Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 6661022, 11 pages
https://doi.org/10.1155/2020/6661022

Research Article

An Energy-Efficient Silicon Photonic-Assisted Deep Learning Accelerator for Big Data

Mengkun Li[1] **and Yongjian Wang**[2]

[1]*School of Management, Capital Normal University, Beijing 100089, China*
[2]*National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China*

Correspondence should be addressed to Mengkun Li; limengkun@cnu.edu.cn and Yongjian Wang; wyj@cert.org.cn

Received 15 November 2020; Revised 7 December 2020; Accepted 10 December 2020; Published 16 December 2020

Academic Editor: Xiaojie Wang

Deep learning has become the most mainstream technology in artificial intelligence (AI) because it can be comparable to human performance in complex tasks. However, in the era of big data, the ever-increasing data volume and model scale makes deep learning require mighty computing power and acceptable energy costs. For electrical chips, including most deep learning accelerators, transistor performance limitations make it challenging to meet computing's energy efficiency requirements. Silicon photonic devices are expected to replace transistors and become the mainstream components in computing architecture due to their advantages, such as low energy consumption, large bandwidth, and high speed. Therefore, we propose a silicon photonic-assisted deep learning accelerator for big data. The accelerator uses microring resonators (MRs) to form a photonic multiplication array. It combines photonic-specific wavelength division multiplexing (WDM) technology to achieve multiple parallel calculations of input feature maps and convolution kernels at the speed of light, providing the promise of energy efficiency and calculation speed improvement. The proposed accelerator achieves at least a 75x improvement in computational efficiency compared to the traditional electrical design.

1. Introduction

In a modern society driven by big data, artificial intelligence (AI) has brought great convenience to human life. As an indispensable part of solving complex problems in the field of AI, deep learning has been used in many applications, e.g., image and speech recognition, machine translation, self-driving, Internet of Things (IoTs), 5th generation (5G) mobile networks, and edge computing [1–13]. Deep learning can use effective learning and training methods to discover the inherent rules in the data model, thus helping machines to perform advanced reasoning tasks like human beings. In deep learning, convolutional neural networks (CNNs) are considered the most representative framework due to its advantages: the simple structure, few parameters, noticeable extraction features, and high recognition rate [14, 15]. Due to the enormous amount of data, the efficient inference of CNNs has high computing requirements. Therefore, the development of the hardware inference accelerator, which can provide strong computing power, is the key to meet the needs of CNNs.

At present, hardware accelerators that perform CNN operation mainly include GPUs, ASICs [16], FPGAs [17], TPU [18], and the emerging near data processing accelerator ISAAC [19]. However, current accelerators rely on a large degree of data movement. The energy consumption of electrical wire-based data movement is even greater than the energy consumed by the computing itself. Due to the widening gap between abundant data and limited power budget, these electric-based accelerators' energy crisis is still unpredictable. Limited by the transmittance rate of the electrical line, the calculation speed and throughput of these accelerators may not be able to keep up with the increase in power, resulting in limited throughput per second per watt.

Recently, silicon photonic technology has emerged as a promising solution to address the issues above [20–25]. Firstly, a certain transistor-based circuit's power consumption has a positive correlation with f^3 (f is the clock frequency). The

photonic circuit only consumes the power proportional to f, so that the photonic circuit can provide ultralow energy consumption [26]. Secondly, light has a very low transmission delay on a chip, typically 0.14 ps for 10 microns, which is 1–2 orders of magnitude faster than the transistor-based circuit [27]. Finally, the photonic circuit is insulated and has strong antielectromagnetic interference performance.

Furthermore, benefitting from the peaceful development of photonic integration technology and manufacturing platform, various mature active and passive building blocks have been demonstrated experimentally, such as modulators, photodetectors, splitters, wavelength multiplexers, and filters [28–31]. Based on these photonic devices, photonic computing elements such as photonic adders, differentiators, integrators, and multipliers can be realized [32–35]. Once the photonic devices can be successfully applied to the CNN accelerator's design, it is expected to improve energy efficiency in deep learning significantly. In addition, by utilizing optical multichannel multiplexing technologies, such as wavelength division multiplexing (WDM) [36–38], we can easily use the speed of light to achieve massively parallel computing to improve the inference speed of CNNs significantly.

Thus, we propose a silicon photonic-assisted CNN accelerator for deep learning. We first use the mature microring resonators (MRs) as the basic unit to design a photonic matrix-vector multiplier (PMVM) to perform the most complex convolution operation on CNNs. Then, we introduce an analytical model to identify the number of MRs used, power consumption, area, and execution time in each layer of the CNNs. At last, we introduce our PMVM-based photonic-assisted CNN accelerator architecture and its workflow. The simulation results show that our accelerator can increase the CNN's inference speed by at least 75 times under the same energy consumption than the current electricity-based accelerators.

The rest of the paper is organized as follows. Section 2 briefly discusses the related works. Section 3 discusses the proposed PMVM and accelerator architectures, followed by Section 4 presenting the performance evaluation of the silicon photonic-assisted accelerator. Section 5 concludes this paper.

2. Related Work

In this section, we first describe CNNs' structure and computing process in deep learning. Then, we introduce photonic devices that might be used. These related works can be used as the guide for our research on the photonic-assisted accelerator design.

2.1. Convolutional Neural Network (CNN) Basics. CNN is comprised of stacking multiple computation layers for feature extraction and classification. Compared to the fully neural networks with simple training but limited scalability, CNN has very deep convolutional (CONV), pooling (POOL), and full connection (FC) layers. Therefore, it can achieve high accuracy [14]. In each CONV layer, the input maps are transformed into highly abstract representation feature maps and convolution with the kernel to generate output feature maps. After nonlinearity and pooling, the output features can be used as the input for the next layer. After multi-CONV and POOL layers, the features are sent to the FC layers and finally output the classification results. The CONV layers take more than 90% of the calculation time [39]. Therefore, the design of an optimization accelerator for CONV layers can significantly improve the entire CNN's performance. Figure 1 shows a CONV layer. It has M 3D convolutional kernels with size $S \times R \times C$ and N input maps with size $W \times H \times C$. M kernels perform M times 3D convolution on the input maps with a sliding stride of S and generate an $E \times F \times M$ output map. In each output map, the value of the element (m, f, e) can be computed as

$$O(m, f, e) = \sigma\left(\sum_{c=0}^{C-1}\sum_{i=0}^{S-1}\sum_{j=0}^{R-1} K[m][c][i][j] \times I[c][f * S + i][e * R + j]\right), \tag{1}$$

where I, K, and O are the input, kernel, and output matrices, respectively. $\sigma(\cdot)$ is an activation function, such as ReLU and sigmoid. The pseudocode to perform this normal convolution operation is shown in Figure 1. Note that in each layer, all kernels share the same input data. Therefore, if the accelerator can support multiple kernels that simultaneously convolve with the same input data, the number of access buffers is reduced. The cycle time can also be reduced, thereby increasing the throughput. As shown in the pseudocode, assuming the input map can be reused by G_m kernels simultaneously, the total convolution cycles can be saved by G_m time. The size of G_m is determined by the accelerator. Therefore, designing the corresponding accelerator architecture to maximize this data reuse capability is the paper's primary motivation.

2.2. Silicon Photonic Devices. Microelectronic devices are the basis of the current CNN accelerator. But with the reduction of feature size, the ability of electronic information processing has approached its limit. Silicon photonic devices offer an exact route to solve the electrical processing bottleneck due to its low loss, high speed, low energy consumption, and compatibility with CMOS platforms. Among the various silicon photonic devices, MRs are considered the most critical devices in photonic computing due to their excellent wavelength selection characteristics, small size, high modulation rate, low energy consumption, and high-quality factors [40, 41]. Figure 2 shows two commonly used MR structures: all-pass MR (Figure 2(a)) and 1×2 cross-MR (Figure 2(e)). All-pass MRs include one straight waveguide and one MR, assuming that the resonant wavelength of the MR is λ_{mr} and the input signal wavelength is λ_{in}. When $\lambda_{in} = \lambda_{mr}$, the input signal will be wholly coupled into the MR, so that the signal power output from the through port is zero (transmittance rate is 0). When $\lambda_{in} \neq \lambda_{mr}$, the coupling ability between the input waveguide and the MR will become weak, and when it is weak enough, the signal will output from the through port (transmittance rate is 1). When the MR's resonance wavelength is between λ_1 and λ_2, the transmittance rate of the MR will be between 0 and 1.

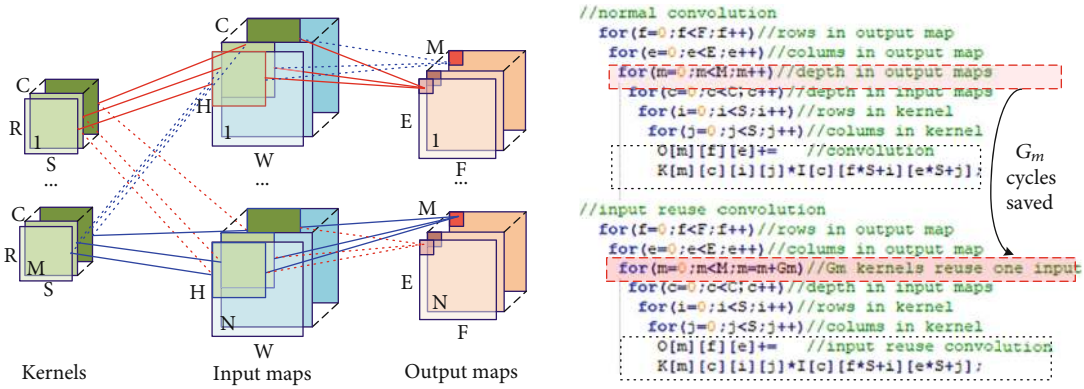


Figure 1: The logical graph and pseudocode for standard convolution and input map reuse convolution of a CONV layer.

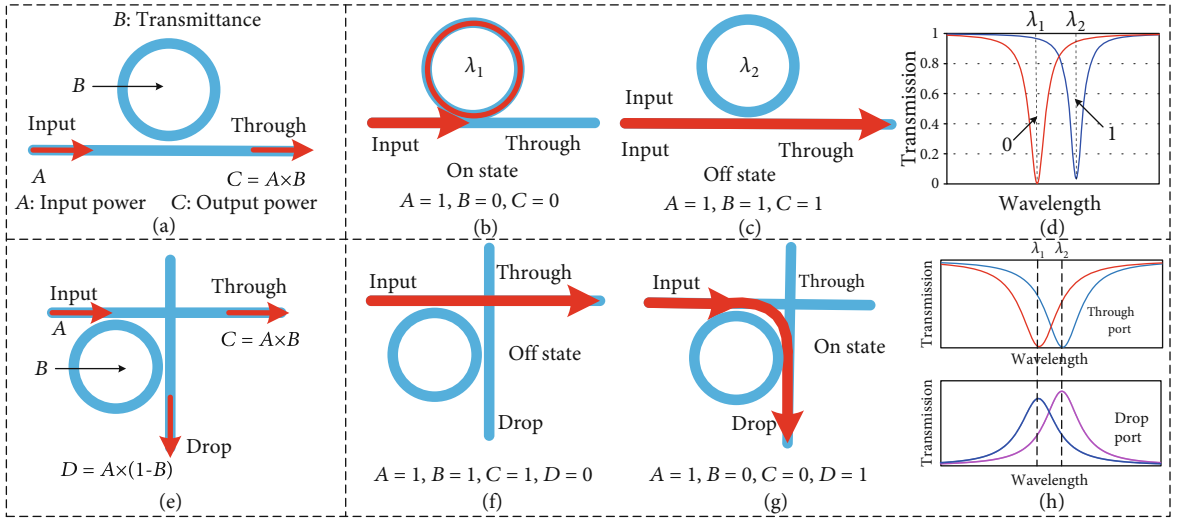


Figure 2: (a) All-pass MR photonic computing unit. (b, c) Computing process with on-state MR and off-state MR. (d) The transmission spectrum lines of the through port for all-pass MR with different wavelengths. (e) 1×2 cross-MR photonic computing unit. (f, g) Computing process with off-state MR and on-state MR. (h) The transmission spectrum lines of through and drop ports for 1×2 cross-MR with different wavelengths.

Therefore, we can use the resonance effect of MR to adjust the output power to realize the photonic multiplication calculation. For instance, as shown in Figure 2(a), assuming that the input optical signal power is A, the transmittance of the MR is B ($0 \le B \le 1$). When the input optical signal passes through the MR, part of the light $(1 - B)$ will be coupled to the MR, and the output optical power of the through port is $C = A \times B$. Usually, by adding a bias voltage to the MR, the transmittance rate of MR (B) can be changed under the thermooptic or electrooptic effect. According to [34], each MR can store more than 16 levels of transmittance rate (i.e., 4 bits). Therefore, for a 16-bit floating-point calculation [19], only 4 MRs are needed. Figure 2(e) shows the structure of 1×2 cross-MR, which has the same working principle as the all-pass MR. The output powers of the through and drop can be controlled by controlling the MR's resonant wavelength, as shown in Figures 2(f)–2(h). Since the multiplication operation of the above two structures can be realized in the optical domain, they have a high processing speed, making them ideal choices for photonic multiplication units.

3. Silicon Photonic-Assisted CNN Accelerator Architecture Design

In order to use silicon photonic technology to improve the calculation rate in deep learning, we first propose a PMVM based on photonic devices in this section. Then, we create a photonic-assisted CNN accelerator architecture based on PMVM.

3.1. Silicon Photonic Matrix-Vector Multiplier. Matrix-vector multiplication is the most important operation in CNN. Therefore, in this section, we will use the essential photonic devices to construct a PMVM and map the input feature map and kernel weight data to the PMVM to complete the parallel multiplication operation.

Figure 3 shows the PMVM architecture. It relies on an all-pass MR-based input matrix and 1×2 cross-MR-based kernel matrix. Current CNNs have tens of kernels in each layer to convolve the same set of input data. Therefore, in PMVM, we multiplex the input data to be convolved with multiple kernels simultaneously, reducing the waste of time

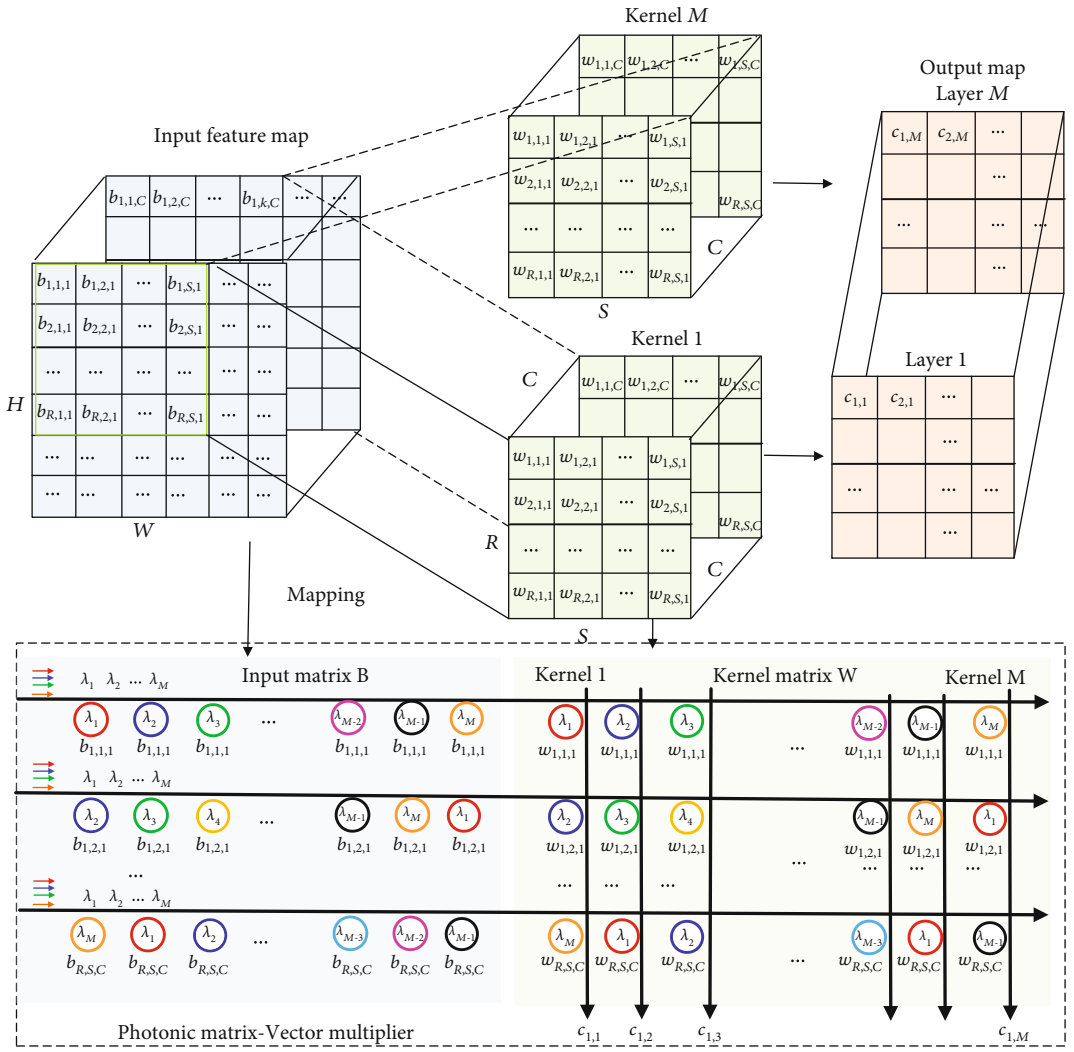


Figure 3: Photonic matrix-vector multiplier.

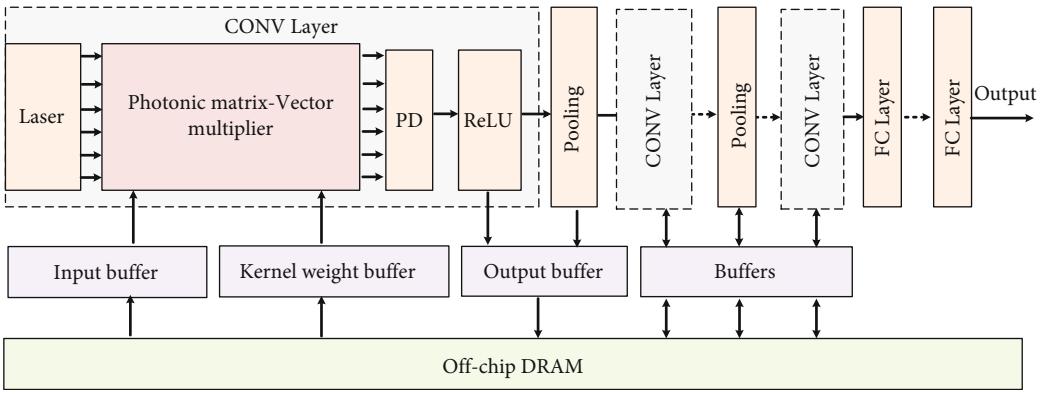
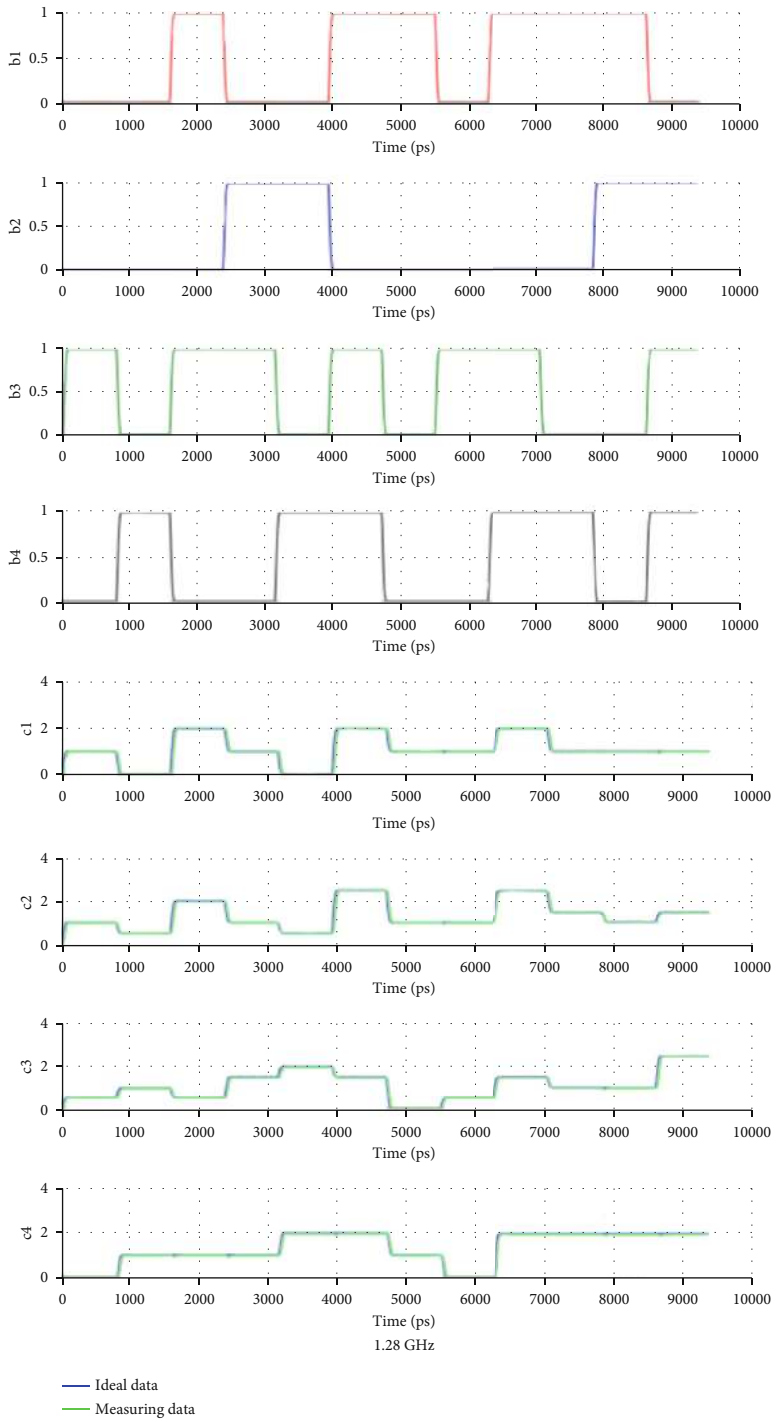


Figure 4: Photonic-assisted CNN inference accelerator architecture.

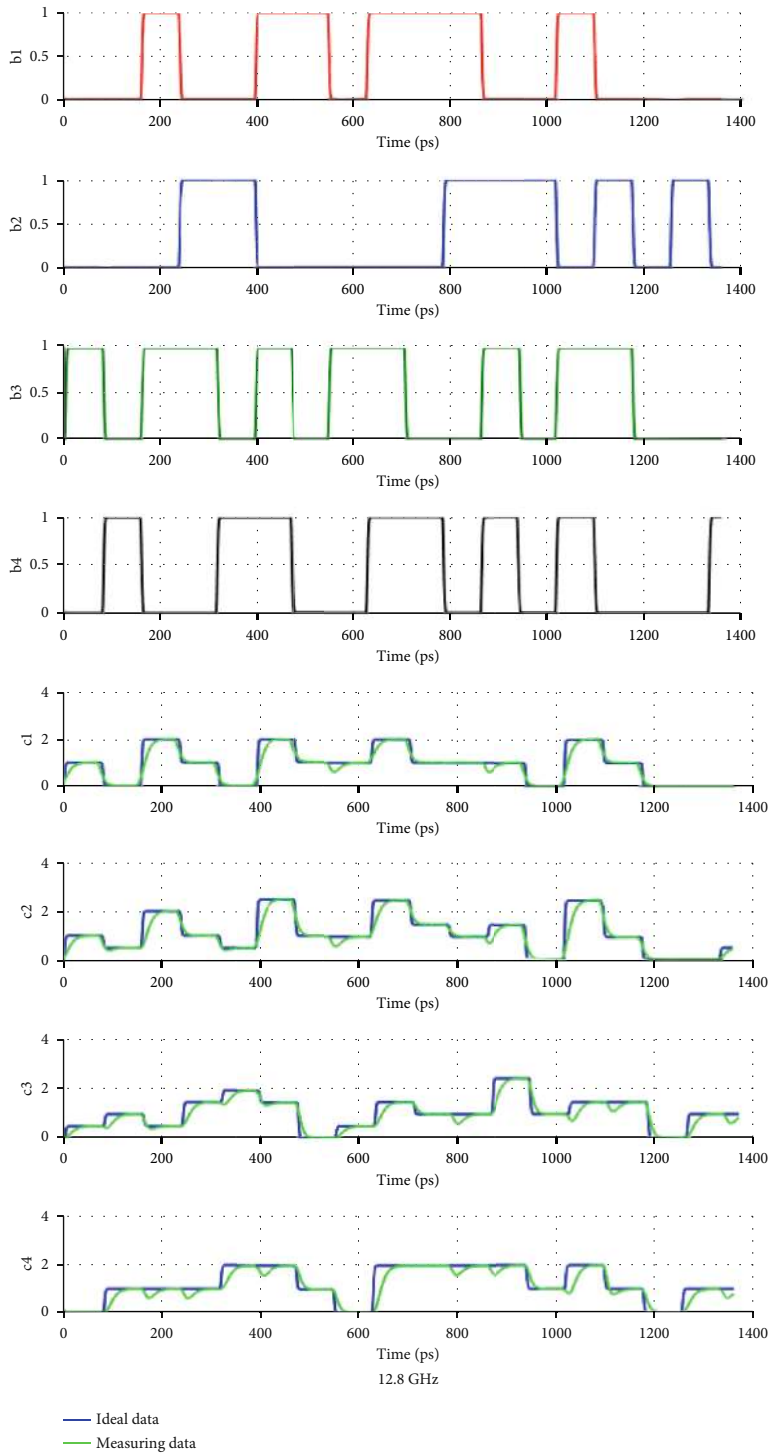
and energy consumption caused by repeated reading of the input data. For convenience, if we assume that the size of each kernel is $R \times S \times C$, the number of the kernels is M. The weight matrix W in PMVM can be composed of an $(R \times S \times C) \times M$ MR-based crossbar array. The MR in the array has different resonance wavelengths to ensure parallel computing. The MR would be on resonance when the wavelength of the light fits a whole number of times inside the optical length of the MRs:

$$\lambda_{\text{res}} = \frac{n_{\text{eff}} L}{m}, \quad L = 2\pi R, m = 1, 2, 3 \cdots . \tag{2}$$



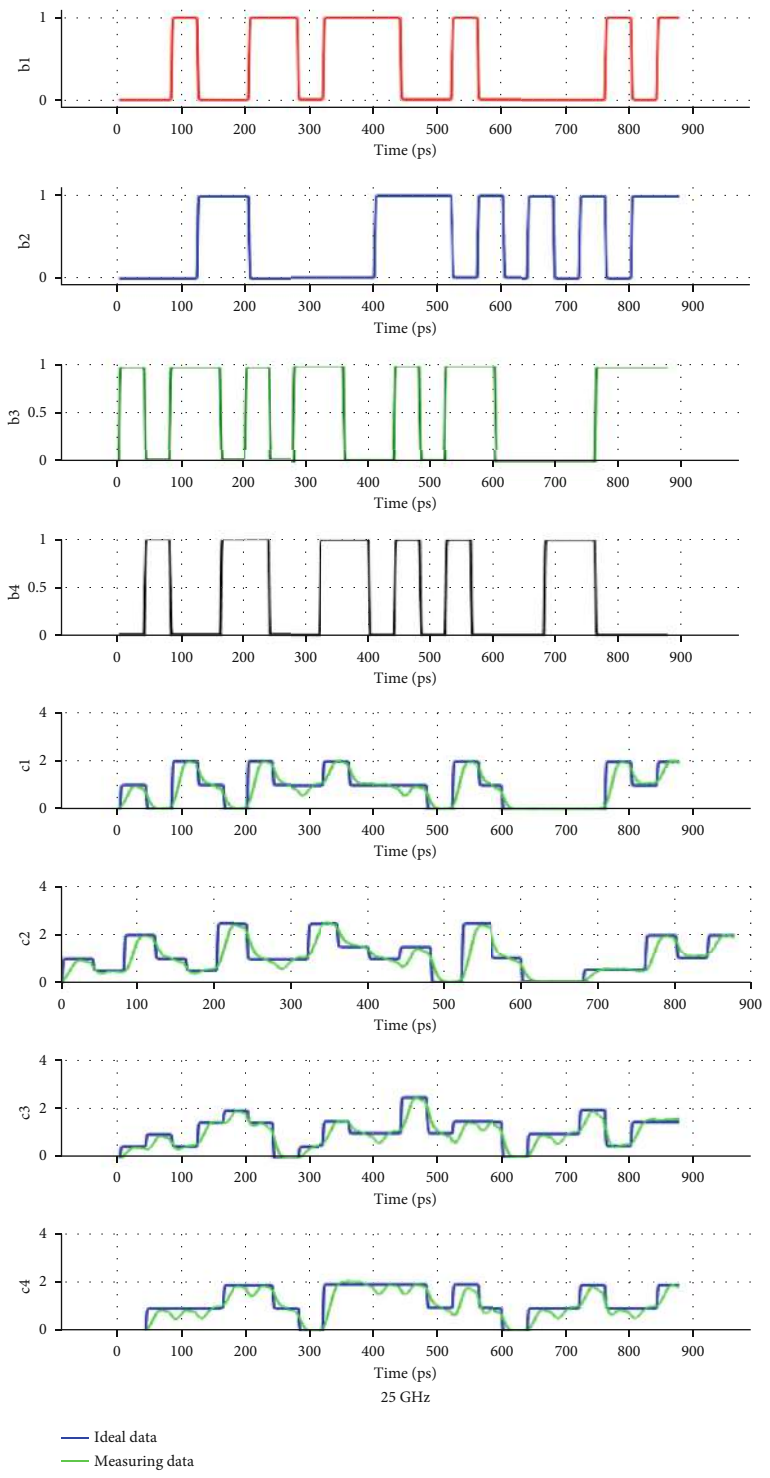
(a)

Figure 5: Continued.



(b)

Figure 5: Continued.



(c)

Figure 5: The simulation waveforms for 4 × 4 PMVM with (a) 1.28 GHz; (b) 12.8 GHz; (c) 25 GHz.

Here, λ_{res} is the resonant wavelength, n_{eff} is the effective refractive index, and R is the radius of the MRs, respectively. Therefore, in this paper, we use MRs with different radii to realize the control of different resonance wavelengths.

As shown in Figure 3, the weight value of the coordinate (i, j, n) in the m-th kernel can be represented by the drop port transmittance rate of the m-column and $((n-1) \times S \times R + (i-1) \times S + j)$ -row MR in the crossbar array, where $0 < i < S, 0 < j < R, \quad 0 < n < C,$ and $0 < m < M$. According to CNN's characteristics, the state of all MRs in the kernel matrix remains unchanged during the inference process. In PMVM, the feature data of the input feature maps are

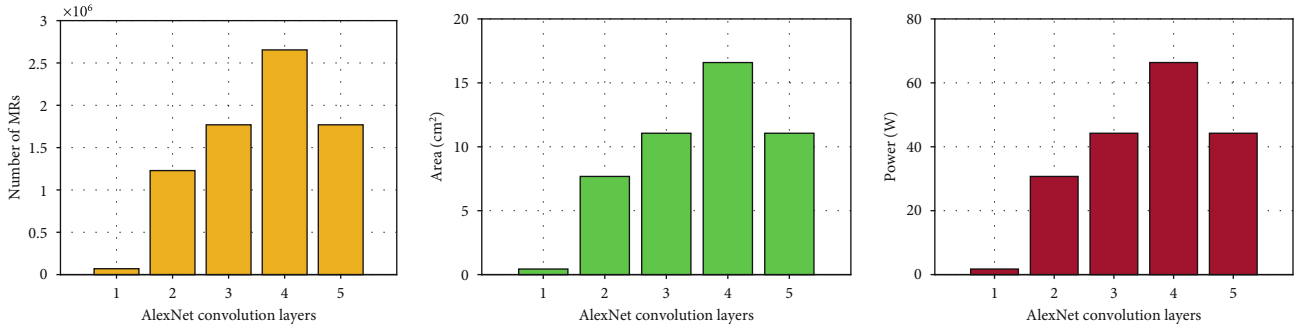


Figure 6: Total number of MRs required, area occupied, and power consumption for different convolutional layers of AlexNet.

mapped to the input matrix in turn. The input matrix comprises all-pass MR, and the size is the same as the kernel matrix. The values of the MR in the input matrix are updated with the sliding window. As shown in Figure 3, assuming the stride of the sliding window is 1, the value of MR with wavelength $\lambda_{1,1}$ is $b_{1,1,1}$ at time t_1, and it will be updated to $b_{1,2,1}$ at time t_2. In this PMVM, the multi-wavelength optical signals emitted by the lasers are injected from the input port of the input matrix and output from the kernel matrix after photonic multiply-accumulate (MAC) operation. The output power is the sum of all wavelength signals. As shown in Figure 3, the calculation process of the PMVM at time t_1 is

$$\begin{aligned}&[b_{1,1,1}, b_{1,2,1}, \cdots, b_{R,S,C}]\\ &\times \begin{bmatrix} w_{1,1,1(\text{kernel }1)} & w_{1,1,1(\text{kernel }2)} & \cdots & w_{1,1,1(\text{kernel }M)} \\ w_{1,2,1(\text{kernel }1)} & w_{1,2,1(\text{kernel }2)} & \cdots & w_{1,2,1(\text{kernel }M)} \\ \cdots & \cdots & \cdots & \cdots \\ w_{R,S,C(\text{kernel }1)} & w_{R,S,C(\text{kernel }2)} & \cdots & w_{R,S,C(\text{kernel }M)} \end{bmatrix}\\ &= [c_{1,1}, c_{1,2}, \cdots, c_{1,M}].\end{aligned} \tag{3}$$

Therefore, the PMVM enables all MAC operations to finish with high parallelism. According to [39], the number of multiplexed wavelengths can reach 128. Thus, the computation speed of the PMVM will be $128 \times 128 \times 10 \times 10^{10} = 1.6384 \times 10^{15}$ MAC/s when all MRs work at 10 Gb/s modulation speed.

3.2. Silicon Photonic-Assisted Accelerator Architecture Design. Based on the PMVM, we propose a photonic-assisted CNN accelerator architecture, as shown in Figure 4. The accelerator consists of multilayer CONV layers, pooling layers, and FC layers, and all layers are processed sequentially. According to different CNN models, the distribution between layers can be adjusted. The proposed PMVM is deployed in the CONV layers. The input matrix and kernel matrix values are read from the off-chip DRAM (the off-chip DRAM data will be sent to the on-chip buffer first). Once the CNN model is sufficiently trained, the weight values of kernels in each layer are determined and programmed into PMVMs by configuring each MR's transmittance rate in the kernel matrix. During the whole process, only the value of the input matrix will be updated. After highly parallel MAC operations, the output optical signals are converted into the electrical signals by photodetectors (PDs) and then activated and pooled. This process can be done very fast because all the photonic-assisted devices' operating frequency can reach tens of GHz, e.g., lasers, MR, and PD. The calculation results are stored back to the off-chip DRAM for reading and calculation of the next layer. After multiple layers of convolution, pooling, and full interconnection operations, the accelerator will output the final inference results.

Table 1: Execution time for convolution layers of AlexNet ($P = 0$, $S = 1$).

CONV layers	Input patch size	Kernel size	Execution time (μs)
1	55×55	11×11	337.561
2	27×27	5×5	19.881
3	13×13	3×3	1.0368
4	13×13	3×3	1.0368
5	13×13	3×3	1.0368

4. Simulation Evaluations

In this section, we used a widely adopted deep learning accelerator simulator, FODLAM [42], to evaluate the performance of our accelerator. FODLAM does total up the latency and energy for each layer, including the storage and read/write costs of the intermediate layers. The simulation of the photonic part of our accelerator structure is performed using a professional optical simulation platform, i.e., Lumerical Solutions [43]. The configuration parameters of other accelerators are obtained from the prior art as referenced.

4.1. Photonic Matrix Multiplication Function Verification. The photonic vector multiplication results of $B \times W$ with different working frequencies are exhibited in Figure 5. Assuming the matrix size is 4×4, we perform the simulation using four CW lasers with different working wavelengths. The input matrix ($B = [b1\,;\, b2\,;\, b3\,;\, b4]$) is modulated by four 2^7-1 pseudorandom binary sequence (PRBS) from the pattern generators. The values in the kernel matrix W are randomly generated once programmed into the corresponding MR

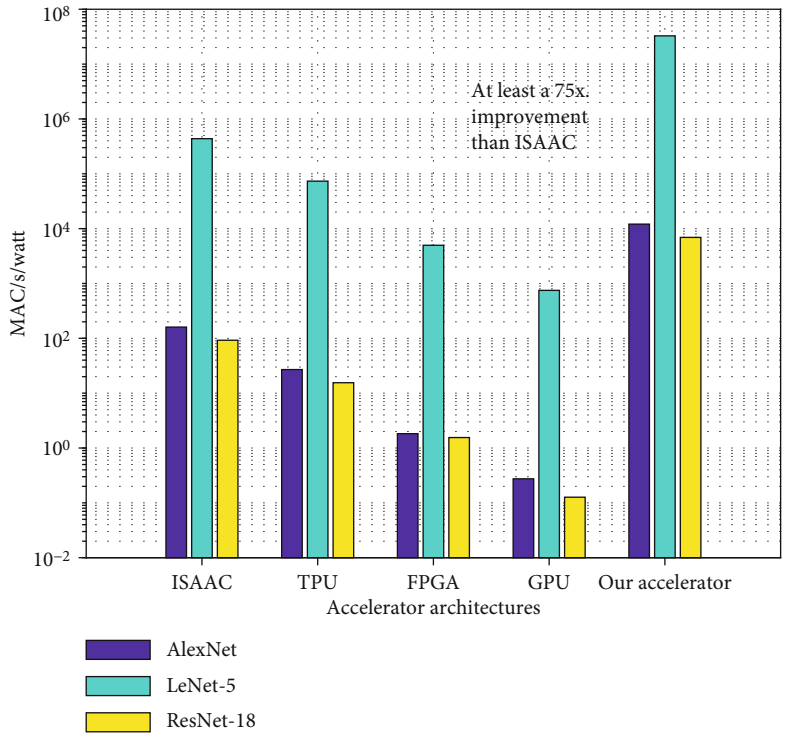


Figure 7: The inference performance of different accelerators under different CNN models.

units with $W = [1, 0, 0.5, 1\,;\, 0, 1, 1, 1\,;\, 1, 0.5, 0, 1\,;\, 0, 1, 1, 0]$, which is fixed throughout the simulation. The simulation output $C = [c1, c2, c3, c4]$ results from the multiply-accumulate of W and B.

It can be seen from Figure 5 that when PMVM works at 1.28 GHz, the simulation results are almost the same as the ideal results. Although a particular error will occur as the operating frequency increases, the designed PMVM can also maintain good calculation accuracy under the operating frequency of 25 GHz.

4.2. Area and Power Consumption Evaluation Models. The area of PMVM is affected by MRs. According to [44], the area of each MR unit is $25\,\mu\text{m} \times 25\,\mu\text{m}$ with 0.025 mW energy consumption. The size of the kernel determines the number of MRs used in PMVM. For example, the first CONV layer of the AlexNet architecture contains 96 kernels, and the size of each kernel is $11 \times 11 \times 3$. Assuming that a set of input data completes all convolution operations of this layer within one cycle, theoretically, the PMVM of this layer needs 69,696 MRs. The area and power of PMVMs in this layer are 43.56 mm^2 and 1.74 W, respectively. Due to the current technological limitations, it is difficult to integrate so many MRs on a single chip. Therefore, multiple interconnected chips are usually used to complete the above functions [19, 39]. Figure 6 shows the number of MRs, occupied area, and power consumption in each convolutional layer of AlexNet. It can be seen that the fourth layer of AlexNet has the largest consumption because this layer has the largest convolution kernel.

4.3. Execution Time Evaluation Models. As mentioned in the previous section, our PMVM can compute convolutions of multiple kernels in parallel for a single input data within one cycle. In AlexNet, the length and width of the input patches are the same. Assuming the size of input patches is $W \times W$, the kernel size is $K \times K$, the padding size is P, and the stride is S. Thus, the number of convolution calculations for each input patch is

$$N_{\text{Calculation}} = \left(\left[\frac{W - K + 2P}{S} \right] + 1 \right)^2. \tag{4}$$

Thus, the computation time of each input patch is

$$T = \frac{N_{\text{Calculation}}}{f_{\text{PMVM}}}, \tag{5}$$

where f_{PMVM} is the operating frequency of the PMVM.

Assuming $P = 0$ and $S = 1$, the execution time results for each layer of AlexNet as shown in Table 1 when the working frequency of the PMVM is 25 GHz.

4.4. Inference Performance. To fully evaluate our accelerator's inference performance, the energy-efficient performance is considered in our simulation, i.e., MAC/s/watt. We compared our accelerator with GPU, FPGA, TPU, and ReRAM-based CNN accelerator ISAAC. The CNN architecture are AlexNet, LeNet-5, and ResNet-18, and the database are ImageNet (AlexNet and ResNet-18) and MNIST (LeNet-5). In the simulation, we use the parameters of the electrical devices listed in Ref. [19]. The simulation results of MAC/s/watt are shown in Figure 7. Compared to other electricity-based accelerators, our accelerator can increase energy efficiency

by at least 75 times because it can use silicon photonics' advantages to increase computing speed while reducing energy consumption.

5. Conclusions

This paper proposed a silicon photonic-assisted CNN accelerator to maximize the inference performance in deep learning. It achieved a high inference throughput by exploiting the high modulation rate MRs and WDM technology. The proposed accelerator achieves at least 75x improvement in computational efficiency compared to the state-of-the-art designs. The photoelectric hybrid CNN accelerator needs to match the operating frequency of the electronic device, which affects the performance of the photonic device. In the future, we will explore the all-optical accelerators to maximize acceleration performance.

Data Availability

Data are available on request. The data are available by contacting Mengkun Li (limengkun@cnu.edu.cn).

Conflicts of Interest

The authors declare that there is no conflict of interest.

Acknowledgments

This research was funded by the Major Technology Project of China National Machinery Industry Corporation (SINOMACH): "Research and Application of Key Technologies for Industrial Environment Monitoring, Early Warning and Intelligent Vibration Control (SINOMAST-ZDZX-2017-05)," and partially supported by the Scientific Research Foundation of the Beijing Municipal Education Commission (KM201810028021).

References

[1] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, 2019.

[2] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: a survey," *IEEE Transactions on Knowledge and Data Engineering*, p. 1, 2020.

[3] X. Wang, Z. Ning, and S. Guo, "Multi-agent imitation learning for pervasive edge computing: a decentralized computation offloading algorithm," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 2, pp. 411–425, 2021.

[4] J. Chen and X. Ran, "Deep learning with edge computing: a review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.

[5] X. Wang, Z. Ning, S. Guo, and L. Wang, "Imitation learning enabled task scheduling for online vehicular edge computing," *IEEE Transactions on Mobile Computing*, p. 1, 2020.

[6] Z. Q. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: a review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[7] Z. Ning, R. Y. K. Kwok, K. Zhang et al., "Joint computing and caching in 5G-envisioned internet of vehicles: a deep reinforcement learning based traffic control system," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2020.

[8] H. Li, K. Ota, and M. Dong, "Learning IoT in edge: deep learning for the internet of things with edge computing," *IEEE Network*, vol. 32, no. 1, pp. 96–101, 2018.

[9] Z. Ning, K. Zhang, X. Wang et al., "Intelligent edge computing in internet of vehicles: a joint computation offloading and caching solution," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2020.

[10] S. Huang, C. Yang, S. Yin, Z. Zhang, and Y. Chu, "Latency-aware task peer offloading on overloaded server in multi-access edge computing system interconnected by metro optical networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 38, no. 21, pp. 5949–5961, 2020.

[11] Z. Ning, P. Dong, X. Wang et al., "Mobile edge computing enabled 5G health monitoring for internet of medical things: a decentralized game theoretic approach," *IEEE Journal on Selected Areas in Communications, To Appear*, pp. 1–6, 2020.

[12] Z. Ning, P. Dong, X. Wang et al., "Partial computation offloading and adaptive task scheduling for 5G-enabled vehicular networks," *IEEE Transactions on Mobile Computing*, p. 1, 2020.

[13] W. Wang, H. Huang, L. Zhang, and C. Su, "Secure and efficient mutual authentication protocol for smart grid under blockchain," *Peer-to-Peer Networking and Applications*, 2020.

[14] Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2017.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[16] A. Graves, G. Wayne, M. Reynolds et al., "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, no. 7626, pp. 471–476, 2016.

[17] C. Farabet, C. Poulet, J. Han, and Y. LeCun, "CNP: an FPGA-based processor for convolutional networks," in *IEEE International Conference on Field Programmable Logic and Applications*, pp. 32–37, Prague, Czech Republic, 2019.

[18] N. P. Jouppi, C. Young, N. Patil et al., "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th Annual International Symposium on Computer Architecture*, pp. 1–12, Toronto, ON, Canada, 2017.

[19] A. Shafiee, A. Nag, N. Muralimanohar et al., "ISAAC: a convolutional neural network accelerator with in-situ analog arithmetic in crossbars," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 14–26, 2016.

[20] L. Guo, Z. Ning, W. Hou, B. Hu, and P. Guo, "Quick answer for big data in sharing economy: innovative computer architecture design facilitating optimal service-demand matching," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 4, pp. 1494–1506, 2018.

[21] P. Guo, W. Hou, L. Guo, Q. Yang, Y. Ge, and H. Liang, "Low insertion loss and non-blocking microring-based optical router for 3d optical network-on-chip," *IEEE Photonics Journal*, vol. 10, no. 2, pp. 1–10, 2018.

[22] J. Feldmann, N. Youngblood, C. Wright, H. Bhaskaran, and W. H. P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature*, vol. 569, no. 7755, pp. 208–214, 2019.

[23] P. Guo, W. Hou, L. Guo et al., "Fault-tolerant routing mechanism in 3d optical network-on-chip based on node reuse," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 3, pp. 547–564, 2020.

[24] Y. Shen, N. C. Harris, S. Skirlo et al., "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, vol. 11, no. 7, pp. 441–446, 2017.

[25] L. Chen, K. Preston, S. Manipatruni, and M. Lipson, "Integrated GHz silicon photonic interconnect with micrometer-scale modulators and detectors," *Optics Express*, vol. 17, no. 17, pp. 15248–15256, 2009.

[26] Z. Ying, C. Feng, Z. Zhao et al., "Electronic-photonic arithmetic logic unit for high-speed computing," *Nature Communications*, vol. 11, no. 1, article 2154, 2020.

[27] Z. Ying, Z. Wang, Z. Zhao et al., "Silicon microdisk-based full adders for optical computing," *Optics Letters*, vol. 43, no. 5, pp. 983–986, 2018.

[28] T. Baba, S. Akiyama, M. Imai et al., "50-Gb/s ring-resonator-based silicon modulator," *Optics Express*, vol. 21, no. 10, pp. 11869–11876, 2013.

[29] J. Michel, J. Liu, and L. C. Kimerling, "High-performance Ge-on-Si photodetectors," *Nature Photonics*, vol. 4, no. 8, pp. 527–534, 2010.

[30] Y. Urino, Y. Noguchi, M. Noguchi et al., "Demonstration of 12.5-Gbps optical interconnects integrated with lasers, optical splitters, optical modulators and photodetectors on a single silicon substrate," *Optics Express*, vol. 20, no. 26, pp. B256–B263, 2012.

[31] H. Jia, L. Zhang, J. Ding, L. Zheng, C. Yuan, and L. Yang, "Microring modulator matrix integrated with mode multiplexer and de-multiplexer for on-chip optical interconnect," *Optics Express*, vol. 25, no. 1, pp. 422–430, 2017.

[32] Z. Ying, S. Dhar, Z. Zhao et al., "Electro-optic ripple-carry adder in integrated silicon photonics for optical computing," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 6, pp. 1–10, 2018.

[33] J. Dong, A. Zheng, D. Gao et al., "High-order photonic differentiator employing on-chip cascaded microring resonators," *Optics Letters*, vol. 38, no. 5, pp. 628–630, 2013.

[34] M. Ferrera, Y. Park, L. Razzari et al., "On-chip CMOS-compatible all-optical integrator," *Nature Communications*, vol. 1, no. 1, article 29, 2010.

[35] L. Yang, R. Ji, L. Zhang, J. Ding, and Q. Xu, "On-chip CMOS-compatible optical signal processor," *Optics Express*, vol. 20, no. 12, pp. 13560–13565, 2012.

[36] F. Liu, H. Zhang, Y. Chen, Z. Huang, and H. Gu, "WRH-ONoC: a wavelength-reused hierarchical architecture for optical network on chips," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 1912–1920, Kowloon, Hong Kong, April 2015.

[37] P. Guo, W. Hou, L. Guo, Z. Cao, and Z. Ning, "Potential threats and possible countermeasures for photonic network-on-chip," *IEEE Communications Magazine*, vol. 58, no. 9, pp. 48–53, 2020.

[38] P. Guo, W. Hou, L. Guo, Z. Ning, M. S. Obaidat, and W. Liu, "WDM-MDM silicon-based optical switching for data center networks," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pp. 1–6, Shanghai, China, May 2019.

[39] W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, and L. Jiang, "Holylight: a nanophotonic accelerator for deep learning in data centers," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1483–1488, Florence, Italy, March 2019.

[40] W. Bogaerts, P. de Heyn, T. van Vaerenbergh et al., "Silicon microring resonators," *Laser & Photonics Reviews*, vol. 6, no. 1, pp. 47–73, 2012.

[41] P. Guo, W. Hou, and L. Guo, "Designs of low insertion loss optical router and reliable routing for 3D optical network-on-chip," *Science China Information Sciences*, vol. 59, no. 10, article 102302, 2016.

[42] A. Sampson and M. Buckler, "FODLAM, a first-order deep learning accelerator model," https://github.com/cucapra/fodlam.

[43] https://www.lumerical.com/cn/.

[44] A. N. Tait, T. F. de Lima, E. Zhou et al., "Neuromorphic photonic networks using silicon photonic weight banks," *Scientific Reports*, vol. 7, no. 1, article 7430, 2017.