An Optical Neural Network Architecture based on Highly Parallelized WDM-Multiplier-Accumulator

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

***Abstract*—Future applications such as anomaly detection in a network and autonomous driving require extremely low, sub-microsecond latency processing in pattern classification. Towards the realization of such an ultra-fast inference processing, this paper proposes an optical neural network architecture which can classify anomaly patterns at sub-nanosecond latency. The architecture fully exploits optical parallelism of lights using wavelength division multiplexing (WDM) in vector-matrix multiplication. It also exploits a linear optics with passive nanophotonic devices such as microring resonators, optical combiners, and passive couplers, which make it possible to construct low power and ultra-low latency optical neural networks. Optoelectronic circuit simulation using optical circuit implementation of multi-layer perceptron (MLP) demonstrates sub-nanosecond processing of optical neural network.**

***Index Terms*—optical neural network, wavelength division multiplexing, multi-layer perceptron**

I. 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 [1] and therefore, the RC time constant cannot be decreased by the transistor downscaling. Post-layout analysis using predictive technology models [2] shows that interconnect performance degradation may dominate over the device speed improvement in a 22 nm technology node and below [1] [3]. 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 advances in nanophotonics, however, make it possible to migrate power-efficient light-based communication into ever-shorter distances and move onto silicon chips as optical networks-on-chip [4].

In this paper, we propose an integrated optical neural network (ONN) architecture as a principle building block for in-network optical computing. The optical networks-on-chip (ONoC) described above [4] is a part of the in-network optical computing concept. Fundamental technologies required for the ONoCs such as memory interface and wavelength division multiplexing (WDM) are also the key technologies for the in-network ONN. An overview of our ONN is depicted in Fig. 1. This architecture fully exploits area efficiency of micro-ring modulators for interfacing electrical memory and exploits optical parallelism of light using WDM. It also exploits low power and ultra-low latency natures of linear optics in

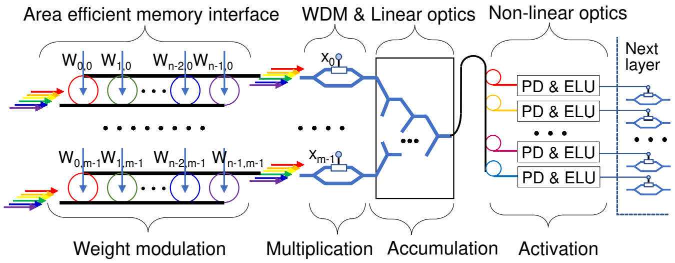


Fig. 1. Overview of our Optical Neural Network

Identify applicable funding agency here. If none, delete this.

optical passive devices such as combiners and couplers for constructing in-network highly parallelized multiplication and accumulation.

The rest of the paper is organized as follows. Section 2 summarizes several previous works on the optical neural networks and shows contributions of this work. Section 3 shows an architecture overview of our optical neural network. Section 4 shows experimental results obtained with a commercial optoelectronic circuit simulator. Section 5 concludes this paper.

II. 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.

A. *Photonic Weight Banks*

An optical circuit structure for calculating a weighted sum is proposed in [5]. This structure is used for vector-matrix multiplication in optical neural networks. The basic structure is depicted in Fig. 2. Incoming WDM signals are weighted by continuous-valued filters called microring (MRR) weight banks and then summed by a photodetector as photocurrent. This is a very area efficient, low power and low latency approach for calculating the weighted sum. It can calculate the weighted sum in a constant order 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, 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 neural network is dominated by the RC time-constant of the electronic low-pass filter. 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.

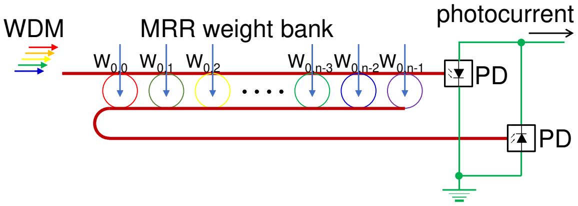


Fig. 2. Calculation of Weighted Sum Using Photonic Weight Banks

In [6], a similar architecture for optical vector-matrix multiplication based on the MRR weight banks is proposed. The architecture is very compact and implements the constant order calculation of the weighted sum using WDM signals and PD-based accumulation. However, it has the oscillation issue in the photodetector which limits the scalability of high-speed vector-matrix multiplication.

Unlike the approaches described above, we use a combiner tree [7] 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.

B. *Reconfigurable Mach-Zehnder Interferometer Array*

In [8], a fully optical neural network (ONN) architecture is presented for implementing general deep neural network algorithms using nanophotonic circuits that process coherent light. 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 [8], the energy consumption introduced by the switching activity is extremely small in this architecture.

However, one big drawback we see in the ONN architecture described above is large energy consumption in laser sources, which may limit the scalability of this architecture. Since the signal power attenuation is exponential to the number of MZIs connected in series, the signal power on the laser sources has to be sufficiently large so that the optical signals at the output can be surely detected by photodetectors even in case that the signals are largely attenuated when they are passing through the MZI array. The order of the number of serial connections in the architecture is $O(n)$, where the n is the number of nodes in a single neural network layer. If the size of the neural network layer increases, the energy consumption in laser sources increases exponentially.

Unlike the architecture in [8], our architecture does not have the circuit structure where a large number of optical devices are serially connected. Although our architecture also exploits

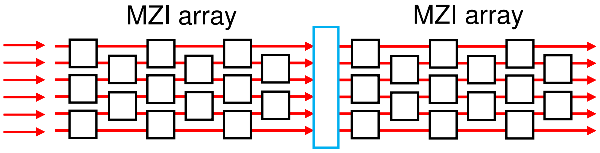


Fig. 3. Calculation of Weighted Sum Using MZI Array

the ultra-high speed nature of the serially connected optical devices, the order of the number of serial connections in our architecture is $O(log_2 n)$.

C. Homodyne Detection-based Vector Matrix Multiplier

A new type of photonic accelerator based on coherent detection is proposed in [9]. 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 [8], both weights and inputs are optically encoded so that the network can be reprogrammed and trained on the fly. This architecture is highly parallelized by the massive spatial multiplexing. However, it does not exploit WDM parallelism in their multiply-and-accumulate operation, which limits the throughput and area efficiency of the architecture.

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.

III. Optical Neural Network Exploiting WDM Parallelism

A. Neural Network Overview

An architectural overview of our optical neural network is depicted in Fig. 4. Each layer is composed of MRR weight bank [5], parallelized Mach Zehnder Modulator (MZM)-based multipliers, a combiner-tree-based accumulator [7], 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.

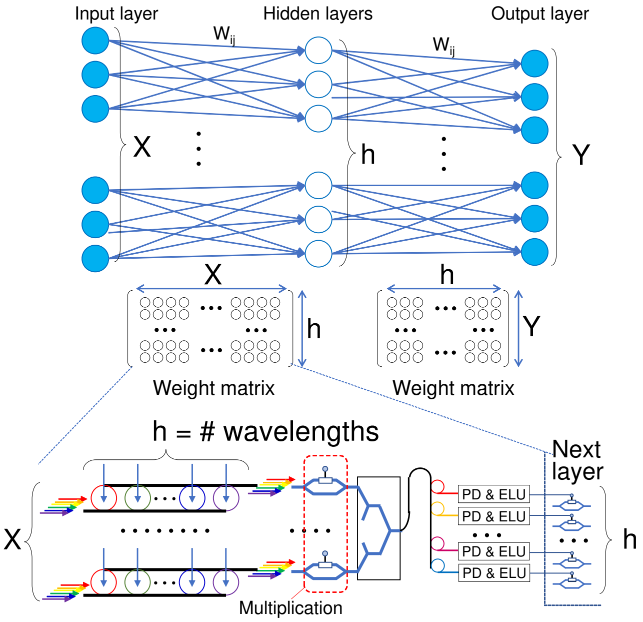


Fig. 4. Optical Neural Network Overview

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 are basically not interfered 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. Since we make decisions whether the corresponding output should be activated or not for every wavelengths separately, the number of activation circuits needed 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 [5]. Both the MZM multiplier and combiner-tree takes only a few picoseconds to propagate the optical signals [7]. The ELU-based activation circuit explained in the following sections also takes a few tens of picoseconds [13]. 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.

B. Vector Matrix Multiplication

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

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

We use an optical vector-matrix multiplier shown in Fig. 5 as a core part of the optical neural network computation. The

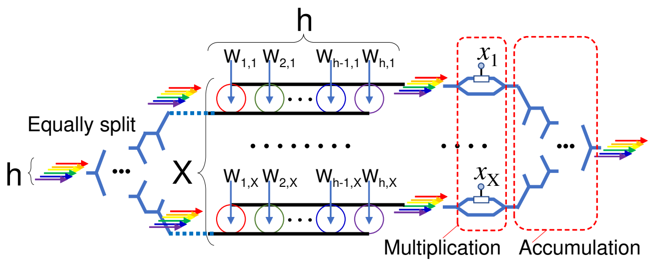

h
$W_{1,1}$ $W_{2,1}$ $W_{h-1,1}$ $W_{h,1}$
x_1
Equally split
h
X
$W_{1,X}$ $W_{2,X}$ $W_{h-1,X}$ $W_{h,X}$
x_X
Multiplication
Accumulation

Fig. 5. Vector Matrix Multiplier Overview

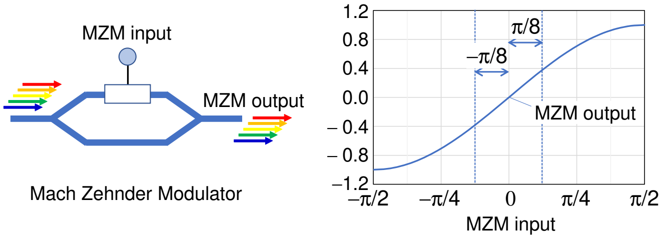


Fig. 6. MZM-based Multiplier Exploiting WDM

leftmost WDM signals are first equally divided into X groups and given to the rows of the MRR weight bank. Therefore, the signal power of all the $X \times h$ optical signals given to the MRR weight bank is the same from 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 (1). The MZM controlled by x_2 performs the multiplication corresponding to the second terms of all the equations in (1). Similarly, all the multiplications which appear in the equations in (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 because the size of the combiner can be made very small [7]. Optical signals with the same wavelength are accumulated in a tournament fashion by the combiner-tree. This accumulation corresponds to each equation in (1). For given WDM signals to the combiner-tree, all the accumulations in (1) are performed individually and concurrently.

1) MZM-based Parallelized Multiplication: We use a Mach-Zehnder Modulator (MZM) as an optical multiplier. By using optical signals with different wavelengths to each other, an MZM can work as a parallel multiplier for the multiple optical signals without mutual interference. It is experimentally demonstrated that optical inter-channel interference is negligible for channels with a wavelength spacing of 1.3 nm [11]. In [12], 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.

For obtaining good linearity of the sinusoidal wave generated by the MZM, we use the first $\pm\pi/8$ parts of the sinusoidal wave as shown in Fig. 6. This makes it possible to handle both plus and minus input values accurately and obtain good linearity in the MZM outputs for both plus and minus input values.

2) Combiner-based Parallelized Analog Accumulation: We use a combiner-tree-based analog accumulator presented in [7]. The accumulation is performed for massive WDM signals in parallel as shown in the leftmost tree structure depicted in Fig. 7. At every combiner in the tree, two input signals with the same wavelength are interfered and added to each other while for any two input signals with different wavelengths do not interfere. As a result, all WDM signals are 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.

Since the accumulation operation is based on the electric field strength instead of the signal power, the result of multiplication and accumulation (MAC) operation is obtained as a value proportional to the electric field strength. Therefore, we use a homodyne detection circuit shown at the middle in Fig. 7 for extracting the electric field strength from the photocurrent which is proportional to the signal power. We use O-E converter presented in [13] 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

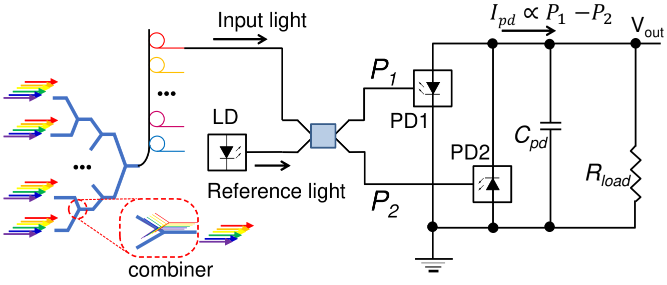


Fig. 7. Combiner-Tree and Homodyne Detection

very energy efficient. Once the photocurrent I_{pd} in Fig. 7, which is proportional to the electric field strength of input light is obtained, the I_{pd} is converted to electrical voltage just using a resistance R_{load}. If the input capacitance of the device connected to the V_{out} in Fig. 7 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 [13].

C. Activation Function

ReLU stands for rectified linear unit, and is a type of activation function. Mathematically, it is defined as (2)

$$y = max(0, x). \tag{2}$$

ReLU is the most commonly used activation function in neural networks, especially in CNNs because of the following reasons;

- It is cheap to compute as there is no complicated math. The model can, therefore, take less time for both training and inference.
- It converges faster. It does not have the vanishing gradient problem suffered by other activation functions like sigmoid or hyperbolic tangent.
- It is sparsely activated. Since ReLU is zero for all negative inputs, it is likely for any given unit to not activate at all. This is often desirable since it makes intuitive sense if we think about the biological neural network that artificial neural networks try to imitate.

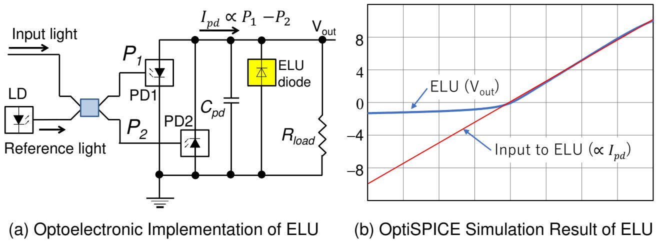
Although ReLU has several advantages over the other activation functions, it has a problem called dying ReLU. A ReLU neuron is dead if it is stuck on the negative side and always outputs 0. Because the slope of ReLU in the negative range is also 0, once a neuron gets negative, it is unlikely for it to recover. Such neurons are not playing any role in discriminating the input and is essentially useless.

Leaky ReLU 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, an exponential linear unit (ELU) has a small slope for negative values [10]. Instead of a straight line, it uses a log curve for negative values as shown in (3), where α is a scaling factor.

$$\begin{aligned} y &= x && (\text{if } x \geq 0), \\ y &= \alpha(e^x - 1) && (otherwise). \end{aligned} \tag{3}$$

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, in terms of implementation, ELU is the most practical activation function for the optoelectronic implementation of neural networks. Since a function which is not differentiable at some points is hard to implement using optical devices or MOS transistors, a differentiable function is preferable as an activation function. Although ReLU and leaky ReLU are not differentiable at zero, ELU is differentiable at any points. Fig. 8 (a) shows an optoelectronic implementation of ELU. This circuit is modified from the homodyne detection circuit depicted in Fig. 7. As explained for the homodyne detection circuit, the input of the ELU function is proportional to the total photocurrent obtained by summing up the photocurrents drawn from PD1 and PD2 in Fig. 8 (a). The yellow colored diode symbol, ELU diode which is newly added, works as an ELU function. For positive input values, ELU diode works as an insulator while it works as a resistor where its resistance is very small for negative input values. The circuit simulation result obtained with a commercial optoelectronic circuit simulator, OptiSPICE, is shown in Fig. 8 (b). The simulation result demonstrates that the ELU circuit shown in Fig. 8 (a) accurately works as the ELU function.



(a) Optoelectronic Implementation of ELU (b) OptiSPICE Simulation Result of ELU

Fig. 8. Optoelectronic Implementation of Activation Function.

IV. Experiments and Results

A. Experimental Setup

This section shows the following two evaluation results;

1) Linearity of MZM-based multiplier and combiner-based accumulator.
2) Functional correctness and accuracy of OMAC (optical multiplication and accumulation) circuit and homodyne detector-based ELU activation circuit combined together.

For evaluating the above two items, we designed an optoelectronic circuit shown in Fig. 9. For the first evaluation, we

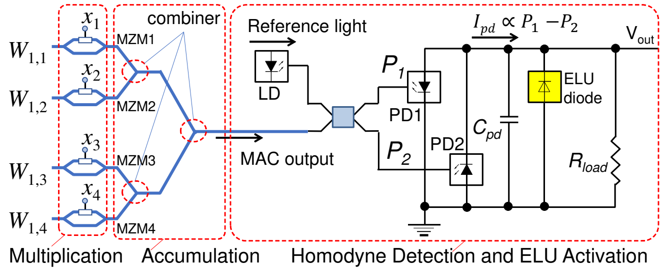


Fig. 9. Test Circuit Composed of Optical MAC Circuit and ELU-based Activation Circuit.

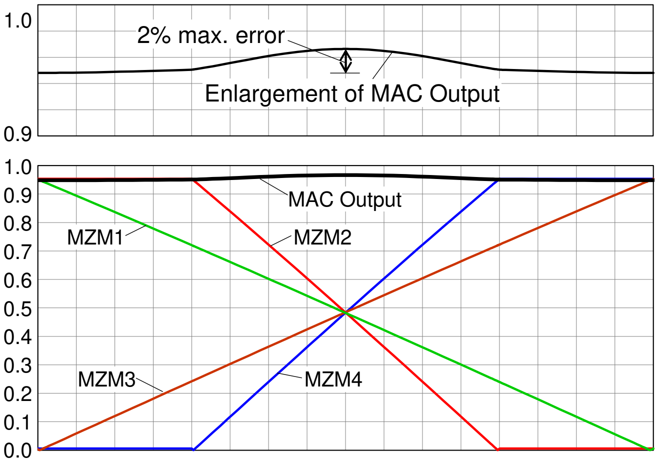


Fig. 10. Nonlinearity Evaluation Results.

use an MZM-based multiplication circuit and a combiner-tree-based accumulation circuit (i.e. left two parts) in Fig. 9. For the second evaluation, we use the entire circuit shown in Fig. 9.

B. Simulation Results

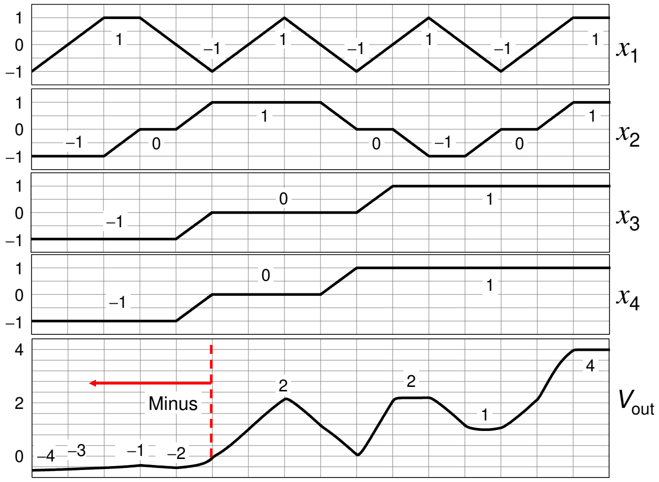


Fig. 11. Simulation Results for Analog Optical MAC and ELU Activation Function.

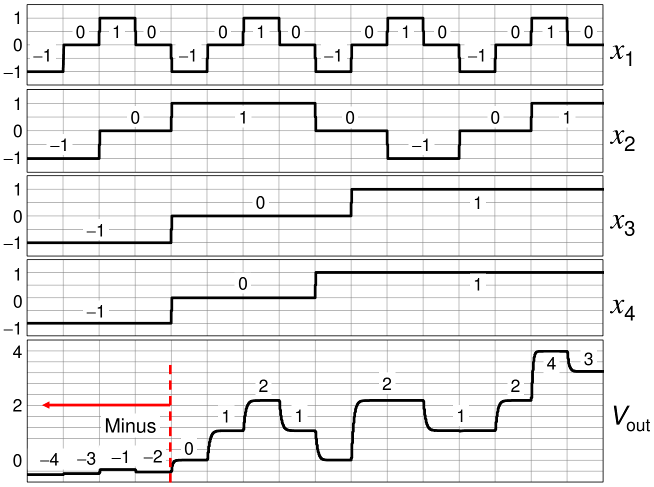


Fig. 12. Simulation Results for Digital Optical MAC and ELU Activation Function.

Fig. 10 shows the electric field strength values of four MZM outputs and MAC output, which correspond to the results of the first evaluation described above. We give electric voltage signals to x_1 to x_4 shown in Fig. 9 so that the electric field strength values in the corresponding MZM outputs exhibit the values shown in the lower part of Fig. 10. As explained with Fig. 6, MZM outputs have nonlinearity since we use a part of the sinusoidal curve produced by the optoelectronic response of the MZM as a linear function. The worst-case nonlinearity appears in the middle, where all the MZM outputs have the intermediate electric field strength. The upper part of Fig. 10 shows an enlargement of the result. At this worst case, 2% error is involved, which is not very big as an analog multiplier.

Fig. 11 and Fig. 12 show the results for the second evaluation. The results for analog inputs are shown in Fig. 11 and those for digital inputs are shown in Fig. 12. The upper four graphs in the figures represent input voltage signals given to the corresponding four inputs $x_1 \cdots x_4$ of MZM multipliers. The lower graphs in the figures show electrical voltage output signals of the ELU activation circuit. The figures demonstrate the functional correctness of the optical MAC operation and optoelectronic ELU activation function.

The reason why we perform the evaluation for digital inputs is because inference accuracy in quantized neural networks rapidly improves in recent years. Several related work [14]–[16] proved that inference accuracy does not widely degrade even if we use binary weights and binary activations while those binarizations largely reduce necessary memory and neural network sizes, and simplifies the memory interface. For example, BinaryConnect proposed in [14] obtains near state-of-the-art accuracy results using binary weights for the permutation-invariant MNIST. Our optical neural network architecture handles both analog values and digital values for weight parameters and inputs. If we use the binary inputs and binary weights, we can largely reduce the hardware complexity and footprint with negligible degradation in inference accuracy.

V. Conclusion

This paper proposes a new optical neural network architecture which fully exploits spatial parallelism and optical parallelism with wavelength division multiplexed (WDM) optical signals. Since, in a single neural network layer, only optical signals are propagated, its latency is extremely low. The MRR weight bank takes a few tens of picoseconds to read the weight parameters and to modulate the optical signals based on the weight values [5]. Both the MZM multiplier and combiner-tree takes only a few picoseconds to propagate the optical signals [7]. The ELU-based activation circuit based on the amp-less O-E-O converter also takes a few tens of picoseconds [13]. 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.

The optoelectronic circuit simulation demonstrated the functional correctness of our optical neural network architecture. Unlike previous works, our architecture has a circuit structure where a small number of optical devices are serially connected. This largely reduces the signal power attenuation along the serially connected path. Therefore, power consumption of laser sources needed for our architecture is very small.

To come up with a new algorithm for canceling out the impacts of the nonlinearity of optical devices and process variation on the inference accuracy through the training of weights is our future work.

References

[1] 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,” Proceedings of Interconnect Technology Conference / Advanced Metallization Conferenc, pp.345–348, May 2014.

[2] Y. Cao, “Predictive Technology Model for Robust Nanoelectronic Design,” Springer, 2011.

[3] S. Sinha, B. Cline, G. Yeric, V. Chandra and Y. Cao, “Design Benchmarking to 7nm with FinFET Predictive Technology Models,” Proceeding of International Symposium on Low Power Electronics and Design, pp.15–20, July 2012.

[4] X. Wu, J. Xu, Y. Ye, Z. Wang, M. Nikdast and X. Wang “SUOR: Sectioned Undirectional Optical Ring for Chip Multiprocessor,” ACM JETC, vol.10, no.4, pp.1–25, April 2014.

[5] A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, M.l A. Nahmias, B. J. Shastri and P. R. Prucnal “Neuromorphic Photonic Networks using Silicon Photonic Weight Banks,” Scientific Reports volume 7, Article number: 7430, August 2017.

[6] N. Janosik, Q. Cheng, M. Glick, Y. Huang, and K. Bergman, “High-resolution Silicon Microring based Architecture for Optical Matrix Multiplication,” in Proceeding of CLEO, SM2J.3, May 2019.

[7] S. Kita, K. Nozaki, K. Takata, A. Shinya, and M. Notomi “Silicon Linear Optical Logic Gates for Low-Latency Computing,” in Proceeding of CLEO, SF1A.2, May 2018.

[8] 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 Photonics, vol. 11, pp. 441-446, June 2017.

[9] R. Hamerly, L. Bernstein, A. Sludds, M. Soljacic, and D. Englund, “Large-Scale Optical Neural Networks Based on Photoelectric Multiplication,” Physical Review X 9, 021032, May 2019.

[10] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs),” arXiv:1511.07289v5, Feb. 2016.

[11] Q. Xu, B. Schmidt, J. Shakya, and M. Lipson, “Cascaded Silicon Micro-ring Modulators for WDM Optical Interconnection,” Optics Express, vol. 14, no. 20, pp. 9431–9436, October 2006.

[12] S. Chen, Y. Shi, S. He, and D. Dai, “Low-Loss and Broadband 2×2 Silicon Thermooptic MachZehnder Switch with Bent Directional Couplers,” Optics Letters, vol. 41, no. 4, pp. 836–839, February 2016.

[13] 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.

[14] M. Courbariaux and Y. Bengio, “BinaryConnect: Training Deep Neural Networks with Binary Weights During Propagations,” arXiv:1511.00363v3, April 2016.

[15] M. Courbariaux, I. Hubara, D. Soudry, R. E.-Yaniv and Y. Bengio, “Binarized Neural Networks: Training Neural Networks withWeights and Activations Constrained to $+1$ or -1,” arXiv:1602.02830v3, March 2016.

[16] M. Rastegariy, V. Ordonezy, J. Redmon, A. Farhadi, “XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks,” arXiv:1603.05279v4, August 2016.