layers by attenuating the resonance wavelength.

Based on the BW protocol, authors presented an optical convolution neural network accelerator (PCNNA) in [46]. The PCNNA can propagate different CNN layers in a same optical circuit because of the using of single layer multiplexing architecture. The authors considered that convolution calculations of different kernels can be executed in parallel because each layer of PCNNA shares the same convolution kernel value. In the overall framework, the PCNNA is configured to run on two clock domains with different speeds because the optical circuits runs faster than the electronic circuits. Thus, convolution results and kernel weights of each layer can be stored in an off-chip Dynamic Random Access Memory (DRAM). Kernel weights are loaded into the ANN by tuning microrings in the MRR banks. The authors showed how their accelerator implements AlexNet, and they claimed a third-order cubic polynomial time of magnitude execution time improvement over electronic engines. An extension of WDM for CNN implementation was

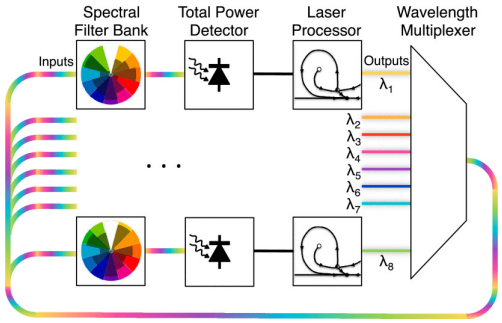


Fig. 4. The Broadcast-and-Weight architecture proposed by [66].

explored in [60] where authors combined MRR and MZI to design an all-optical multiplication and accumulation. Based on the optical resonance, an array of tuned MRRs is utilized to realize WDM and optical *AND* operations. The cascade MZIs can operate pure optical shift accumulation on each sequential *AND* operation.

The optical resonance is also widely used in the implementations of SNN. In [66], the spiking neuron unit is named processing network node (PNN), and each PNN is composed of weight banks, photodetectors (PDs) and laser diodes (LDs). The weights are divided into excitatory and inhibitory weights that are stored in the weight banks. Different types of weights are received by two PDs to complete the accumulation of PNN. Finally, a broadcast loop is used to propagate the WDM signals between these PNNs. In addition to the BW architecture, authors in [10] mentioned that the integration of MRRs and PCM material $(\mathrm{Ge_2Sb_5Te_5})$ were used to fire neurons. The bidirectional multi-ports integrating action of MRR was used to figure up the film potential in the effect of the weighted sum. Spikes output when the film potential of a neuron exceeds the thresholds. There is a few works on optical resonance implementation for reservoir computing (RC). A 4×4 swirl topology-based reservoir was proposed in [21]. The work utilized MRRs and basic Boolean operations, in which non-linear elements (MRRs) are the nodes of the recurrence network. The input signals/weight matrix is mixed in the swirl to realize computing.

3.2. Optical Diffraction based Neural Network Accelerators

In optical network, diffraction effects tend to limit the performance of optical units, while efficient optical neural networks can be realized by designing diffraction-based architectures using appropriate optical elements. For example, a holographic optical element (HOE), which is usually used for information storage, can be utilized to store weights and propagation directions in neural network [51]. Using HOEs, authors in [84] explored an all optical neural network. Firstly, according to the number of input optical signals, the spatial light modulators are divided into several regions, and the holograms of input signals can be obtained by superimposing phase gratings in front of each region. The optical matrix multiplication is implemented by diffraction of the optical signal in HOEs, where the weights of the ANN are loaded in the direction of the input optical signals. The convex lens are set behind the HOE to perform Fourier transform on the diffracted signal. Finally, all signals are output to the receiving plane to complete the accumulation operation.

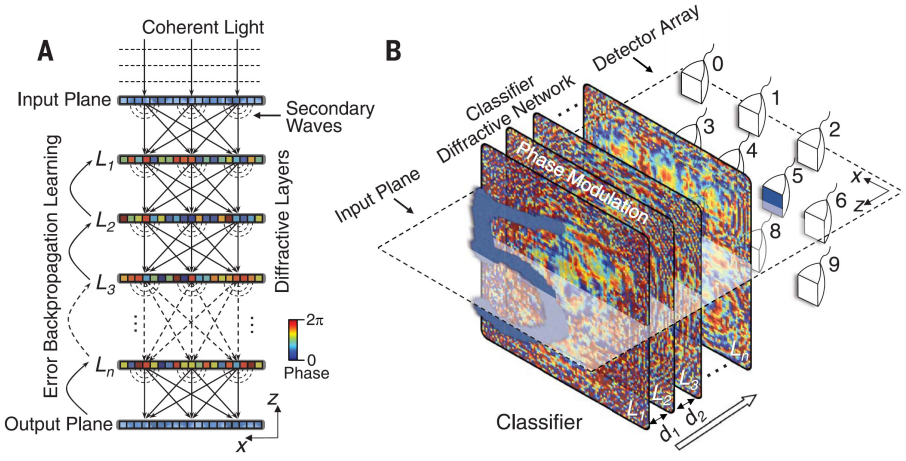


Fig. 5. Diffractive deep neural networks (D^2NN) depicted by [40].

Apart from holograms, an optical ANN with cascade phase mask structure was proposed in [40], which is called D^2NN. As illustration of Fig. 5, the fully connected layers consist of several sequential and hierarchical phase masks. These phase masks are 3D printed, and each mask represents a layer in the fully connected network. The grids inside the mask represent different neurons. The refractive index of the grids can be changed by setting different thickness. Hence, the D^2NN maps the weights of neurons to the state of grids and phase of masks. When the input optical signals pass through the mask, the matrix multiplication will perform because of diffraction effect. The output signals will directly enter the next mask that represents a direct fully connection with the next layer in ANN. Therefore, the cascade mask array forms a multi-layer fully connected optical neural network with variable weight. Finally, an array of detectors in the last mask is deposited to measure the intensity of the output light, which can be defined as the classification results of D^2NN. Specifically, Lin et al. [40] assumed that the light wavelength is λ, the size of neurons is usually about 0.5λ, and the axial distance between phase masks is usually set as about 40λ, such that 200×200 neurons can be packed in an area of $8 \times 8cm^2$ each layer with an axial distance of $3cm$ between layers. Considering there are five layers, approximately 8 billion connections were implemented. The feasibility of D^2NN network has also been confirmed by microwave [52] and broadband incoherent light source [44] experiments. Whereas, this architecture would be more sensitive to the assembling errors, which causes it difficult to manufacture.

To the best of our knowledge, there have been no studies of the SNNs implementation based on diffractive optics. Apart from SNNs, some works focused on the realization of the diffractive-based Reservoir Computing. A reservoir used a 4×4 swirl topology was discussed in [31], where the readout layer is composed of a nonlinear optical modulator. The authors in [21] also proposed an RC architecture using pillar silicon scatterers and cavities as passive elements. The work of [8] described an all optical large system that used digital micromirrors for diffraction in its output layer. However, due to the nonlinearity of electrical domain, the update rate is severely limited to 5Hz. This study demon-

strated a system with 2025 nonlinear nodes, and implemented in the form of pixels in a spatial light modulator. The SLM will show the status of the reservoir in the form of a specklegram that will be received by a camera, and then calculate the following step required for the reservoir and encode it into the system. Finally, the optical RC system is demonstrated through experiments. The system realizes that the random interconnection between neurons in the reservoir is a random diffraction behavior through the diffractive optical element (DOE). The authors believed that random diffraction has been proved to be suitable for optical nanocrystals. Nevertheless, diffraction-based ANNs are mostly realized by free-space optic devices, which take up a lot of space and do not support large-scale neural network acceleration.

3.3. Optical Interference based Neural Network Accelerators

Different from diffraction with a large amount of beams input, interference-based optical computation requires only a small number of lights to carry information, and the lights need to transmit through the waveguide. Optical matrix multiplication of neural network based on interference is mainly realized by cascading MZIs, in which the interference of light is carried out in the directional coupler and phase shifter in MZIs. The phase, amplitude and power of the input optical signals can be changed to achieve weight loading in the initial data by adjusting the couplers and phase shifters. The authors presented a pioneering work in [58], which a coherent nano-photonic circuit is designed for all-optical neural networks, and this work lays the foundation for future interference-based neural network accelerators. As shown in Fig. 6, the optical matrix multiplication between input data and weight is realized by singular value decomposition (SVD) [37]. Specifically, the matrix M is decomposed into $M = U \sum V$, where U and V are two unitary matrices, and $\sum$ is a diagonal matrix. Accordingly, MZIs are assembled as a cascaded array with three segments that implement the matrices U, $\sum$ and V, respectively. The cascade MZIs array represents the fully connected layers of ANN. If an input optical signal passes through MZI, two parallel lights will be applied to the two phase arms of MZI, and then the occurrence of parallel light interference will realize the matrix accumulative operation of ANN. The weight information is loaded into optical neural network by interference effect, and the weight can be changed by adjusting the shifters and amplitude of the coherent wave. Hence, the energy consumption is low in the entire training process. However, the depth of the ONN in [58] is limited to $2N - 1$ affected by light attenuation, which also means that the reduction in the size of the ANN may reduce the classification accuracy. Therefore, the authors in [22] improved unitary matrix multiplier, controlled the depth of MZI layer by cluster grid and statistical Fast Fourier Transform (FFT). The authors claimed that FFT-based network is inherently more robust than grid-based network, because it has a much smaller number of MZI layers for realizing a same unitary matrix multiplier. For example, a multiplier can be realized by using a FFT network with only $log_2(N)$ layers instead of the grid network with N layers, that means the depth of grid network is 32 times that of FFT network when $N = 2^8$.

Due to the area of MZI device and the large demand for nodes in the RC network, the interference effect has not received much attention in the research of RC. Authors in [36] provided an electrical-optical nonlinear modulation transfer function by using the delay coupling technology integrated MZIs. A long distance optical fiber is used to realize a

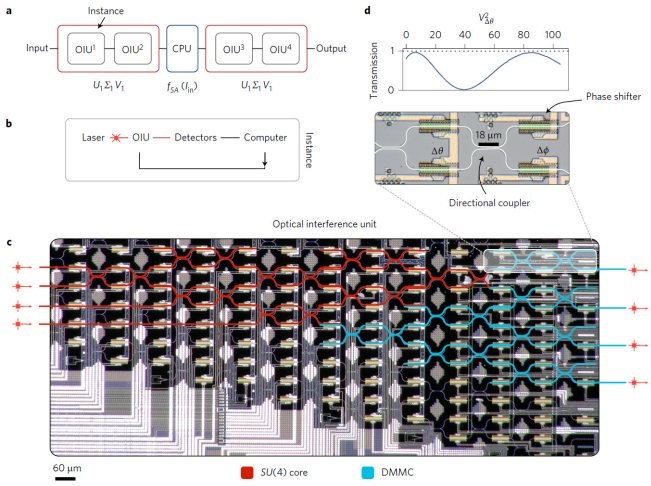


Fig. 6. Interference based photonic integrated circuit depicted by [58].

delay feedback loop, and a photodiode is used for optical detection. The electronic feedback circuit output is connected to MZI input electrode. Therefore, long distance optical fibers are divided into a number of subintervals to define virtual nodes. By extracting the virtual node state at the end of each subinterval, [36] simulated the nodes of conventional RC network.

3.4. Summary

The literature concerning photonic neural network architectures is vast, and so are the techniques and devices used to realize these architectures. In this section, we reviewed different architectures and divided the literature into resonance-based implementations, interference-based implementations, diffraction-optics based implementations. We have provided a summary of the literature on architectures covered as part of Section 3 in Table 1. The table has the devices prominently used in the architecture (first column); a brief summary of the advantages (second column); a brief summary of the disadvantages (third column); the references to the works (fourth column).

4. Optical Implementations for Communication

Existing ANNs have been challenged by the fact of high computational complexity, large amount of computational data, frequent memory access and high parallelism requirements that are widespread in current neural network workloads. In the latest ANNs, tens to hundreds of megabytes of parameters are required to execute a single inference pass. Over one billions of operations will generate large amounts of memory access requirements from the processing elements (PE) which makes existing architectures face the challenge of memory wall. In the processing of model training, a large amount of reusable data are usually generated. For example, a huge amount of filter data, input feature map data

Table 1. A summary of selected proposed optical ANN using computing and communication acceleration

	Implementation	Advantages	Disadvantages	References
computing acceleration	Use MRR weight banks for synapse; photodetector for optoelectrical conversions	Use WDM to offer high bandwidth; Use passive devices compatible with low power consumption	consumes power in electro-optical and optoelectrical conversions; low cascadability; sensitive to temperature	[66], [10], [60]
	Use HOEs	Use passive devices compatible with low power consumption, low energy loss	sensitive to the assembling errors, difficult to manufacture; area inefficient	[40], [84], [52]
	Splitters and MZIs	Higher reliability and low power overhead in comparison with using several wavelengths	Low bandwidth because of using one wavelength; area inefficient; exact splitting ratios are hard to achieve after fabrication due to variations; susceptible to noise in phase and splitting ratios	[36], [58], [22]
communication acceleration	use optical circuit switch-based topology for data reusing	high workloads parallelism; demand of ANN workloads and the singleness and repeatability of transmis-sion in workloads,		

and partial sum data are created in the processing of convolution in CNN, in which these data can be regarded as reusable resources. Hence, the CNN accelerators using dataflow optimization with Network-on-Chip (NoC) architecture have been proposed in [9] which has good acceleration performance and system efficiency.

However, electrical signal based ANN accelerators with NoC architecture still face the challenges of energy consumption and time delay. To break the communication bottleneck, recent advances in CMOS-compatible optical devices have suggested that optical networks on chip (ONoC) could be a promising solution [41]. In contrast to electrical NoC, the data is is transmitted in the optical domain via waveguides, which has lower power consumption and higher performance than electrical signal transmission. This makes ONoC uniquely capable of performing data-intensive and high-throughput off/on-chip communications, in which needs a huge data movement among processing units or chips to accelerate parallel processing of ANN workloads.

4.1. Off-Chip Communication for Neural Network Accelerators

The research on photonic interconnection in datacenter has a long history. To improve communication performance, exist works have showed the improvement of communica-

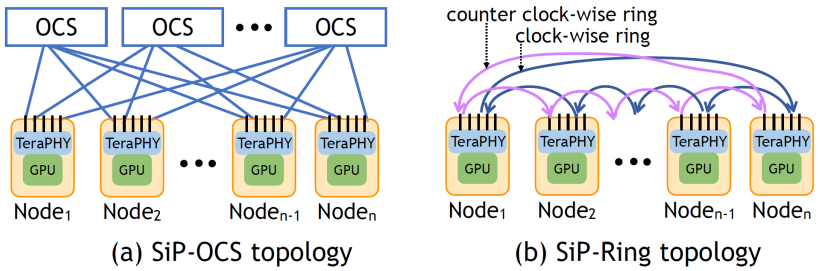


Fig. 7. Two topologies for SiP-ML proposed by [33].

tion performance in datacenter networks by designing photoelectric hybrid connection in reconfigurable topologies [24,47] or designing all-optical interconnects [3,11]. Whereas, compared with the optical technology in computing implementation [79], only a handful of researches focus on using photonic interconnection to accelerate ANNs communication. Authors in [33] presented all photonic interconnects for ANNs acceleration named Silicon Photonic Machine Learning (SiP-ML) which has powerful scalability of ANN training jobs by using SiP chips. The authors argued that ANN training jobs are predictable and periodical that include mostly large data transfers instead of unpredictable behavior and short data flow workloads in conventional datacenter. Authors considered the parallel demand of ANN workloads and the singleness and repeatability of transmission in workloads, and then explored two data-reusing topologies. As illustration of Fig. 7, an Optical Circuit Switch (OCS) based topology called SiP-OCS is designed with commercially available optical switches. Each OCS is linked to each GPUs by port in a flat topology. By setting a 10ms reconfiguration delay, Sip-OCS can transfer data infrequently across the ANN workloads. Furthermore, a switch-free topology without any switching elements was proposed by embedding MRRs in SiP ports named Sip-Ring. As a filter, the activated MRR enables to replace the switch by selecting and forwarding optical signals. Compared to SiP-OCS, SiP-Ring can reuse the signals in non-overlapping portion of MRR, and can reconfigure the resonant wavelength in each port to enrich logically topologies. However, with the increase of process units, the communication costs of large ANN model training increase greatly in SiP-ML, and the communication mapping performance of ANN training is affected by the parallelization strategy and AllReduce.

Motivated by the optical switch based topologies in [33], authors in [70] proposed a co-optimizes network topology and a parallelization strategy for ANN training system named TOPOOPT. The proposed scheme searches over the parallelization strategy space with a fixed topology, and returns the communication demands to the system. A topology is then reconfigured with the searched parallelization strategy, which enables system to alternate between optimizing the parallelization strategy and optimizing the network topology. This looped process helps system to find an optimized parallelization strategy and an optimized topology. The TOPOOPT system is an optical shareable interconnect, in which interfaces of server are connected to the processing unit layer by optical switches. The optical switches enable to achieve the target topology by partitioning the cluster dedicated partitions for each training workload. When the system finds an optimized parallelization

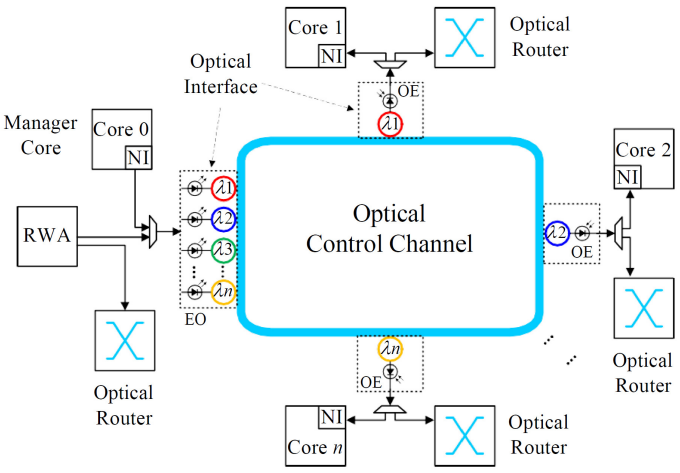


Fig. 8. The optical control channel in [17].

strategy or network topology, interconnection between the server and the processing unit can be changed to the corresponding topology by optical switches reconfiguration.

In addition to all optical network architecture for ANNs workloads, authors proposed a hybrid optical/electrical network architecture that optical switches are employed to offer long-term ANN training communication in [69]. Zhu et al. [83] proposed a silicon photonic reconfigurable architecture with a fat tree topology, which optical switches are applied to link top-of-rack interfaces and aggregate electronic packet interfaces. The optimized optical switches control scheme was designed to reduce the complexity of control implementation, which enables optical switches to apply for large-scale systems integrating. The experiments in hardware test platform show that the silicon photonic reconfigurable architecture can execute ANNs training jobs efficiently. A similar distributed ANN training application with fat topology was proposed in [27], where the experiments were built in commercial rack servers to test the performance of optical switch based distributed ANN acceleration.

Furthermore, authors in [75] proposed an Inter/Intra-Chip silicon photonic network for rack-scale computing systems. To void the challenge of photonic buffering, the architecture employs circuit switching for the ONoC that can also avoid the large overhead in optical devices assembly and unloading. [75] utilized the inter-node interface as the medium to coordinate the request from both local ONoC and optical switch. A channel partition and dynamic path priority control scheme is designed to reduce the control complexity and arbitration overhead. Feng et al. in [25] proposed a variant architecture that is optimized by floorplan optimized delta optical network switch architecture and the preemptive chain feedback scheme. In [43], an arrayed waveguide grating routers based hybrid optical-electrical architecture was proposed, in which a complete bipartite graph is employed to enhance the transmission bandwidth and interconnection scale of machine learning.

4.2. On-Chip Communication for Neural Network Accelerators

In [35], the authors considered that electrical interconnection in the existing manycore platform would not be sustainable for handling the massively increasing bandwidth demand of big data driven AI applications. Hence, a rapid topology generation and core mapping of ONoC (REGO) for heterogeneous multicore architecture was proposed. Based on the genetic algorithm, REGO receives an application task graph including the number of cores and ONoC parameters as inputs, which further includes the available router structure, loss and noise factors of the optical elements. Thus, the REGO can accommodate various router structures and optical elements because it calculates the worst-case OSNR through loss and noise parameters obtained in advance through the parameters of optical.

A fine-grained parallel computing model for ANNs training was depicted in [17] on ONoC, in which the trade-off between computation and communication can be analyzed to support the ANN acceleration. As shown in 8, The optical control channel was designed to configure the state of cores and optical routers. To minimize the total training time, three mapping strategies were designed in each ANN training stage which has the optimal number of cores. The advantages and disadvantages for each mapping strategy are discussed and analyzed in terms of hotspot level, memory requirement and state transitions. Furthermore, an optoelectronic hybrid on chip architecture was demonstrated for CPU and GPU heterogeneous systems in [12], which the first layer is connected by waveguide, the second layer is a electrical mesh with 8×8 nodes, and all layers are linked by the through-silicon-via. The proposed architecture utilized the reservation-based single write multiple reader bus to reduce the number of optical switches that can reduce energy consumption.

Table 2. Challenges and Future research directions

Scalability	Low complexity architectures
	Noise resilient optical devices
	Low loss optical devices
Robustness	Effective photonic crosstalk mitigation
	Phase noise correction
	Noise resilient photodetection
Design Space Exploration	Parallelism of models
	Devices reuse Architectures
	Data reuse topologies
Optical Nonlinear Activation	All-electronic nonlinear activation function
	Photoelectric hybrid nonlinear activation function
	All-optical nonlinear activation function

5. Challenges and Opportunities

In this paper, we reviewed the optical approaches to accelerate neural networks from two aspects, i.e., computing and communication. In recent years, with the maturity of ANN theory and the development of silicon optical technology, one of the areas with growing concerns is the implementations of optical ANNs. Meanwhile, the addition of some sophisticated optical devices, such as optical frequency comb [57], makes it possible for accelerators to train ANN models at extremely high speeds. Nevertheless, there are still some outstanding challenges that limit the inference accuracy, reliability and scalability of optical ANNs. Hence, we summarize the challenges and opportunities to offer suggestions for future research, as shown in Table 2.

Scalability: The exiting works that have been discussed in this review mainly focus on three approaches to accelerate ANNs model training, which are small optical neural network implementation, matrix vector multiplication acceleration and optical network architectures for communication accelerating. The two major issues of the above approaches are high area consumption and energy attenuation of the optical devices. The schemes in [53] and [15] described that the optical depth (the number of MZI units traversed through the longest path) for the unitary matrix is limited to $2N - 3$ and N, in an ANN with N number of neurons, respectively. Therefore, the optical depth of singular value decomposition encoding based ANNs is also limited to $2N - 1$ and $2N + 1$, in which the diagonal matrix is realized by MZIs. The optical depth is positively associated with the number of layers in ANNs that will cause the additional loss as the optical depth increases. The additional loss could exceed the budget and dramatically increase the ratio of signal to noise, which will reduce the computing accuracy in an optical network with limited power consumption. Therefore, more studies are expected to design photonic devices that are noise resilient and low loss to improve scalability for the large scale ANNs, and design novel architectures to reduce the optical integration complexity.

Robustness: With the optical integration scale up, robustness is becoming an important factor for system stability. For example, the performance of MZI-based computing architecture is directly affected by crosstalk, environment temperature and manufacturing process, in which the slight phase change will cause a cascaded calculation error. The experiments in [58] showed that the accuracy in small optical ANN outperform that of large scale optical ANN about 20%. Moreover, the smaller ANN also shows better robustness if added signal noise on optical devices. Whereas the on-chip thermal crosstalk can be suppressed, the finite encoding precision on phase settings will remain as the fundamental limitation for the optical ANNs with high computational complexity. The phase errors, in particular, accumulate when the lightwave signal traverses the MZI mesh with an optical depth of $2N + 1$. In addition, such errors propagate through each layer of the network, which ultimately restricts the depth of the neural network. In order to realize robust photonic accelerator, research is needed to achieve effective photonic crosstalk mitigation, phase noise correction, and noise resilient photodetection.

Design Space Exploration: Early demonstrations of photonic solutions for ANN and RC acceleration were implemented with bulky free-space optics [39], which have strict requirements for accurate phase matching and great difficulty for optical devices footprint reducing. Even in recent optical neural networks based on singular value decomposition encoding, a $m \times n$ weight matrix needs the number of $m(m-1)/2 + n(n-1)/2 + \max(m, n)$ MZIs to realize. This hardware complexity can limit the actually implemen-

tation scale of optical ANNs, especially when the size of an MZI reaches up to 100 μm. Moreover, The extensive use of optical control switches will also cause energy loss in off-chip optical network. Research is thus needed to consider the predictability and periodicity of ANN workloads, parallelism of models, and design architectures or topologies with data and optical devices reusability.

Optical Nonlinear Activation: There are mainly three types for nonlinear activation function implementation in optical ANNs: 1) all electronic nonlinearity, 2) photoelectric hybrid nonlinearity and 3) all-optical nonlinearity. The traditional full electronic nonlinearity receives the weight and output data from the buffer pool. While the photoelectric hybrid way requires the support of optical to electronic converter, and the optical outcomes from modulator will be converted to the electrical results. Examples include semiconductor excitable lasers [19] and electro-absorption modulators [26]. However, the optical-electrical conversion noise and energy loss limit the computing power and expansion of photoelectric hybrid based ANNs. All-optical nonlinear activation functions are still the most promising solutions, which can improve the throughput of ANNs and reduce the latency and power consumption in computation. Currently, the generally used all optical nonlinear activation is saturated absorption of optical materials including monolayer graphene, two photon excitation as well as photonic superlattices [5,56,62]. In addition, the nonlinearities of MRRs can also be designed for nonlinear activation implementation [21]. Whereas, the all optical solutions could reduce the computing accuracy and efficiency of the nonlinear modulation due to the space of nonlinear unit and the speed of devices operational. The implementations of all optical ANNs represent the long term goals, and the hybrid optoelectronic or full electric architectures remain a promising alternative to all-optical networks in the short term.

6. Conclusion

In this paper, we provide a comprehensive survey for optical implementation of ANN accelerators, including photonic computing acceleration and photonic communication acceleration. We first review the fundamental photonic devices that are employed to realize optical accelerator. For the optical neural networks, we present the current ANN accelerators that are realized by the optical effects, including resonance based optical ANN accelerators, diffraction based optical ANN accelerators and interference based optical ANN accelerators. For the optical interconnection, we introduce the existing studies from the perspectives of off-chip communication and on-chip communication for ANN accelerator. Furthermore, we point out the open challenges and the future research opportunities for photonic neural network accelerator, which is expected to provide guidance and insight for future researchers and developers on this research field.

Acknowledgments. This work is supported by the National Natural Science Foundation of China under Grant Nos. 62106052 and 62072118. The authors wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance computing facilities as part of this research.

References

1. Abu-Mostafa, Y.S., Psaltis, D.: Optical neural computers. Scientific American 256(3), 88–95 (1987)
2. Ankit, A., Hajj, I.E., Chalamalasetti, S.R., Ndu, G., Foltin, M., Williams, R.S., Faraboschi, P., Hwu, W.m.W., Strachan, J.P., Roy, K., et al.: Puma: A programmable ultra-efficient memristor-based accelerator for machine learning inference. In: Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems. pp. 715–731 (2019)
3. Ballani, H., Costa, P., Behrendt, R., Cletheroe, D., Haller, I., Jozwik, K., Karinou, F., Lange, S., Shi, K., Thomsen, B., et al.: Sirius: A flat datacenter network with nanosecond optical switching. In: Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication. pp. 782–797 (2020)
4. Banerjee, A., Park, Y., Clarke, F., Song, H., Yang, S., Kramer, G., Kim, K., Mukherjee, B.: Wavelength-division-multiplexed passive optical network (wdm-pon) technologies for broadband access: a review. Journal of optical networking 4(11), 737–758 (2005)
5. Bao, Q., Zhang, H., Ni, Z., Wang, Y., Polavarapu, L., Shen, Z., Xu, Q.H., Tang, D., Loh, K.P.: Monolayer graphene as a saturable absorber in a mode-locked laser. Nano Research 4(3), 297–307 (2011)
6. Bhardwaj, R., Saxena, S.B., Sharma, P., Jaiswal, V., Mehrotra, R.: Experimental realisation of parallel optical logic gates and combinational logic using multiple beam interference. Optik 128, 253–263 (2017)
7. Bogaerts, W., De Heyn, P., Van Vaerenbergh, T., De Vos, K., Kumar Selvaraja, S., Claes, T., Dumon, P., Bienstman, P., Van Thourhout, D., Baets, R.: Silicon microring resonators. Laser & Photonics Reviews 6(1), 47–73 (2012)
8. Bueno, J., Maktoobi, S., Froehly, L., Fischer, I., Jacquot, M., Larger, L., Brunner, D.: Reinforcement learning in a large-scale photonic recurrent neural network. Optica 5(6), 756–760 (2018)
9. Bytyn, A., Ahlsdorf, R., Leupers, R., Ascheid, G.: Dataflow aware mapping of convolutional neural networks onto many-core platforms with network-on-chip interconnect. arXiv preprint arXiv:2006.12274 (2020)
10. Chakraborty, I., Saha, G., Sengupta, A., Roy, K.: Toward fast neural computing using all-photonic phase change spiking neurons. entific Reports 8(1) (2018)
11. Chen, L., Chen, K., Zhu, Z., Yu, M., Porter, G., Qiao, C., Zhong, S.: Enabling wide-spread communications on optical fabric with megaswitch. In: 14th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 17). pp. 577–593 (2017)
12. Cheng, T., Wu, N., Yan, G., Zhang, X., Zhang, X.: Poet: A power efficient hybrid optical noc topology for heterogeneous cpu-gpu systems. In: IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society. vol. 1, pp. 3091–3095. IEEE (2019)
13. Chi, P., Li, S., Xu, C., Zhang, T., Zhao, J., Liu, Y., Wang, Y., Xie, Y.: Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory. ACM SIGARCH Computer Architecture News 44(3), 27–39 (2016)
14. Choi, H., Park, S.: A survey of machine learning-based system performance optimization techniques. Applied Sciences 11(7), 3235 (2021)
15. Clements, W.R., Humphreys, P.C., Metcalf, B.J., Kolthammer, W.S., Walmsley, I.A.: Optimal design for universal multiport interferometers. Optica 3(12), 1460–1465 (2016)
16. Crawley, D., Nikolic, K., Forshaw, M.: 3D Nanoelectronic Computer Architecture and Implementation. CRC Press (2020)
17. Dai, F., Chen, Y., Zhang, H., Huang, Z.: Accelerating fully connected neural network on optical network-on-chip (onoc). arXiv preprint arXiv:2109.14878 (2021)

18. De Lima, T.F., Peng, H.T., Tait, A.N., Nahmias, M.A., Miller, H.B., Shastri, B.J., Prucnal, P.R.: Machine learning with neuromorphic photonics. Journal of Lightwave Technology 37(5), 1515–1534 (2019)
19. De Lima, T.F., Shastri, B.J., Tait, A.N., Nahmias, M.A., Prucnal, P.R.: Progress in neuromorphic photonics. Nanophotonics 6(3), 577–599 (2017)
20. De Marinis, L., Cococcioni, M., Castoldi, P., Andriolli, N.: Photonic neural networks: A survey. IEEE Access 7, 175827–175841 (2019)
21. Denis-Le Coarer, F., Sciamanna, M., Katumba, A., Freiberger, M., Dambre, J., Bienstman, P., Rontani, D.: All-optical reservoir computing on a photonic chip using silicon-based ring resonators. IEEE Journal of Selected Topics in Quantum Electronics 24(6), 1–8 (2018)
22. Fang, M.Y.S., Manipatruni, S., Wierzynski, C., Khosrowshahi, A., DeWeese, M.R.: Design of optical neural networks with component imprecisions. Optics express 27(10), 14009–14029 (2019)
23. Farhi, E., Neven, H.: Classification with quantum neural networks on near term processors. arXiv preprint arXiv:1802.06002 (2018)
24. Farrington, N., Porter, G., Radhakrishnan, S., Bazzaz, H.H., Subramanya, V., Fainman, Y., Papen, G., Vahdat, A.: Helios: a hybrid electrical/optical switch architecture for modular data centers. In: Proceedings of the ACM SIGCOMM 2010 Conference. pp. 339–350 (2010)
25. Feng, J., Wang, Z., Wang, Z., Chen, X., Chen, S., Zhang, J., Xu, J.: Scalable low-power high-performance rack-scale optical network. In: 2019 IEEE/ACM Workshop on Photonics-Optics Technology Oriented Networking, Information and Computing Systems (PHOTONICS). pp. 1–6. IEEE (2019)
26. George, J.K., Mehrabian, A., Amin, R., Meng, J., De Lima, T.F., Tait, A.N., Shastri, B.J., El-Ghazawi, T., Prucnal, P.R., Sorger, V.J.: Neuromorphic photonics with electro-absorption modulators. Optics express 27(4), 5181–5191 (2019)
27. Glick, M., Wu, Z., Yan, S., Zhu, Z., Bergman, K.: Flexible optical interconnects for efficient resource utilization and distributed machine learning training in disaggregated architectures. In: Proc. of SPIE Vol. vol. 12027, pp. 1202703–1 (2022)
28. Hamerly, R., Bernstein, L., Sludds, A., Soljačić, M., Englund, D.: Large-scale optical neural networks based on photoelectric multiplication. Physical Review X 9(2), 021032 (2019)
29. Heck, M.J., Bowers, J.E.: Energy efficient and energy proportional optical interconnects for multi-core processors: Driving the need for on-chip sources. IEEE Journal of Selected Topics in Quantum Electronics 20(4), 332–343 (2013)
30. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural computation 18(7), 1527–1554 (2006)
31. Katumba, A., Freiberger, M., Laporte, F., Lugnan, A., Sackesyn, S., Ma, C., Dambre, J., Bienstman, P.: Neuromorphic computing based on silicon photonics and reservoir computing. IEEE Journal of Selected Topics in Quantum Electronics 24(6), 1–10 (2018)
32. Khan, K., Pasricha, S., Kim, R.G.: A survey of resource management for processing-in-memory and near-memory processing architectures. Journal of Low Power Electronics and Applications 10(4), 30 (2020)
33. Khani, M., Ghobadi, M., Alizadeh, M., Zhu, Z., Glick, M., Bergman, K., Vahdat, A., Klenk, B., Ebrahimi, E.: Sip-ml: high-bandwidth optical network interconnects for machine learning training. In: Proceedings of the 2021 ACM SIGCOMM 2021 Conference. pp. 657–675 (2021)
34. Kim, J.Y., Kang, J.M., Kim, T.Y., Han, S.K.: All-optical multiple logic gates with xor, nor, or, and nand functions using parallel soa-mzi structures: theory and experiment. Journal of Lightwave Technology 24(9), 3392 (2006)
35. Kim, Y.W., Choi, S.H., Han, T.H.: Rapid topology generation and core mapping of optical network-on-chip for heterogeneous computing platform. IEEE Access 9, 110359–110370 (2021)

36. Larger, L., Soriano, M.C., Brunner, D., Appeltant, L., Gutiérrez, J.M., Pesquera, L., Mirasso, C.R., Fischer, I.: Photonic information processing beyond turing: an optoelectronic implementation of reservoir computing. Optics express 20(3), 3241–3249 (2012)
37. Lawson, C.L., Hanson, R.J.: Solving least squares problems. SIAM (1995)
38. Li, N., Mahalingavelar, P., Vella, J.H., Leem, D.S., Azoulay, J.D., Ng, T.N.: Solution-processable infrared photodetectors: materials, device physics, and applications. Materials Science and Engineering: R: Reports 146, 100643 (2021)
39. Liang, Y.Z., Liu, H.K.: Optical matrix–matrix multiplication method demonstrated by the use of a multifocus hololens. Optics letters 9(8), 322–324 (1984)
40. Lin, X., Rivenson, Y., Yardimci, N.T., Veli, M., Luo, Y., Jarrahi, M., Ozcan, A.: All-optical machine learning using diffractive deep neural networks. Science 361(6406), 1004–1008 (2018)
41. Liu, F., Zhang, H., Chen, Y., Huang, Z., Gu, H.: Wrh-onoc: A wavelength-reused hierarchical architecture for optical network on chips. In: 2015 IEEE Conference on Computer Communications (INFOCOM). pp. 1912–1920. IEEE (2015)
42. Liu, J., Wu, Q., Sui, X., Chen, Q., Gu, G., Wang, L., Li, S.: Research progress in optical neural networks: theory, applications and developments. PhotoniX 2(1), 1–39 (2021)
43. Lu, Y., Gu, H., Yu, X., Chakrabarty, K.: Lotus: A new topology for large-scale distributed machine learning. ACM Journal on Emerging Technologies in Computing Systems (JETC) 17(1), 1–21 (2020)
44. Luo, Y., Mengu, D., Yardimci, N.T., Rivenson, Y., Veli, M., Jarrahi, M., Ozcan, A.: Design of task-specific optical systems using broadband diffractive neural networks. Light: Science & Applications 8(1), 1–14 (2019)
45. Markram, H., Muller, E., Ramaswamy, S., Reimann, M.W., Abdellah, M., Sanchez, C.A., Ailamaki, A., Alonso-Nanclares, L., Antille, N., Arsever, S., et al.: Reconstruction and simulation of neocortical microcircuitry. Cell 163(2), 456–492 (2015)
46. Mehrabian, A., Al-Kabani, Y., Sorger, V.J., El-Ghazawi, T.: Pcnna: A photonic convolutional neural network accelerator. In: 2018 31st IEEE International System-on-Chip Conference (SOCC). pp. 169–173. IEEE (2018)
47. Mellette, W.M., McGuinness, R., Roy, A., Forencich, A., Papen, G., Snoeren, A.C., Porter, G.: Rotornet: A scalable, low-complexity, optical datacenter network. In: Proceedings of the Conference of the ACM Special Interest Group on Data Communication. pp. 267–280 (2017)
48. Nurvitadhi, E., Sheffield, D., Sim, J., Mishra, A., Venkatesh, G., Marr, D.: Accelerating binarized neural networks: Comparison of fpga, cpu, gpu, and asic. In: 2016 International Conference on Field-Programmable Technology (FPT). pp. 77–84. IEEE (2016)
49. O'Connor, I., Nicolescu, G.: Integrated optical interconnect architectures for embedded systems. Springer Science & Business Media (2012)
50. Ohno, S., Toprasertpong, K., Takagi, S., Takenaka, M.: Si microring resonator crossbar array for on-chip inference and training of optical neural network. arXiv preprint arXiv:2106.04351 (2021)
51. Psaltis, D., Brady, D., Wagner, K.: Adaptive optical networks using photorefractive crystals. Applied Optics 27(9), 1752–1759 (1988)
52. Qian, C., Lin, X., Lin, X., Xu, J., Sun, Y., Li, E., Zhang, B., Chen, H.: Performing optical logic operations by a diffractive neural network. Light: Science & Applications 9(1), 1–7 (2020)
53. Reck, M., Zeilinger, A., Bernstein, H.J., Bertani, P.: Experimental realization of any discrete unitary operator. Physical review letters 73(1), 58 (1994)
54. Van der Sande, G., Brunner, D., Soriano, M.C.: Advances in photonic reservoir computing. Nanophotonics 6(3), 561–576 (2017)
55. Sasikala, V., Chitra, K.: All optical switching and associated technologies: a review. Journal of Optics 47(3), 307–317 (2018)
56. Schirmer, R.W., Gaeta, A.L.: Nonlinear mirror based on two-photon absorption. JOSA B 14(11), 2865–2868 (1997)

57. Scott, A., Diddams: The evolving optical frequency comb [invited]. Journal of the Optical Society of America B 27(11), B51–B62 (2010)
58. Shen, Y., Harris, N.C., Skirlo, S., Prabhu, M., Baehr-Jones, T., Hochberg, M., Sun, X., Zhao, S., Larochelle, H., Englund, D., et al.: Deep learning with coherent nanophotonic circuits. Nature Photonics 11(7), 441–446 (2017)
59. Shiflett, K., Karanth, A., Bunescu, R., Louri, A.: Albireo: Energy-efficient acceleration of convolutional neural networks via silicon photonics. In: 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA). pp. 860–873. IEEE (2021)
60. Shiflett, K., Wright, D., Karanth, A., Louri, A.: Pixel: Photonic neural network accelerator. In: 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA). pp. 474–487. IEEE (2020)
61. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al.: Mastering the game of go with deep neural networks and tree search. nature 529(7587), 484–489 (2016)
62. Soljačić, M., Ibanescu, M., Johnson, S.G., Fink, Y., Joannopoulos, J.D.: Optimal bistable switching in nonlinear photonic crystals. Physical Review E 66(5), 055601 (2002)
63. Sorrentino, T., Quintero-Quiroz, C., Torrent, M., Masoller, C.: Analysis of the spike rate and spike correlations in modulated semiconductor lasers with optical feedback. IEEE Journal of Selected Topics in Quantum Electronics 21(6), 561–567 (2015)
64. Spuesens, T., Liu, L., de Vries, T., Romeo, P.R., Regreny, P., Van Thourhout, D.: Improved design of an inp-based microdisk laser heterogeneously integrated with soi. In: 2009 6th IEEE International Conference on Group IV Photonics. pp. 202–204. IEEE (2009)
65. Stanley, A., Singh, G., Eke, J., Tsuda, H.: Mach–zehnder interferometer: A review of a perfect all-optical switching structure. In: Proceedings of the International Conference on Recent Cognizance in Wireless Communication & Image Processing. pp. 415–425. Springer (2016)
66. Tait, A.N., Nahmias, M.A., Shastri, B.J., Prucnal, P.R.: Broadcast and weight: an integrated network for scalable photonic spike processing. Journal of Lightwave Technology 32(21), 4029–4041 (2014)
67. Tait, A.N., Wu, A.X., De Lima, T.F., Zhou, E., Shastri, B.J., Nahmias, M.A., Prucnal, P.R.: Microring weight banks. IEEE Journal of Selected Topics in Quantum Electronics 22(6), 312–325 (2016)
68. Totović, A.R., Dabos, G., Passalis, N., Tefas, A., Pleros, N.: Femtojoule per mac neuromorphic photonics: An energy and technology roadmap. IEEE Journal of selected topics in Quantum Electronics 26(5), 1–15 (2020)
69. Truong, T.N., Takano, R.: Hybrid electrical/optical switch architectures for training distributed deep learning in large-scale. IEICE TRANSACTIONS on Information and Systems 104(8), 1332–1339 (2021)
70. Wang, W., Khazraee, M., Zhong, Z., Jia, Z., Mudigere, D., Zhang, Y., Kewitsch, A., Ghobadi, M.: Topoopt: Optimizing the network topology for distributed dnn training. arXiv preprint arXiv:2202.00433 (2022)
71. Wang, Y.G.: Applications of memristors in neural networks and neuromorphic computing: A review. Int. J. Mach. Learn. Comput 11, 350–356 (2021)
72. Xia, C., Chen, Y., Zhang, H., Zhang, H., Wu, J.: Photonic computing and communication for neural network accelerators. In: International Conference on Parallel and Distributed Computing: Applications and Technologies. pp. 121–128. Springer (2022)
73. Xiang, S., Wen, A., Pan, W.: Emulation of spiking response and spiking frequency property in vcsel-based photonic neuron. IEEE Photonics Journal 8(5), 1–9 (2016)
74. Xu, R., Lv, P., Xu, F., Shi, Y.: A survey of approaches for implementing optical neural networks. Optics & Laser Technology 136, 106787 (2021)
75. Yang, P., Pang, Z., Wang, Z., Wang, Z., Xie, M., Chen, X., Duong, L.H., Xu, J.: Rson: An inter/intra-chip silicon photonic network for rack-scale computing systems. In: 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE). pp. 1369–1374. IEEE (2018)

76. Yao, Z., Wu, K., Tan, B.X., Wang, J., Li, Y., Zhang, Y., Poon, A.W.: Integrated silicon photonic microresonators: emerging technologies. IEEE Journal of Selected Topics in Quantum Electronics 24(6), 1–24 (2018)
77. Zhang, C., Li, P., Sun, G., Guan, Y., Xiao, B., Cong, J.: Optimizing fpga-based accelerator design for deep convolutional neural networks. In: Proceedings of the 2015 ACM/SIGDA international symposium on field-programmable gate arrays. pp. 161–170 (2015)
78. Zhang, H., Gu, M., Jiang, X., Thompson, J., Cai, H., Paesani, S., Santagati, R., Laing, A., Zhang, Y., Yung, M., et al.: An optical neural chip for implementing complex-valued neural network. Nature Communications 12(1), 1–11 (2021)
79. Zhang, Q., Yu, H., Barbiero, M., Wang, B., Gu, M.: Artificial neural networks enabled by nanophotonics. Light: Science & Applications 8(1), 1–14 (2019)
80. Zhao, Y., Zhao, H., Lv, R.q., Zhao, J.: Review of optical fiber mach–zehnder interferometers with micro-cavity fabricated by femtosecond laser and sensing applications. Optics and Lasers in Engineering 117, 7–20 (2019)
81. Zhao, Z., Gu, J., Ying, Z., Feng, C., Chen, R.T., Pan, D.Z.: Design technology for scalable and robust photonic integrated circuits: Invited paper. In: 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). pp. 1–7 (2019)
82. Zhou, Z., Yin, B., Michel, J.: On-chip light sources for silicon photonics. Light: Science & Applications 4(11), e358 (2015)
83. Zhu, Z., Teh, M.Y., Wu, Z., Glick, M.S., Yan, S., Hattink, M., Bergman, K.: Distributed deep learning training using silicon photonic switched architectures. APL Photonics 7(3), 1–11 (2022)
84. Zuo, Y., Li, B., Zhao, Y., Jiang, Y., Chen, Y.C., Chen, P., Jo, G.B., Liu, J., Du, S.: All-optical neural network with nonlinear activation functions. Optica 6(9), 1132–1137 (2019)

Chengpeng Xia received B.E. degree from Lanzhou Jiaotong University in 2017 and M.E. degree from Guangdong University of Technology in 2020. He is now working toward Ph.D. degree in Computer Science from University of Otago. His main research interests include optical computing, optical neural network accelerator and distributed computing.

Yawen Chen (Member, IEEE) received the PhD degree in computer science from the University of Adelaide, Adelaide, Australia, in 2008. She is a senior lecturer at the University of Otago in New Zealand. Her research interests include resource optimization and performance evaluation in computer networking and computer architecture (optical network-on-chips, interconnection network, wired, and wireless networking).

Haibo Zhang (Senior Member, IEEE) received the PhD degree from the University of Adelaide, Adelaide, Australia, in 2009. From 2009 to 2010, he was a postdoctoral research associate with the Automatic Control Lab, Royal Institute of Technology in Sweden. He is currently a senior lecturer at the Department of Computer Science, University of Otago, New Zealand. His current research interests include wireless communication, optical network-on-chips, wireless body sensor networks, protocol design.

Hao Zhang received B.E. degree and M.E. degree from Shandong University of Science and Technology in 2017 and 2020, respectively. He is now working toward Ph.D. degree in Computer Science from University of Otago. His main research interests include optical network on chip, optical computing and parallel computing.

Fei Dai obtained BSc of computer science and MSc of software engineering in 2014 and 2019 from Guilin University of Technology, China. He is currently working towards the PhD degree in computer science at University of Otago, New Zealand. His research interests include optical interconnect communications, multicore architecture, parallel computation, deep learning accelerators, IoT system, etc.

Jigang Wu received B.Sc. degree from Lanzhou University, and Ph.D. degree from the University of Science and Technology of China. Now, he is distinguished professor of School of Computer Science and Technology, Guangdong University of Technology. His research interests include network computing and machine learning.

Received: January 31, 2022; Accepted: December 10, 2022.