IEICE TRANSACTIONS

on Fundamentals of Electronics, Communications and Computer Sciences

VOL. E104-A NO. 11
NOVEMBER 2021

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER Special Section on Circuits and Systems

Neural Network Calculations at the Speed of Light Using Optical Vector-Matrix Multiplication and Optoelectronic Activation*

Naoki HATTORI[†a)], ***Nonmember***, **Jun SHIOMI**[††], **Yutaka MASUDA**[†], **Tohru ISHIHARA**[†], ***Members***, **Akihiko SHINYA**[†††,††††], ***and*** **Masaya NOTOMI**[†††,††††], ***Nonmembers***

SUMMARY With the rapid progress of the integrated nanophotonics technology, the optical neural network architecture has been widely investigated. Since the optical neural network can complete the inference processing just by propagating the optical signal in the network, it is expected more than one order of magnitude faster than the electronics-only implementation of artificial neural networks (ANN). In this paper, we first propose an optical vector-matrix multiplication (VMM) circuit using wavelength division multiplexing, which enables inference processing at the speed of light with ultra-wideband. This paper next proposes optoelectronic circuit implementation for batch normalization and activation function, which significantly improves the accuracy of the inference processing without sacrificing the speed performance. Finally, using a virtual environment for machine learning and an optoelectronic circuit simulator, we demonstrate the ultra-fast and accurate operation of the optical-electronic ANN circuit.

key words: *neural network, optical circuit, multi-layer perceptron, wavelength division multiplexing*

1. Introduction

Today's highly sophisticated information society, with low latency access to the Internet, would not be realizable without optical communication technologies and CMOS LSI technologies. According to Moore's Law, the propagation delay of CMOS gates in the LSI circuits has drastically decreased. Historically, the delays of local level wires also decreased with transistor downscaling since the delays are determined by RC time constant which can be reduced along the transistor downscaling. At ultra-scaled dimensions, however, the effective resistivities (R) of local level wires increase more rapidly than a decrease of wire capacitance (C) due to size effects [2] and therefore, the RC time constant cannot be decreased by the transistor downscaling. Post-layout analysis using predictive technology models [3] shows that interconnect performance degradation may dominate over the device speed improvement in a 22 nm technology node and below [2], [4]. This means that technology scaling itself cannot resolve the latency issue of CMOS LSI circuits in advanced technology nodes such as 7 nm and below.

Concurrently, optical communication technologies have also been rapidly growing over the past several decades. Although optical communication technologies are widely used for the long-distance communications, electronics still remain as major players for short-distance communications. Recent innovation in nanophotonics, however, makes it possible to migrate power-efficient light-based communication into ever-shorter distances and move onto silicon chips as optical networks-on-chip [5]. More recently, significant efforts have been made on architecture development of optical neural networks (ONN) [6]–[11]. We have also proposed an ONN architecture in [1] based on an optical vector-matrix multiplication (VMM) circuit which can fully exploit the ultra-fast nature of light.

This paper is an extension of our previous work [1]. An overview of the architecture is depicted in Fig. 1. The architecture is composed of activation circuits, the optical VMM circuit, and micro-ring resonators for interfacing the weight matrix. Although ultra-fast inference is mainly achieved by the optical VMM, batch normalization and activation function are integral parts for accurate inference processing in artificial neural network (ANN). One of the major challenges for the ONN is developing an ultra-low latency circuit for an activation function to achieve sufficient inference accuracy while maintaining the speed of light. To achieve this goal, this paper newly proposes optoelectronic circuit implementation for batch normalization combined with activation function, which significantly improves the accuracy of the inference processing without sacrificing the speed performance. The detailed circuit for the batch normalization is explained in Sect. 4.5. Another key challenge for the ONN is to reduce circuit footprint for providing sufficient scalability. Towards this challenge, we use ternary values for weights in the ONN and propose an area efficient interface circuit for

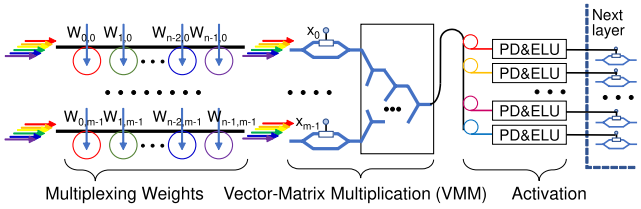


Fig. 1 Overview of our optical neural network.

Manuscript received November 23, 2020.
Manuscript revised March 27, 2021.
Manuscript publicized May 17, 2021.

[†]The authors are with Nagoya University, Nagoya-shi, 464-8601 Japan.
[††]The author is with Kyoto University, Kyoto-shi, 606-8501 Japan.
[†††]The authors are with NTT Nanophotonics Center, Atsugi-shi, 243-0198 Japan.
[††††]The authors are with NTT Basic Research Laboratories, Atsugi-shi, 243-0198 Japan.

*This work was partly presented in the conference [1].

a) E-mail: naok1_h@ertl.jp

DOI: 10.1587/transfun.2020KEP0016

the weights. This largely reduces the circuit area required for interfacing the weight matrices. The detailed structure for the interface circuit is explained in Sect. 4.2.

The rest of the paper is organized as follows. Section 2 summarizes several previous works on the ONNs. Section 3 explains basic operation of optical arithmetic circuits used in our ONN. Section 4 shows a detailed architecture of our ONN. Section 5 shows experimental results obtained with a commercial optoelectronic circuit simulator and a virtual environment for machine learning. Section 6 concludes this paper.

2. Related Work and Motivation

Neural networks are a continued staple of machine learning and alternative computing, with applications ranging from classification, anomaly detection and regression to general-purpose computation. The following subsections summarize recent architectures proposed for optical neural networks.

2.1 Photonic Weight Banks

An optical circuit structure for calculating a weighted sum is proposed in [6]. This structure is used for vector-matrix multiplication in optical neural networks. The basic structure is depicted in Fig. 2. Incoming wavelength division multiplexed (WDM) signals are weighted by continuous-valued filters called microring (MRR) weight banks and then summed as photocurrent by a photodetector. This is a very area efficient and low power approach for calculating the weighted sum. The complexity is $O(1)$ regardless of the number of weights. However, this approach has the following drawbacks. If more than one optical signals having different wavelengths (i.e., WDM signals) are given to the photodetector (PD), undesirable oscillation in photocurrent occurs. One of the most straightforward approaches to eliminate the oscillation is low-pass filtering with an electronic low-pass filter. This is very simple but prevents us from exploiting the ultra-high speed nature of lights since the time-constant of the low-pass filter is more than one order of magnitude bigger than that of optical-electrical-optical (O-E-O) repeater. Another approach for eliminating the oscillation is using wavelengths which are sufficiently apart from each other. Since the oscillation frequency depends on the difference between the wavelengths of the lights, we can make the frequency to be too high for the photodetector to oscillate by setting the wavelengths sufficiently apart from each other. However, this approach limits the number of different wavelengths used in WDM signals and limits the scalability of the vector-matrix multiplication. In [7]–[9], similar optical vector-matrix multipliers based on the MRR weight banks are proposed. The architectures are very compact and implement the $O(1)$ calculation of the weighted sum using WDM signals and PD-based accumulation. However, they also have the oscillation issue in the photodetector which prevents the light-speed operation of optical vector-matrix multiplication.

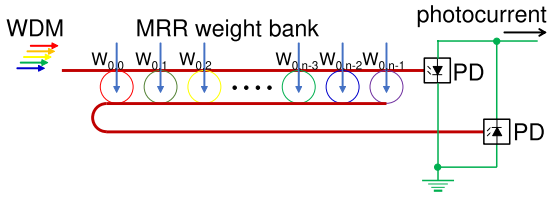


Fig. 2 Calculation of weighted sum using photonic weight banks.

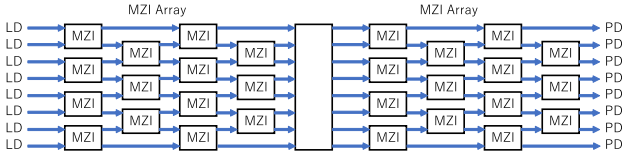


Fig. 3 Calculation of weighted sum using MZI array.

The primary goal of the approaches described above is providing an area-efficient ONN architecture while the goal of our approach proposed in this paper is providing an ultra-fast ONN architecture. Unlike the works presented in [6]–[9], we use a combiner tree [12] for accumulating WDM signals in parallel instead of using PD-based accumulation. Our approach does not have the oscillation issue in the photocurrent, since only coherent light signals having a single wavelength are given to a photodetector.

2.2 Reconfigurable Mach-Zehnder Interferometer Array

In [10], a fully optical neural network (ONN) architecture is presented for implementing general deep neural network algorithms using nanophotonic circuits that process coherent light from laser diode (LD). The core part of the ONN architecture is a matrix multiplication unit which is composed of a reconfigurable Mach-Zehnder Interferometer (MZI) array as shown in Fig. 3. Once a neural network is trained, the architecture can be passive, and computation can be performed at the speed of light. In addition to the light-speed processing, the computation can be performed without additional energy input. These features could enable ONNs that are substantially more energy-efficient and faster than their electronic counterparts. As described in [10], the energy consumption introduced by the switching activity is extremely small in this architecture. However, one big drawback in the ONN architecture described above is high photonic component utilization and area cost. Considering a single fully-connected layer with an $n \times m$ weight matrix, the ONN architecture in [10] requires $O(n^2 + m^2)$ MZIs for implementation. If the number of neurons in the network increases, the area for the implementation increases quadratically. In [11], a more compact ONN architecture based on fast Fourier transform (FFT) is proposed. It improves the area efficiency of the ONN by a factor of 2.2 to 3.7. However, the area required for the implementation is still very large as the number of MZIs required is still quadratic to the number of neurons.

Unlike the architecture in [10], [11], our architecture does not have the circuit structure where the number of MZIs

required is quadratic to the number of neurons. Although our architecture also exploits the ultra-high speed nature of the Mach-Zehnder Modulator (MZM), the order of the number of MZMs required for the implementation is linear to the number of neurons.

2.3 Homodyne Detection-based Vector Matrix Multiplier

A new type of photonic accelerator based on coherent detection is proposed in [13]. It is scalable to large ($N \geq 10^6$) networks and can be operated at high speeds (GHz) and very low energies (sub-aJ) per multiply-and-accumulate (MAC), using the massive spatial multiplexing enabled by standard free-space optical components. In contrast to previous approaches [10], both weights and inputs are optically encoded so that the network can be reprogrammed and trained on the fly. However, it does not exploit WDM parallelism in their multiply-and-accumulate operation, which limits the throughput and area efficiency of the architecture. A technique proposed in [9] combines the WDM method with the coherent detection to exploit the optical parallelism. However, as explained in Sect. 2.1, it suffers from the oscillation issue in photodetectors, which limits the scalability of high-speed vector-matrix multiplication.

Unlike the architecture presented above, our ONN architecture largely exploits WDM parallelism as well as the circuit-level spatial parallelism using an optical combiner tree which can be functioned for accumulating WDM signals in parallel.

3. Basics of Optical Arithmetic Operation

3.1 Analog Multiplier based on Mach-Zehnder Modulator

We use a Mach-Zehnder Modulator (MZM) as an optical multiplier. The MZM is a popular device as an analog multiplier to calculate the product of an electrical voltage input and an optical signal input [9]. By using optical signals with different wavelengths to each other, the MZM can work as a parallel multiplier for the multiple optical signals as shown in Fig. 4(a) without mutual interference. It is experimentally demonstrated that optical inter-channel interference is negligible for channels with a wavelength spacing of 1.3 nm [14]. In [15], a broadband silicon Mach Zehnder Switch (MZS in the following) which operates over a wide wavelength range from 1510 nm to 1650 nm is proposed. In this case, the MZS can be functional for more than 100 WDM optical signals.

3.2 Analog Adder based on Optical Combiner

We use a combiner-tree-based analog accumulator presented in [12]. The analog accumulation operation is performed for massive WDM signals in parallel as shown in the tree structure depicted in Fig. 4(b). At every combiner in the tree, two input signals with the same wavelength interfere and add up together while for any two input signals with different wavelengths do not interfere. As a result, all WDM signals are

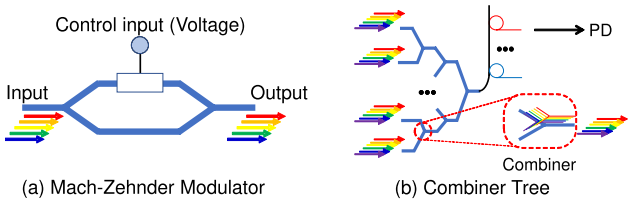


Fig. 4 Optically parallelized multiplication and accumulation.

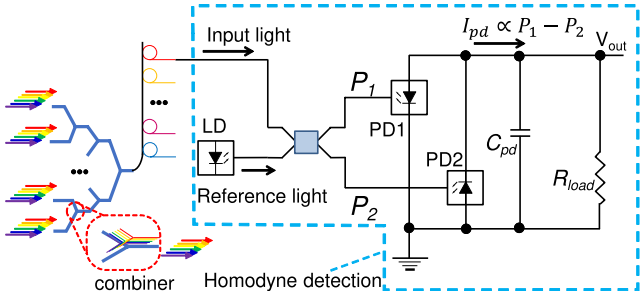


Fig. 5 Combiner-tree and homodyne detection.

individually accumulated in parallel through the combiner-tree. We use a phase shift for representing a minus value. Assuming a base optical signal A represents a plus value and if the phase of an optical signal B is π (i.e., 180 degree) out of phase from the base signal, the signal B represents a minus value. With this representation for the plus and minus values, we can correctly accumulate both plus and minus values using the combiner-tree.

3.3 Arithmetic Operation based on Electric Field Strength

Since the output value of our multiplication and accumulation (MAC) circuit is expressed by the electric field strength instead of the signal power, we need to extract the electric field strength from the optical signal. We use a homodyne detection circuit shown in Fig. 5 for extracting the electric field strength from the photocurrent which is proportional to the signal power. We use O-E converter presented in [16] for the homodyne detection. This O-E converter does not need an amplifier to convert the optical signal to the electrical signal and therefore, it is very energy efficient. Once the photocurrent I_{pd} in Fig. 5, which is proportional to the electric field strength of input light is obtained, the I_{pd} is converted to electrical voltage just using a load resistance R_{load}. If the input capacitance of the device connected to the V_{out} in Fig. 5 is an order of femtofarad, the RC time constant can be very small and the O-E-O conversion delay is, as a result, around 25 picosecond [16].

4. Optical Neural Network Using WDM Parallelism

4.1 Neural Network Overview

An architectural overview of our optical neural network is depicted in Fig. 6. Each layer is composed of MRR weight

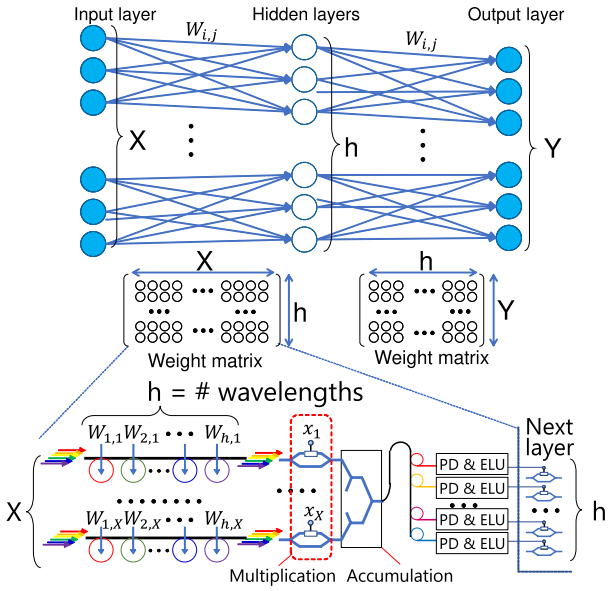


Fig. 6 Optical neural network overview.

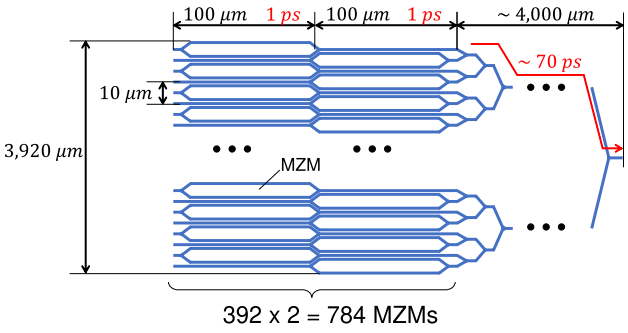


Fig. 7 Propagation delay of optical vector-matrix multiplier.

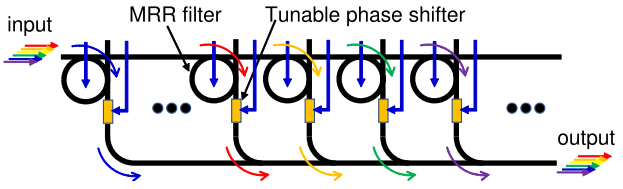


Fig. 8 Interface for ternary weight matrix.

bank [6], parallelized Mach Zehnder Modulator (MZM)-based multipliers, a combiner-tree-based accumulator [12], and optoelectronic activation circuits. The output signals of the activation circuits are passed to the next layer as electrical voltage signals which are used as input operands of the MZM-based multipliers in the next layer.

Assuming the number of nodes in the input layer is X and that in the next layer is h, we need $X \times h$ micro-ring array as the MRR weight bank. The number of different wavelengths needed is h and the number of rows in the bank is X. The h different light signals are given to every row in parallel. Note that the signal power of the h lights given as inputs is the same from each other. Then the $X \times h$ weight values are individually multiplied to the light signals in parallel. Once the weighted signals are given, the MZM works as a parallel multiplier. The number of MZMs is X and the number of wavelengths used in the WDM signals given to each MZM is h. Therefore, $X \times h$ multiplications are performed in the MZM-based multipliers in parallel by exploiting both spatial parallelism and WDM parallelism. The next step is accumulation. Once the outputs of the MZM-based multipliers are given to the combiner-tree as WDM signals, accumulation operations are performed in parallel. Optical signals with the same wavelength are accumulated in a tournament fashion in the combiner-tree. This accumulation is performed for every different wavelength in parallel. Since the optical signals with different wavelengths do not basically interfere with each other, the h different accumulations can be independently performed in parallel. Finally, the accumulated values are extracted by micro-ring resonators, wavelength selective splitters or arrayed waveguide gratings (AWG), and given to activation circuits separately. The number of activation circuits needed in the layer is h. The outputs of the activation circuits are passed to the next layer as inputs of the MZM-based multiplier in the next layer.

The MRR weight bank takes a few tens of picoseconds to read weight parameters [6]. Note that the process for reading the weights through the MRR weight bank is not on the critical-path since the weight values are unchanged once the weights have been learned. Therefore, the delay of the optical neural network calculations is determined by the delay required for the VMM calculation and activation. Since the VMM calculation is performed by just propagating the optical signals through MZM multipliers and a combiner-tree, the delay is determined by the total length of the critical-path along the MZM and the combiner-tree as shown in Fig. 7. Note that the speed of light traveling through the MZM and the combiner is about 100 μm per picosecond. The MZM multiplier takes only a few picoseconds to propagate the optical signals since the length of the MZM is about 100 μm. The delay of the combiner-tree depends on the number of inputs [12]. If the number of inputs is 784, the delay is about 70 ps as shown in Fig. 7. If the number of inputs is 100 as is used in the hidden layers, the delay is down to about 8 ps. The ELU-based activation circuit explained in the following sections also takes 25 ps [16]. As a result, the propagation delay of the single neural network layer is less than 100 picoseconds in total approximately and if we construct a neural network with less than 10 layers, a sub-nanosecond neural network is realizable, which is extremely fast.

4.2 Interface for Quantized Weight Matrix

In this paper, we use a quantized weight matrix to save a circuit area required for implementation. Specifically targeting a ternary weight matrix, we propose a compact interface circuit for the weight matrix. Each element of the matrix can have ternary values of -1, 0, or 1. These values are encoded with two bits of 11, 00, and 01, respectively. Figure 8 shows the interface circuit. The right and left bits of the code are

applied to the MRR filter and tunable phase shifter in Fig. 8, respectively. If the weight is 0 whose code is 00, the optical input with a wavelength corresponding to the MRR filter does not come to the output. If the weight is 1 whose code is 01, the input with a wavelength corresponding to the MRR filter comes to the output without a phase shift. If the weight is −1 and the code is 11, the optical signal with a wavelength corresponding to the MRR filter comes to the output with a phase shift of a 180 degree from the base signal which represents the positive values. As we explained in Sect. 3.2, we use a phase shift by a 180 degree for representing a minus value. With this representation for the minus values, we can correctly accumulate both plus and minus values using a combiner-tree.

4.3 Vector Matrix Multiplication

We perform vector-matrix multiplication as shown in Eq. (1), where a set of $W_{i,j}$ values forms a weight matrix and a set of x_i values forms an input vector.

$$\begin{aligned} y_1 &= x_1 \cdot W_{1,1} + x_2 \cdot W_{1,2} + \cdots + x_X \cdot W_{1,X}, \\ y_2 &= x_1 \cdot W_{2,1} + x_2 \cdot W_{2,2} + \cdots + x_X \cdot W_{2,X}, \\ \cdots &= \quad \cdots \quad + \quad \cdots \quad + \cdots + \quad , \\ y_h &= x_1 \cdot W_{h,1} + x_2 \cdot W_{h,2} + \cdots + x_X \cdot W_{h,X}. \end{aligned} \tag{1}$$

We use an optical vector-matrix multiplier shown in the bottom of Fig. 6 as a core part of the optical neural network. The input WDM signals are first equally divided into X groups and given to the rows of the MRR weight bank. Therefore, the power values of all the $X \times h$ optical signals given to the MRR weight bank are equal to each other. These WDM signals are then multiplied by the weight parameters. For example in the topmost row, $W_{1,1}, W_{2,1}, \cdots, W_{h-1,1}, W_{h,1}$ are individually weighted by the micro-ring modulators. Then the WDM output of the row is passed to the MZM which is controlled by the electric voltage signal proportional to the value of x_1. Since the input of the MZM is weighted WDM signals which do not interfere each other, the value of x_1 is individually and concurrently multiplied to all the WDM signals. This multiplication corresponds to the first terms of all the equations in Eq. (1). The MZM controlled by x_2 performs the multiplication corresponding to the second terms of all the equations in Eq. (1). Similarly, all the multiplications which appear in the equations in Eq. (1) are performed throughout the MZMs in parallel.

All the output WDM signals of the MZMs are passed to the combiner-tree with being multiplexed for the optically parallelized accumulation. The number of leaves needed for the combiner-tree in this example is X. Since it is a tree structure, the order of the propagation delay in this combiner-tree is $O(log_2 X)$. The propagation delay of each combiner is very small since the size of the combiner can be made very small [12]. Optical signals with the same wavelength are accumulated in a tournament fashion by the combiner-tree. This accumulation corresponds to each equation in Eq. (1). For given WDM signals to the combiner-tree, all the accumulations in Eq. (1) are performed individually and concurrently.

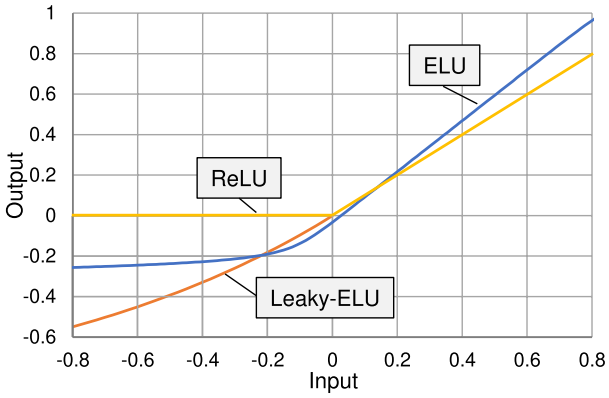


Fig. 9 Comparison of activation functions.

4.4 Activation Function

Figure 9 compares the representative activation functions, i.e., Rectified Linear Unit (ReLU) [17], exponential linear unit (ELU) [18], and the Leaky-ELU. ReLU is the most commonly used activation function in neural networks, especially in CNNs since it is cheap to compute as there is no complicated math. Although ReLU has several advantages over the other activation functions, it has a problem called dying ReLU. A ReLU neuron is dead if the input is stuck on the negative side and always outputs 0 since the output y of ReLU is $y = max(0, x)$ for the input x. Because the slope of ReLU in the negative range is also 0, once a neuron gets negative, it is unlikely for it to recover. Leaky ReLU [19] has been proposed to fix the dying ReLU problem. It has a small slope for negative values, instead of altogether zero. For example, leaky ReLU may have $y = 0.01x$ when $x < 0$. Similar to leaky ReLU, ELU has a small slope for negative values. Instead of a straight line, it uses a log curve for negative values as shown in Eq. (2), where α is a scaling factor.

$$\begin{aligned} y &= x && (\text{if } x \geq 0), \\ y &= \alpha(e^x - 1) && (otherwise). \end{aligned} \tag{2}$$

It is designed to combine the good parts of ReLU and leaky ReLU, that is, while it does not have the dying ReLU problem, it saturates for large negative values, allowing them to be essentially inactive.

In many neural network applications, ReLU, leaky ReLU and ELU do not have a big difference in training speed and inference accuracy. However, ELU is the most suitable activation function for circuit implementation since it is differentiable at every point and the output changes gently. Figure 10 shows an example of optoelectronic implementation of ELU. This circuit is modified from the homodyne detection circuit depicted in Fig. 5. The yellow colored diode symbol, an ELU diode which is newly added, works as an ELU function. For positive input values, the ELU diode

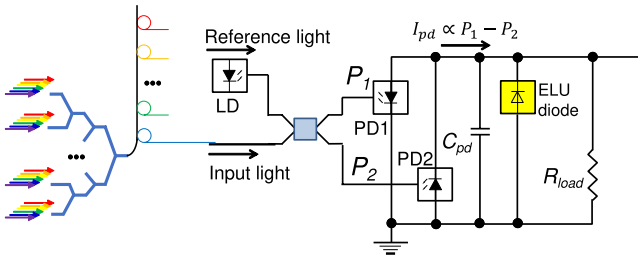


Fig. 10 Optoelectronic implementation of activation function.

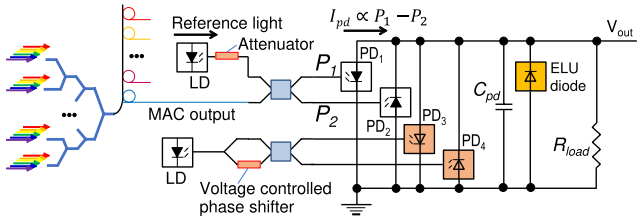


Fig. 11 Activation function circuit with batch normalization.

works as an insulator while it works as a resistor where its resistance is very small for negative input values. Since the circuit presented in Fig. 10 does not behaves exactly as the ELU function, we refer to this function as a Leaky-ELU. In order to use the Leaky-ELU function in the machine learning process, we perform the regression analysis and fit the Leaky-ELU function model to the simulation results with the SPICE-based circuit as shown in Fig. 10. The curve fitting accuracy is very good. The coefficient of determination, which represents how well the regression predictions approximate the real data points is more than 99.9% for the Leaky-ELU function model.

4.5 Batch Normalization

Batch normalization [20] is a commonly used technique to normalize activations using a scaling factor γ and a shifting factor β in intermediate layers of neural networks for improving inference accuracy. The values of β and γ are learned for every neurons individually in the learning process. At the inference phase, every outputs of the vector-matrix multiplication are individually scaled by γ and shifted by β to normalize the activations. Figure 11 shows a circuit for the batch normalization. This circuit is modified from the activation circuit depicted in Fig. 10. Since the signal power values of P_1 and P_2 are both proportional to the signal power of the laser diode (LD) used as the reference light, the photocurrent I_{pd} can be scaled by tuning the attenuator connected to the LD. The photocurrent I_{pd} can be shifted by β using PD_3 and PD_4. The photocurrent drawn from PD_3 and PD_4 can be controlled by tuning the input voltage to the voltage controlled phase shifter shown in Fig. 11.

5. Experimental Evaluation

This section evaluates the area and inference accuracy of the proposed architecture. Section 5.1 estimates the trade-off relationship between the area and accuracy with a Python-based deep learning simulator, i.e., TensorFlow. Section 5.2 examines the functional behavior of the proposed architecture via the optoelectronic circuit simulator.

5.1 TensorFlow Simulation

As previously explained in Sect. 4, the proposed architecture integrates the highly-parallelized VMM thanks to WDM and the quantized weight matrix interface. To evaluate the proposed architecture's area-efficiency, we examine the inference accuracy and circuit area with and without quantization in this subsection. Section 5.1.1 explains the evaluation setup for Tensorflow simulation. Section 5.1.2 shows the evaluation results for the trade-off analysis between the circuit area and inference accuracy. Section 5.1.3 discusses the impact of activation function and Batch Normalization on the inference accuracy.

5.1.1 Evaluation Setup

As a target circuit, we select the multi-layer perceptron (MLP) with MNIST dataset [21]. This neural network consists of one input layer, several hidden layers, and one output layer. The input layer has 784 (28×28) nodes, and the output layer has 10 nodes. As the activation function, ELU is adopted to each node in hidden layers. Besides, the Batch Normalization is added before the activation function in hidden layers. We introduce stochastic quantization [22] and quantize the values of the weight matrices to binary (1-bit) or ternary (2-bit), which eliminates the need for a digital-to-analog converter (DAC) and thus significantly saves the circuit area. By stochastic quantization of the weights, it is expected that the information of continuous weights can be reproduced to some extent in calculating weighted linear sums at each layer during inference.

We construct a test framework of the target circuit using TensorFlow, the Python-based open-source simulator. We initialize all trainable weights with a random uniform initializer, adopt the Adam optimizer [23] with initial learning rate=3E-03 and a stepwise exponential-decay learning rate schedule with decay rate=0.99. All NN models are trained for 1,000 epochs with a minibatch size of 100 until fully converged. The TensorFlow is run on a computer with an AMD Ryzen Threadripper 2920X processor under the Ubuntu 16.04.6 LTS operating system with 128-GB memory.

We evaluate the inference accuracy and the circuit area of the test circuits with the different number of nodes in the middle layers, the different number of hidden layers, and the different quantization precision. The numbers of nodes in the middle layers evaluated are 4,096, 2,048, 1,024, 512, 256, 128, and 100, as shown in Fig. 12. The evaluated numbers of hidden layers are ranging from one to seven, as shown in Fig. 13. In the experiment, we prepared three different precision scenario, i.e., 32-bit floating-point without

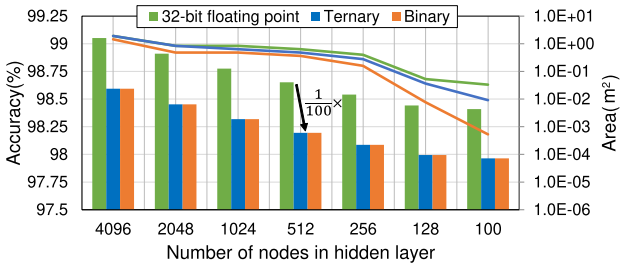


Fig. 12 Inference accuracy and area on MNIST dataset with different network configurations, i.e., the number of nodes in hidden layers and the quantization precision. The number of hidden layers is set to three. The line graphs show accuracy, and the bar graphs show the optical circuit area. Binary and ternary represent the quantization precision for the weights.

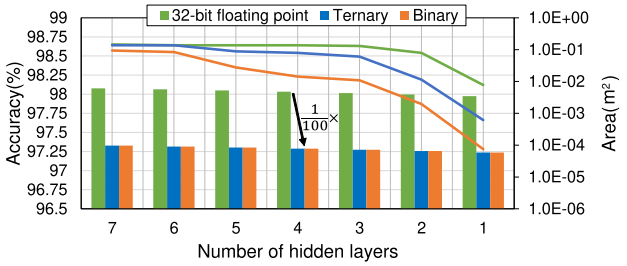


Fig. 13 Inference accuracy and area on MNIST dataset with different network configurations, i.e., the number of nodes in hidden layers and the quantization precision. The number of nodes in every hidden layer is set to 100.

Table 1 Optical component sizes used in the area estimation.

Optical Component	Length(μm)	Width(μm)
Digital-to-Analog Converter (DAC) [24]	39	1,100
Micro-Ring Resonator (MRR) [6]	25	25
Mach-Zehnder Modulator (MZM) [10]	10	100

quantization, ternary, and binary. In the ternary and binary quantized weights, the weights are quantized into $(-1, 0, 1)$ and $(-1, 1)$, respectively. Note that the maximum number of wavelengths multiplexed in an optical arithmetic unit is about 100 [15].

To evaluate the entire circuit area, we estimate the area in each layer from Eq. (3). The circuit of each layer is mainly composed of MZMs, MRRs, and DACs, and the number of these components depends on the number of nodes. Therefore, by estimating the total area of these components, we can approximate the total area of the proposed circuit. In Eq. (3), N_{node}^{1st} and N_{node}^{2nd} are the numbers of nodes in the first and the next second layer. As for the area of DAC, MRR, and MZM, i.e., A_{DAC}, A_{MRR}, and A_{MZM}, we estimate the area by referring to Table 1.

$$\begin{aligned} Area =&(N_{node}^{1st} \times A_{MZM}) + (N_{node}^{1st} \times N_{node}^{2nd} \times A_{MRR}) \\ &+ (N_{node}^{1st} \times N_{node}^{2nd} \times A_{DAC}). \end{aligned} \tag{3}$$

5.1.2 Accuracy and Area Estimation Results

Figure 12 shows the correlation between the optical circuit area, the number of nodes in the hidden layer, and the inference accuracy in MLP. In Fig. 12, line graphs show the accuracy, and bar graphs show the optical circuit area. Besides, "Binary" and "Ternary" represent quantization precision the weights, and "32-bit floating-point (fp32)" corresponds to the weight matrix without quantization. As shown in Fig. 12, the accuracy degrades as the number of nodes decreases, and the "Ternary" achieves a sufficiently good accuracy, which is a negligible degradation compared with "fp32". For example, based on "fp32 - 4,096 nodes", the inference accuracy with "fp32 - 100 nodes" degrades only 0.39% from 99.04% to 98.63%, and that with "Ternary-100 nodes" does about 0.55% from 99.04% to 98.49%. As the number of nodes in the hidden layer increases, the number of elements in the weight matrix increases. Thus, the number of optical components, such as MRR and MZM used in weighting and multiplication, also increases. Note that the number of DACs is the same as that of weight matrix elements since each weight matrix value is read individually.

Another key observation of Fig. 12 is that the optical circuit area decreases exponentially as the number of nodes decreases. For example, compared with "fp32 - 2,048nodes", the area of "fp32 - 1,024-node" is reduced by 70.8% from 4.36E-01 m^2 to 1.27E-01 m^2, and the area of "fp32 - 512-node" is reduced by 90.7% from 3.09E-01 m^2 to 4.05E-02 m^2. Since the maximum number of wavelengths multiplexed in an optical arithmetic complement is equal to 100, the same circuit needs to be added for every time the number of nodes, which is the same as that of wavelength, exceeds 100. Moreover, when the weight matrix is quantized, the area is reduced by about two orders of magnitude. For example, while the area of "fp32 - 100-node" is 4.34E-03 m^2, the area of "Ternary - 100-node" is 7.16E-05 m^2.

Figure 13 shows the correlation between the optical circuit area and the number of hidden layers and the inference accuracy in MLP. When the number of hidden layers is more than 3, the inference accuracy can be maintained even when the number of nodes is 100. For example, based on "fp32 - 7-layer", the accuracy with "fp32 - 3-layer" decreases by only 0.02%, from 98.65% to 98.63%. We also observe that the area can be effectively reduced by quantizing the weight matrix to binary or ternary. From the observation above, we experimentally confirm that if the application allows for some accuracy degradation, the optical circuit area can be dramatically reduced by quantizing the weight matrix to ternary. We note that the better inference accuracy can be obtained without quantization in the scenario where the area is not a top-priority constraint.

5.1.3 Discussion

The previous section experimentally evaluates the proposed architecture's trade-off characteristics in terms of the area

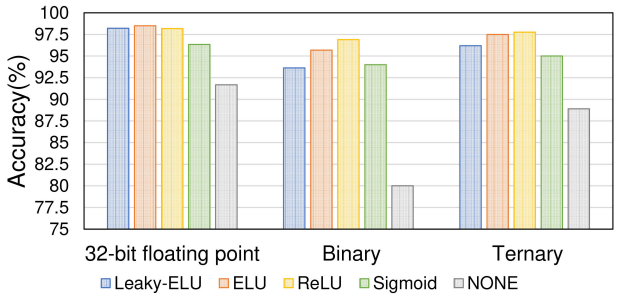


Fig. 14 Inference accuracy comparisons with configurations using different activation functions without Batch Normalization based on MNIST dataset. The MLP consists of 784 (28×28) inputs, three hidden layers with 100 neurons in each layer, and 10 neurons for the last layer. NONE indicates that the activation function is not used for MLP calculations. The proposed architecture adopts Leaky-ELU as an activation function.

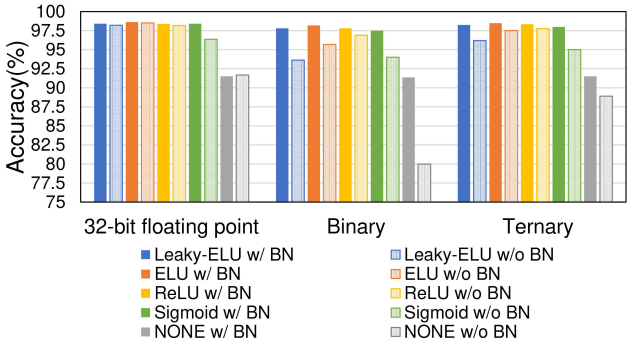


Fig. 15 Inference accuracy comparisons of different configurations with and without Batch Normalization based on MNIST dataset. The MLP configurations are the same as Fig. 14 (i.e., the number of hidden layers is 3 and the number of nodes in every hidden layer is 100).

and inference accuracy. In this section, we focus on the inference accuracy and discuss the impact of activation function and Batch Normalization for investigating the effectiveness of the proposed architecture.

Firstly, let us investigate the impact of activation function on accuracy. In the experiment, we compare the inference accuracy with different activation functions without Batch Normalization. Specifically, we use Leaky-ELU, which was proposed in Sect. 4.4, and three representative activation functions, i.e., ReLU, ELU, and Sigmoid. Figure 14 shows the comparison results. From Fig. 14, we can see that the inference accuracy of ReLU and ELU is the highest in all network configurations. In 32-bit floating-point and Ternary cases, Leaky-ELU achieves a high inference accuracy with little difference compared to ReLU and ELU. Here, please remind that Leaky-ELU is an implementation-friendly function. Therefore, we experimentally confirm that the proposed activation function achieves a similar inference accuracy as ReLU while more suitable to implement.

Then, the impact of Batch Normalization on the accuracy is discussed. Figure 15 shows the inference accuracy with and without Batch Normalization. The graphs in darker colors indicate the inference accuracy of MLPs with Batch Normalization, while those in lighter colors show the inference accuracy of MLPs without it. From Fig. 15, we can see that the Batch Normalization improves the inference accuracy in almost all configurations. In particular, when we quantize the weight matrix, the Batch Normalization significantly enhances accuracy. For example, in Leaky-ELU with the binary quantization, the Batch Normalization increases the accuracy by about 4.2% from 93.64% to 97.8%. Moreover, the differences in accuracy between activation functions, such as ReLU, ELU, and sigmoid, are reduced. Therefore, we demonstrate that the Batch Normalization, which is incorporated in the proposed architecture, significantly improves the accuracy. In summary, in this subsection, we experimentally confirm that the Batch Normalization is highly compatible with the quantized neural networks. Hence, the proposed architecture achieves good accuracy with a significant area reduction.

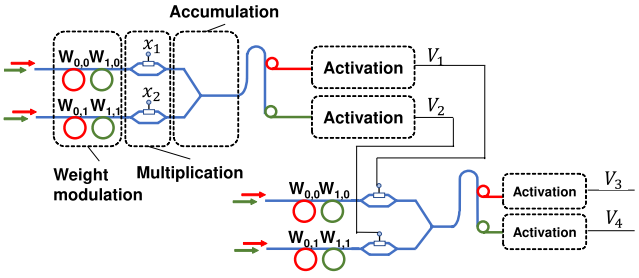


Fig. 16 Test Circuit Composed of Optical MAC Circuit and ELU-based Activation Circuit. The upper part represents the first layer, and the lower part is the second layer in MLP.

5.2 Optoelectronic Circuit Simulation

Lastly, this subsection examines the functional behavior of the proposed architecture via the optoelectronic circuit simulator. Section 5.2.1 explains the evaluation setups. Section 5.2.2 shows the evaluation results.

5.2.1 Evaluation Setup

As a target circuit, we design an optical multiplication and accumulation (OMAC) circuit, as shown in Fig. 16. This circuit represents two layers in MLP. In Fig. 16, x_1 and x_2 are inputs to the first layer in MLP, and $W_{0,0}$, $W_{0,1}$, $W_{1,0}$, and $W_{1,1}$ are weights in each layer of MLP. The two WDM optical signals given from the left of the upper section are weighted $W_{0,0}$, $W_{0,1}$, $W_{1,0}$, and $W_{1,1}$, respectively. Then, those signals are individually multiplied with the electrical signal inputs x_1 and x_2. The accumulation results are divided into two wavelengths, then passed to the activation function (ELU), and finally are given to the next layer as V_1 and V_2.

For verifying the behavior of the designed circuit, we use Optisystem and OptiSPICE. These are commercial optoelectronic circuit simulators and can analyze integrated optoelectronic circuits. In addition to MOS transistors, Optisystem and OptiSPICE can simulate optoelectric conversion in photodetectors and linear interference in MZMs and combiners at the transistor level. In this simulation, $W_{0,0}$,

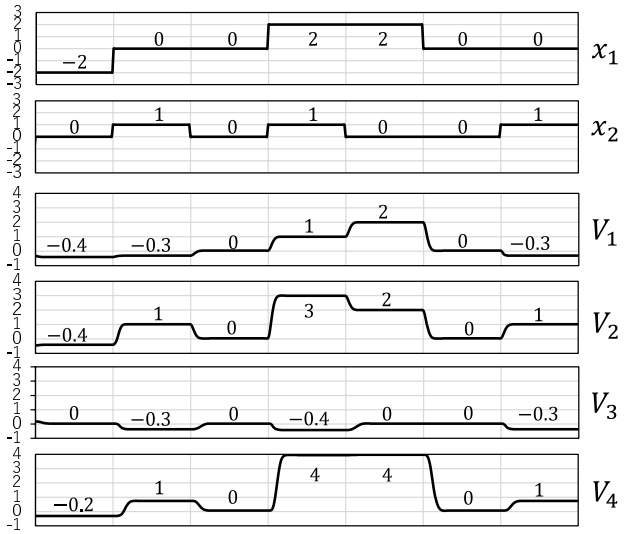


Fig. 17 Simulation results for Digital Optical MAC and ELU Activation Function. x_1 and x_2 represent electrical inputs to MZMs in Fig. 16. V_1 and V_2 are output from activation functions in the first layer, and V_3 and V_4 are that in the second layer.

$W_{1,0}$, and $W_{1,1}$ are all set to 1, and only $W_{0,1}$ is set to -1. Therefore, from V_1 to V_4 are represented by Eq. (4). In this evaluation, we confirm the functional behavior of the proposed circuit by observing all the values from V_1 to V_4.

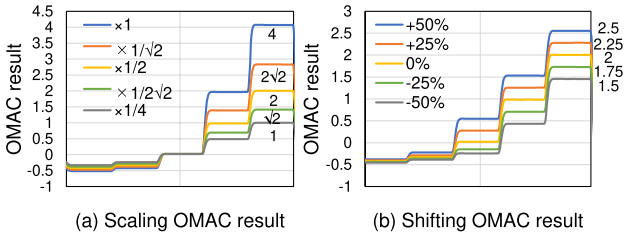
$$\begin{aligned} V_1 &= ELU(W_{0,0} \times x_1 + W_{0,1} \times x_2) = ELU(x_1 - x_2), \\ V_2 &= ELU(W_{1,0} \times x_1 + W_{1,1} \times x_2) = ELU(x_1 + x_2), \\ V_3 &= ELU(W_{0,0} \times V_1 + W_{0,1} \times V_2) = ELU(V_1 - V_2), \\ V_4 &= ELU(W_{1,0} \times V_1 + W_{1,1} \times V_2) = ELU(V_1 + V_2). \end{aligned} \tag{4}$$

5.2.2 Evaluation Results

This section evaluates the functional behavior of the designed circuit using optoelectronic circuit simulators for verifying whether the fundamental components in the proposed architecture, i.e., OMAC circuit and homodyne detector-based ELU activation circuit, work correctly.

Figure 17 shows the simulation results for the designed circuit in Fig. 16. Figure 17 indicates that the output, i.e., each value from V_1 to V_4, is identical to the accumulation result if the result is positive, which is expected behavior. Similarly, we observe that the output value reduces with the activation function if the accumulation result is negative. For example, when x_1 is equal to -2 and x_2 is equal to 0, the activation function circuit in the top row in Fig. 16 receives -2 as an OMAC result and outputs -0.4 as V_1. We experimentally confirm that the functional correctness of the optical MAC operation and optoelectronic ELU activation function through this evaluation.

Lastly, we simulate the functionality of the homodyne detector-based ELU activation circuit with Batch Normalization. Figure 18 shows the simulation results of the V_1 in Fig. 16. V_1 is scaled from 1/4 times to 1 time in Fig. 18(a) and shifted from -50% to $+50\%$ in Fig. 18(b), respectively. Figure 18 demonstrates that the outputs are scaled and shifted as expected by the Batch Normalization circuit. For example, in Fig. 18(a), $\times 1/2$ represents the reference light's electric field strength is halved by tuning the attenuator. Since the photocurrent I_{pd} is proportional to the reference light's electric field strength, the operation result is halved. Similarly, in Fig. 18(b), +50% indicates that the total photocurrent is increased by 50% from the total photocurrent in Fig. 17. For each 25% change in the current, the operation result also shifts by about 0.25. In other words, it is possible to scale and shift the operation result by tuning the LD power and increasing (decreasing) photocurrent drawn from newly added PDs. From the above, we experimentally confirm the functional correctness of the homodyne detector-based ELU activation circuit with Batch Normalization.



(a) Scaling OMAC result (b) Shifting OMAC result

Fig. 18 Simulation results for the scaling (the left part) and shifting (the right part) of V_1 by an activation function circuit with Batch Normalization. The configurations are the same as Fig. 17. $\times 1/2$ indicates that the reference light electric field is half in Fig. 16, which is the same result as V_1 in Fig. 17. 0% indicates that the total photocurrent is not changed from Fig. 17, which is the same result as V_1 in Fig. 17.

6. Conclusion

This paper proposes a new optical neural network architecture that fully exploits spatial parallelism and optical parallelism with wavelength division multiplexed (WDM) optical signals. Since only optical signals are propagated in a single neural network layer, its latency is extremely low. The MRR weight bank takes a few tens of picoseconds to read the weight parameters and modulate the optical signals based on the weight values [6]. Both the MZM multiplier and combiner-tree takes only a few picoseconds to propagate the optical signals [12]. The ELU-based activation circuit based on the amp-less O-E-O converter also takes a few tens of picoseconds [16]. As a result, the single neural network layer's propagation delay is less than 100 picoseconds in total approximately. If we construct a neural network with less than 10 layers, a sub-nanosecond neural network is realizable, which is extremely fast.

We investigate the trade-off relationship between the area and the accuracy of the proposed architecture with TensorFlow. The TensorFlow simulation indicates that our proposed architecture achieves high area-efficiency with satisfying good accuracy thanks to WDM, Batch Normalization, and the weight matrix quantization. Moreover, we also demonstrate the functional correctness of the proposed architecture's fundamental components, i.e., OMAC circuit and

homodyne detector-based ELU activation circuit, via opto-electronic circuit simulators. Unlike previous works, our architecture has a circuit structure where a small number of optical devices are serially connected.

Our future work will be devoted to developing a methodology for reducing the size of ONN without sacrificing the inference accuracy by pruning the neurons and their inputs appropriately.

Acknowledgments

This work is partly supported by JST CREST Grant Number JPMJCR15N4 and MEXT/JSPS KAKENHI Grant Number 20H04155.

References

[1] T. Ishihara, J. Shiomi, N. Hattori, Y. Masuda, A. Shinya, and M. Notomi, "An optical neural network architecture based on highly parallelized WDM-multiplier-accumulator," Proc. IEEE/ACM Workshop on Photonics-Optics Technology Oriented Networking, Information and Computing Systems, pp.15–21, Nov. 2019.

[2] A. Ceyhan, M. Jung, S. Panth, S.K. Lim, and A. Naeemi, "Impact of size effects in local interconnects for future technology nodes: A study based on full-chip layouts," Proc. IEEE Interconnect Technology Conference/Advanced Metallization Conference, pp.345–348, May 2014.

[3] Y. Cao, Predictive Technology Model for Robust Nanoelectronic Design, Springer, New York, NY, 2011.

[4] S. Sinha, B. Cline, G. Yeric, V. Chandra, and Y. Cao, "Design benchmarking to 7 nm with FinFET predictive technology models," Proc. International Symposium on Low Power Electronics and Design, pp.15–20, July 2012.

[5] X. Wu, J. Xu, Y. Ye, Z. Wang, M. Nikdast, and X. Wang, "SUOR: Sectioned undirectional optical ring for chip multiprocessor," J. Emerg. Technol. Comput. Syst., vol.10, no.4, pp.1–25, April 2014.

[6] A.N. Tait, T.F. de Lima, E. Zhou, A.X. Wu, M.A. Nahmias, B.J. Shastri, and P.R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," Scientific Reports, vol.7, no.1, Aug. 2017.

[7] W. Liu, W. Liu, Y. Ye, Q. Lou, Y. Xie, and L. Jiang, "HolyLight: A nanophotonic accelerator for deep learning in data centers," Proc. Design Automation, and Test in Europe, pp.1483–1488, March 2019.

[8] N. Janosik, Q. Cheng, M. Glick, Y. Huang, and K. Bergman, "High-resolution silicon microring based architecture for optical matrix multiplication," Proc. Conference on Lasers and Electro-Optics, no.SM2J.3, May 2019.

[9] M.B. On, H. Lu, H. Chen, R. Proietti, and S.J.B. Yoo, "Wavelength-space domain high-throughput artificial neural networks by parallel photoelectric matrix multiplier," Proc. IEEE Optical Fiber Communications Conference and Exhibition, pp.1–3, March 2020.

[10] Y. Shen, N.C. Harris, S. Skirlo, M. Prabhu, T.B.-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljacic, "Deep learning with coherent nanophotonic circuits," Nature, vol.11, no.7, pp.441–446, June 2017.

[11] J. Gu, Z. Zhao, C. Feng, Z. Ying, M. Liu, R.T. Chen, and D.Z. Pan, "Towards hardware-efficient optical neural networks: Beyond FFT architecture via joint learnability," IEEE Trans. Computer-Aided Des. Integr. Circuits Syst., vol.40, no.9, pp.1796–1809, 2021.

[12] S. Kita, K. Nozaki, K. Takata, A. Shinya, and M. Notomi, "Ultrashort low-loss ψ gates for linear optical logic on Si photonics platform," Commun. Phys., vol.3, no.33, March 2020.

[13] R. Hamerly, L. Bernstein, A. Sludds, M. Soljacic, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," Phys. Rev. X, vol.9, no.2, pp.21–32, May 2019.

[14] Q. Xu, B. Schmidt, J. Shakya, and M. Lipson, "Cascaded silicon micro-ring Modulators for WDM optical interconnection," Opt. Express, vol.14, no.20, pp.9431–9436, Oct. 2006.

[15] S. Chen, Y. Shi, S. He, and D. Dai, "Low-loss and broadband 2×2 silicon thermooptic Mach–Zehnder switch with bent directional couplers," Opt. Lett., vol.41, no.4, pp.836–839, Feb. 2016.

[16] K. Nozaki, S. Matsuo, T. Fujii, K. Takeda, A. Shinya, E. Kuramochi, and M. Notomi, "Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions," Nature Photonics, vol.13, pp.454–459, July 2019.

[17] V. Nair and G.E. Hinton, "Rectified linear units improve restricted Boltzmann machines," Proc. ICML'10, pp.807–814, 2010.

[18] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," arXiv preprint arXiv:1511.07289, 2015.

[19] A.L. Maas, A.Y. Hannun, and A.Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," Proc. ICML, vol.30, no.1, Citeseer, p.3, 2013.

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Proc. Int'l Conference on Machine Learning, pp.448–456, Dec. 2015.

[21] Y. Lecun, "The mnist database of handwritten digits," 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[22] M. Courbariaux, Y. Bengio, and J.-P. David, "BinaryConnect: Training deep neural networks with binary weights during propagations," Proc. Int'l Conference on Neural Information Processing Systems, pp.3123–3131, Dec. 2015.

[23] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[24] X. Ming, H. Zunkai, T. Li, W. Ning, Z. Yongxin, W. Hui, and F. Songlin, "An area-efficient 10-Bit buffer-reused DAC for AMOLED column driver ICs," Electronics, vol.9, no.2, p.208, Sept. 2020.

Naoki Hattori received the B.E. degree in Electronics and Information Engineering in 2020 from Nagoya University, Nagoya, Japan. He is currentry working toward the M.E. degree at Informatics from Nagoya University. His research interests include optical neural network (ONN) circuit architecture and design methodologies for ONN circuits.

Jun Shiomi received the B.E. degree in Electronic Engineering in 2014, the M.E. degree in Communications and Computer Engineering in 2016, and the Ph.D. degree in Informatics in 2017 all from Kyoto University, Kyoto, Japan. In 2017, he joined Kyoto University, where he is currently an Assistant Professor in the Department of Communications and Computer Engineering. His research interests include modeling and computer-aided design for low power and low voltage system-on-chips. Dr. Shiomi is a member of the IEEE and IPSJ.

Yutaka Masuda received the B.E., M.E., and Ph.D. degrees in Information Systems Engineering from the Osaka University, Osaka, Japan, in 2014, 2016, and 2019, respectively. He is currently an Assistant Professor in Center for Embedded Computing Systems, Graduate School of Informatics, Nagoya University. His research interests include low-power circuit design. He is a member of IEEE, IEICE, and IPSJ.

Tohru Ishihara received his Dr.Eng. degree in computer science from Kyushu University in 2000. For the next three years, he was a Research Associate in the University of Tokyo. From 2003 to 2005, he was with Fujitsu Laboratories of America as a Research Staff of an Advanced CAD Technology Group. From 2005 to 2011, he was with Kyushu University and for the next seven years he was with Kyoto University as an Associate Professor. In October 2018, he joined Nagoya University where he is currently a Professor in the Department of Computing and Software Systems. His research interests include low-power design methodologies and power management techniques for embedded systems. Dr. Ishihara is a member of the IEEE, ACM and IPSJ.

Akihiko Shinya received the B.E., M.E., and Ph.D. degrees in electrical engineering from Tokushima University in 1994, 1996, and 1999, respectively. In 1999, he joined NTT Basic Research Laboratories. He has been involved in R&D of photonic crystal devices. Dr. Shinya is a Member of the Japan Society of Applied Physics and the Laser Society of Japan.

Masaya Notomi received his B.E., M.E. and Ph.D. degrees in applied physics from University of Tokyo in 1986, 1988, and 1997, respectively. After joined NTT laboratories in 1988, his research interest has been photonic nanostructures for novel devices. He is entitled as Senior Distinguished Scientist in NTT, a director of NTT Nanophotonics Center, heading Photonic Nanostructure Research Group in NTT Basic Research Laboratories, and a guest professor of Tokyo Institute of Technology. He received IEEE/LEOS Distinguished Lecturer Award (2006), Japan Society for Promotion of Science Prize (2009), Japan Academy Medal (2009), and the Commendation by Japanese Minister of Education, Science and Technology (2010). IEEE Fellow since 2013.